DataScienceLab

2024-07-13

Contents

Load required libraries

```r
library(car)
library(readr)
library(MASS)
library(pscl)
library(ggplot2)
```

Import the dataset and rename selected columns for clarity

```r
ds <- read_csv("Financialliteracy.csv")
colnames(ds)[c(99:104)] <- c("Gender","Household","Age","Education","Employment","Country")
head(ds)
```

```
## # A tibble: 6 x 106
##      id pesofitc   qf1   qf2 qf3_1 qf3_3 qf3_4 qf3_6 qf3_7 qf3_8 qf3_99   qf4
##   <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1     1    0.707     2     0     0     0     0     0     0     1      0   -98
## 2     2    1.22      2     1     0     0     0     1     0     0      0     1
## 3     3    1.80      1     0     0     0     0     0     0     1      0     1
## 4     4    1.52      2     0     0     0     0     0     0     1      0   -98
## 5     5    0.245     2     0     0     0     0     0     0     1      0     0
## 6     7    2.12      2     0     0     0     0     0     0     0      1   -99
## # i 94 more variables: qf8 <dbl>, qf9_1 <dbl>, qf9_2 <dbl>, qf9_3 <dbl>,
## #   qf9_4 <dbl>, qf9_5 <dbl>, qf9_6 <dbl>, qf9_7 <dbl>, qf9_8 <dbl>,
## #   qf9_9 <dbl>, qf9_10 <dbl>, qf9_99 <dbl>, qprod1c_1 <dbl>, qprod1c_2 <dbl>,
## #   qprod1c_10 <dbl>, qprod1c_11 <dbl>, qprod1c_12 <dbl>, qprod1c_3 <dbl>,
## #   qprod1c_5 <dbl>, qprod1c_6 <dbl>, qprod1c_14 <dbl>, qprod1c_7 <dbl>,
## #   qprod1c_8 <dbl>, qprod1c_99 <dbl>, qprod1_d <dbl>, qprod2 <dbl>,
## #   qprod3_1 <dbl>, qprod3_2 <dbl>, qprod3_3 <dbl>, qprod3_4 <dbl>, ...
```

Convert appropriate variables to factors for categorical analysis

```r
cols_to_factor <- colnames(ds)[c(3:100,102,104:106)]
ds[cols_to_factor] <- lapply(ds[cols_to_factor], factor)
```

Provide an overview of socio-demographic variables

```r
require(skimr)
skim_without_charts(ds[99:106])
```

Table 1: Data summary

Name	ds[99:106]
Number of rows	2376
Number of columns	8
Column type frequency:	
factor	6
numeric	2
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Gender	0	1	FALSE	2	0: 1212, 1: 1164
Household	0	1	FALSE	6	2: 634, 3: 623, 4: 614, 1: 290
Education	0	1	FALSE	6	3: 925, 4: 717, 1: 537, 5: 171
Country	0	1	FALSE	2	1: 2314, 0: 62
SM	0	1	FALSE	2	0: 1420, 1: 956
AREA5	0	1	FALSE	5	1: 637, 4: 529, 2: 483, 3: 456

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Age	0	1	50.34	17.09	18	38	50	64	92
Employment	0	1	3.91	2.35	1	2	4	6	10

1 Descriptive Analysis

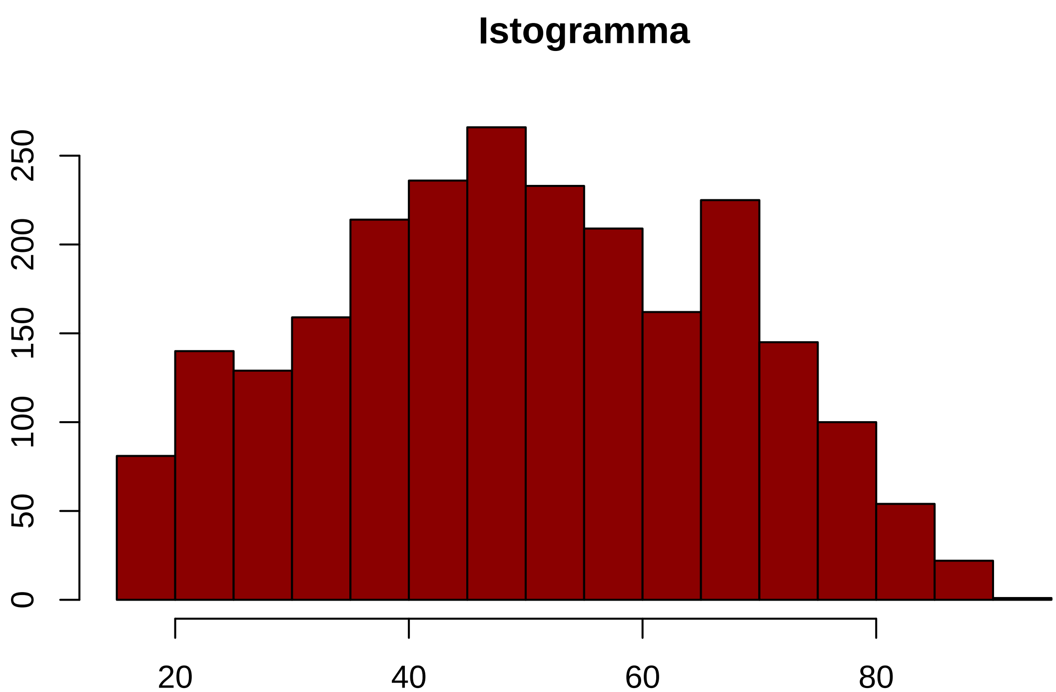
1.1 Univariate Analysis (Continuous Variables)

Duplicate the Age variable for further processing

```r
ds$Age1 <- ds$Age
```

Create a histogram of the Age variable

```r
hist(ds$Age , main="Istogramma",  xlab="", ylab="", col="red4")
```



Discretize the Age variable into categories based on age ranges

```
ds$Aged <- ifelse(ds$Age < 35,1,0)
ds$Aged <- ifelse(ds$Age >= 35 & ds$Age < 50,2,ds$Aged)
#ds$Aged <- ifelse(ds$Age >= 40 & ds$Age < 50,3,ds$Aged)
ds$Aged <- ifelse(ds$Age >= 50 & ds$Age < 65,3,ds$Aged)
ds$Aged <- ifelse(ds$Age >= 65,4,ds$Aged)
ds$Aged <- ordered(ds$Aged, levels= c(1:4))
table(ds$Aged)
```

```
##
##   1   2   3   4
## 480 688 621 587
```

1.2 Univariate Analysis (Discrete Variables)

1.2.1 Employment

Analyze the Employment variable and recategorize its levels table

```
table(ds$Employment)
```

```
##
##   1   2   4   5   6   9  10
## 263 867 264 229 571 161  21
```

1 Self-employed 2 In paid employment 4 -> 3 Looking after the home 5 -> 4 Looking for work
6 -> 5 Retired
9 -> 6 Student
10 -> 7 Other

```
#ds$Employment <- as.numeric(ds$Employment)
ds$Employment[ds$Employment == 4] <- 3
ds$Employment[ds$Employment == 5] <- 4
ds$Employment[ds$Employment == 6] <- 5
```

```
ds$Employment[ds$Employment == 9] <- 6
ds$Employment[ds$Employment == 10] <- 7

table(factor(ds$Employment))
```

```
##
##   1   2   3   4   5   6   7
## 263 867 264 229 571 161  21
```

```
ds$Employment <- factor(ds$Employment)
```

Create a binary variable for employment status (0 = unemployed, 1 = employed)

```
ds$Employment1 <- ifelse(ds$Employment %in% c(1, 2), 1, 0)
table(ds$Employment1)
```

```
##
##    0    1
## 1246 1130
```

1.2.2 Education

Analyze the Education variable and unify categories with low frequencies

1 University-level education 3 Complete secondary school 4 Some secondary school 5 Complete primary school 6 Some primary school 7 No formal education

```
table(ds$Education)
```

```
##
##   1   3   4   5   6   7
## 537 925 717 171  25   1
```

The variable "Education" is highly imbalanced.

Calculate the mean age for education levels 6 and 7

```
ds[which(ds$Education == 6),c(99:106)]
```

```
## # A tibble: 25 x 8
##    Gender Household   Age Education Employment Country SM    AREA5
##    <fct>  <fct>     <dbl> <fct>     <fct>      <fct>   <fct> <fct>
##  1 0      4            75 6         5          1       0     3
##  2 0      2            69 6         3          1       0     1
##  3 1      2            73 6         5          1       0     4
##  4 0      5            76 6         3          1       0     4
##  5 0      3            86 6         5          1       0     1
##  6 1      3            78 6         5          1       0     5
##  7 1      5            70 6         5          1       0     4
##  8 0      1            81 6         5          1       0     5
##  9 0      2            72 6         5          1       0     4
## 10 1      2            72 6         5          1       0     1
## # i 15 more rows
```

```
ds[which(ds$Education == 7),c(99:106)]
```

```
## # A tibble: 1 x 8
##   Gender Household   Age Education Employment Country SM    AREA5
##   <fct>  <fct>     <dbl> <fct>     <fct>      <fct>   <fct> <fct>
```

```
## 1 0      6             23 7         4          1        0      4
```

```r
mean(ds[which(ds$Education == 6),]$Age)
```

```
## [1] 77.32
```

```r
mean(ds[which(ds$Education == 7),]$Age)
```

```
## [1] 23
```

The only individual with "No formal Education" (7) is 23 years old. Given the compulsory education until 16 years in 2006, it is highly unlikely that this individual has no education. The average age of individuals with "Some primary school" (6) is 77 years old.

Given the low frequency of individuals with "Some primary school" (6) and "No formal education" (7), I will unify these categories into one category (5).

```r
ds$Education[ds$Education == 6 | ds$Education == 7 ] <- 5
table(ds$Education)
```

```
##
##   1   3   4   5   6   7
## 537 925 717 197   0   0
```

Make the variable suitable for analysis by assigning a value from 1 to 3 to each category. As the level of education increases, the value increases. Therefore:

1 Some secondary school - Complete primary school - Some primary school - No formal education 2 Complete secondary school 3 University-level education

```r
Education <- ds$Education
ds$Education <- as.numeric(ds$Education)
ds$Education[Education == 4| Education == 5]<-1
ds$Education[Education == 3]<-2
ds$Education[Education == 1]<-3
ds$Education <- ordered(factor(ds$Education),levels=c(1:3))
table(ds$Education)
```

```
##
##   1   2   3
## 914 925 537
```

1.2.3 Households

Simplify household categories by merging groups with low frequencies

```r
table(ds$Household)
```

```
##
##   1   2   3   4   5   6
## 290 634 623 614 161  54
```

Given the low frequency of families with 5 or more members, I will unify categories 5 and 6 into one category (4).

```r
ds$Household[ds$Household == 6 | ds$Household == 5] <- 4
ds$Household <- ordered(ds$Household, levels = c(1:4))
table(ds$Household)
```

```
##
##   1   2   3   4
## 290 634 623 829
```

1.2.4 Gender

Check if the Gender variable is balanced across categories

```
prop.table(table(ds$Gender))
```

```
##
##        0        1
## 0.510101 0.489899
```

The variable is balanced.

1.2.5 Country

Analyze the Country variable

```
table(ds$Country)
```

```
##
##    0    1
##   62 2314
```

The variable is highly imbalanced. We have very few individuals not born in Italy.

1.2.6 Area5

Examine geographic area distribution

```
table(ds$AREA5)
```

```
##
##   1   2   3   4   5
## 637 483 456 529 271
```

The categories related to geographic area are balanced, except for category 5 corresponding to the Islands, which has a slightly lower frequency.

1.3 Multivariate Analysis

1.3.1 Education & Area5

Cross-tabulate Education and Area5 variables to analyze relationships

```
tab<-table(Education = ds$Education, area = ds$AREA5)
tab
```

```
##          area
## Education   1   2   3   4   5
##         1 242 194 153 195 130
##         2 253 188 197 206  81
##         3 142 101 106 128  60
```

```
# Relative frequencies
prop.table(tab)
```

```
##          area
## Education          1          2          3          4          5
##         1 0.10185185 0.08164983 0.06439394 0.08207071 0.05471380
##         2 0.10648148 0.07912458 0.08291246 0.08670034 0.03409091
##         3 0.05976431 0.04250842 0.04461279 0.05387205 0.02525253
```

```r
# Margin relative frequencies
prop.table(tab,margin=2)
```

```
##          area
## Education         1         2         3         4         5
##         1 0.3799058 0.4016563 0.3355263 0.3686200 0.4797048
##         2 0.3971743 0.3892340 0.4320175 0.3894140 0.2988930
##         3 0.2229199 0.2091097 0.2324561 0.2419660 0.2214022
```

```r
prop.table(tab,margin=1)
```

```
##          area
## Education          1          2          3          4          5
##         1 0.26477024 0.21225383 0.16739606 0.21334792 0.14223195
##         2 0.27351351 0.20324324 0.21297297 0.22270270 0.08756757
##         3 0.26443203 0.18808194 0.19739292 0.23836127 0.11173184
```

1.3.2 Education & Gender

Cross-tabulate Education and Gender variables; visualize with mosaic plot

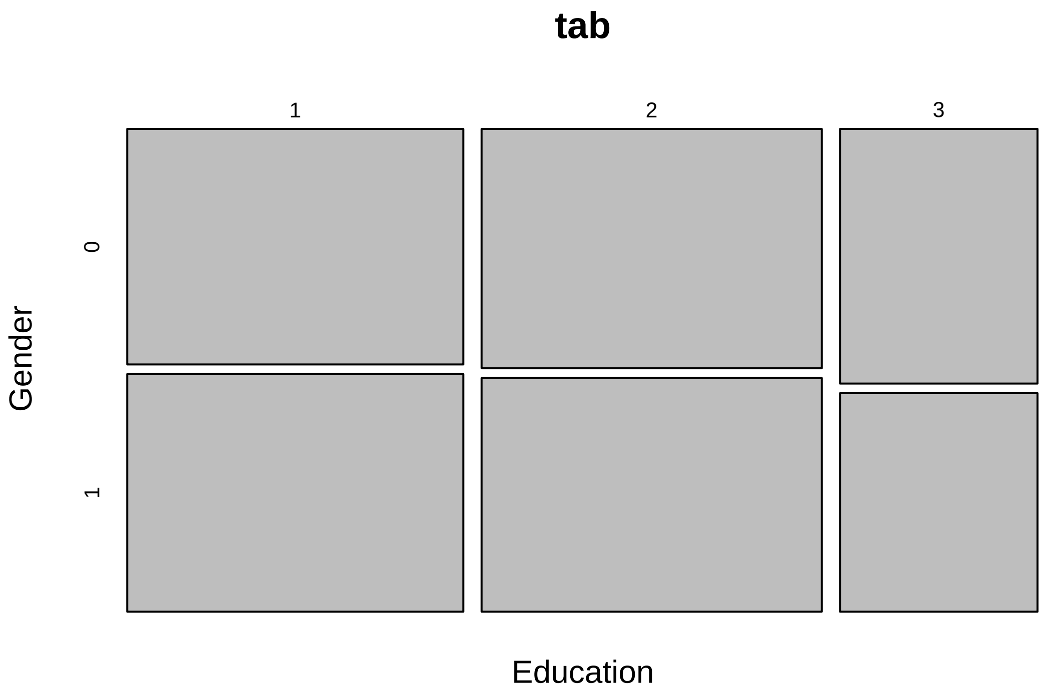
```r
tab<-table(Education = ds$Education, Gender = ds$Gender)
tab
```

```
##          Gender
## Education   0   1
##         1 455 459
##         2 468 457
##         3 289 248
```

```r
# Relative frequencies
prop.table(tab, margin=2)
```

```
##          Gender
## Education         0         1
##         1 0.3754125 0.3943299
##         2 0.3861386 0.3926117
##         3 0.2384488 0.2130584
```

```r
mosaicplot(tab)
```



1.3.3 Education & Household

Cross-tabulate Education and Household variables; visualize with mosaic plot

```
tab<-table(Education = ds$Education, Household = ds$Household)
tab
```

```
##          Household
## Education   1   2   3   4
##         1 132 308 204 270
##         2  93 221 275 336
##         3  65 105 144 223
```

```
# Relative frequencies
prop.table(tab)
```

```
##          Household
## Education          1          2          3          4
##         1 0.05555556 0.12962963 0.08585859 0.11363636
##         2 0.03914141 0.09301347 0.11574074 0.14141414
##         3 0.02735690 0.04419192 0.06060606 0.09385522
```

```
prop.table(tab, margin=1)
```

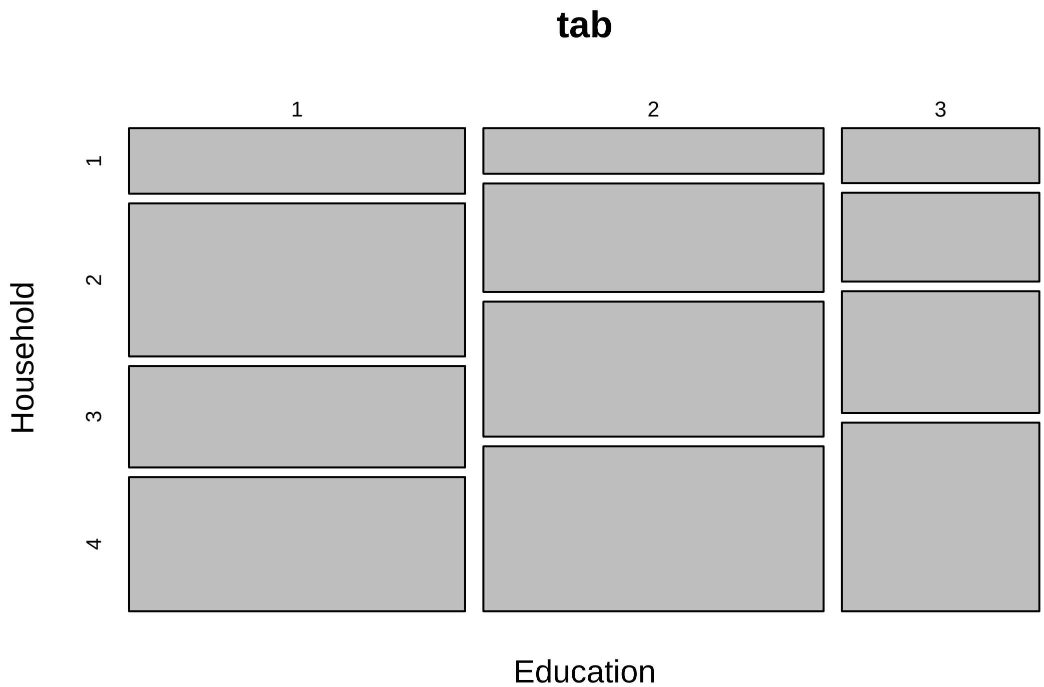
```
##          Household
## Education         1         2         3         4
##         1 0.1444201 0.3369803 0.2231947 0.2954048
##         2 0.1005405 0.2389189 0.2972973 0.3632432
##         3 0.1210428 0.1955307 0.2681564 0.4152700
```

```
prop.table(tab, margin=2)
```

```
##          Household
## Education         1         2         3         4
##         1 0.4551724 0.4858044 0.3274478 0.3256936
```

```
##          2 0.3206897 0.3485804 0.4414125 0.4053076
##          3 0.2241379 0.1656151 0.2311396 0.2689988
```

```
mosaicplot(tab)
```



1.3.4 Household & Area5

Analyze the relationship between household size and geographic area

```
tab<-table(Household = ds$Household, Area= ds$AREA5)
prop.table(tab, margin = 1) # Conditional frequencies by household size
```

```
##           Area
## Household          1          2          3          4          5
##         1 0.34137931 0.26206897 0.17241379 0.14482759 0.07931034
##         2 0.29022082 0.22239748 0.22082019 0.15772871 0.10883281
##         3 0.25682183 0.24558587 0.18940610 0.21990369 0.08828250
##         4 0.23401689 0.13630881 0.17852835 0.30156815 0.14957780
```

Observing the table of conditional frequencies, we can see that there are more families with a household size of 5 (32%) in Southern Italy (Area=4) compared to other geographical areas. This is an opposite trend to Southern Italy, where we have a lower conditional frequency for families with only one individual.

2 Financial Knowledge Analysis

Perform a preliminary analysis of financial knowledge questions. Summarize the structure and missing values in the dataset for variables qk3 to qk7_3.

```
skim_without_charts(ds[92:98])
```

Table 4: Data summary

Name	ds[92:98]
Number of rows	2376

Number of columns	7
Column type frequency:	
factor	7
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
qk3	0	1	FALSE	5	3: 1152, 2: 723, -97: 372, 1: 78
qk4	0	1	FALSE	16	0: 1321, -97: 798, -99: 224, 10: 12
qk5	0	1	FALSE	42	102: 1141, -97: 638, -99: 129, 100: 98
qk6	0	1	FALSE	6	1: 803, 2: 575, -97: 481, 3: 255
qk7_1	0	1	FALSE	4	1: 1790, -97: 360, 0: 195, -99: 31
qk7_2	0	1	FALSE	4	1: 1728, -97: 356, 0: 260, -99: 32
qk7_3	0	1	FALSE	4	1: 918, -97: 825, 0: 590, -99: 43

No missing values

Check for missing values (-99) in financial knowledge questions

```
tab_99<-data.frame(
  c(
    length(which(ds$qk3 == -99)),
    length(which(ds$qk4 == -99)),
    length(which(ds$qk5 == -99)),
    length(which(ds$qk6 == -99)),
    length(which(ds$qk7_1 == -99)),
    length(which(ds$qk7_2 == -99)),
    length(which(ds$qk7_3 == -99))
  ),
  row.names = colnames(ds[92:98])
)
colnames(tab_99) <- "N_noAnsware"
tab_99
```

```
##       N_noAnsware
## qk3            51
## qk4           224
## qk5           129
## qk6            86
## qk7_1          31
## qk7_2          32
## qk7_3          43
```

Assign scores to financial knowledge questions based on correct answers: All answers with value "-99" are assigned a score of 0. Calculate the knowledge score by assigning 1 point if: - qk3 = 3 - qk4 = 0(%) - qk5 = 102 - qk6 = 1 - qk7_1 = 1 - qk7_2 = 1 - qk7_3 = 1 For the remaining values, assign a score of 0. The knowledge score ranges from 0 to 7.

```
know<-ds[1]
know$qk3 <- ifelse(ds$qk3 == 3,1,0)
```

```
know$qk4 <- ifelse(ds$qk4 == 0,1,0)
know$qk5 <- ifelse(ds$qk5 == 102 ,1,0)
know$qk6 <- ifelse(ds$qk6 == 2 ,1,0)
know$qk7_1 <- ifelse(ds$qk7_1 == 1,1,0)
know$qk7_2 <- ifelse(ds$qk7_2 == 1,1,0)
know$qk7_3 <- ifelse(ds$qk7_3 == 1,1,0)

# Calculate total score (0-7) based on correct answers
know$tot <- unlist(know$qk3+know$qk4+know$qk5+know$qk6+know$qk7_1+know$qk7_2+know$qk7_3)
know$tot <- ordered(know$tot, levels = c(0:7))
```

2.1 Analyze Financial Knowledge Score

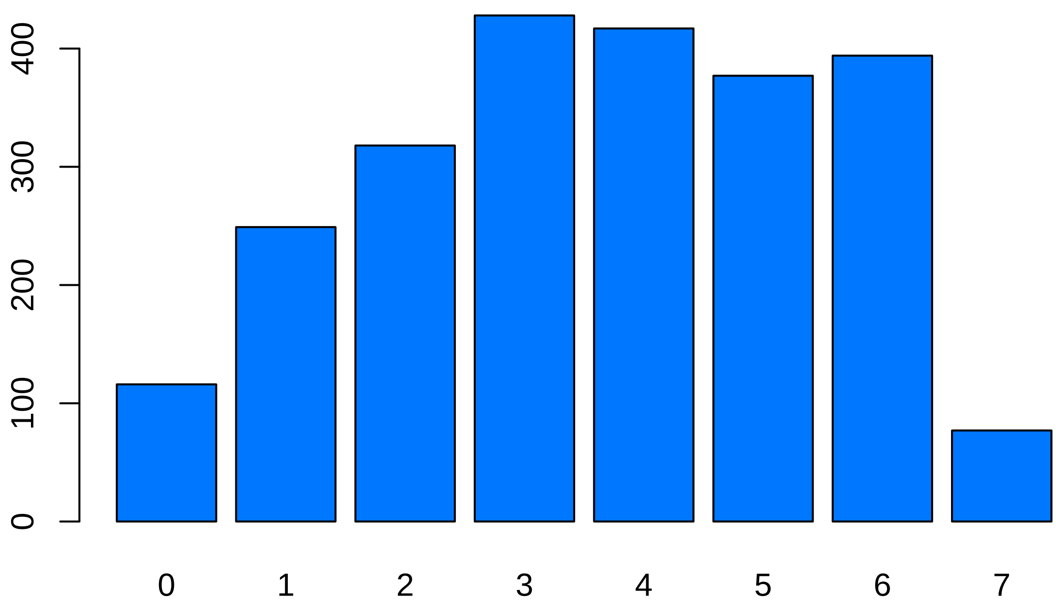
Display frequency distribution of financial knowledge scores

```
table(know$tot)
```

```
##
##   0   1   2   3   4   5   6   7
## 116 249 318 428 417 377 394  77
```

Plot the distribution of scores

```
plot(know$tot,col=c("#0077FF"))
```

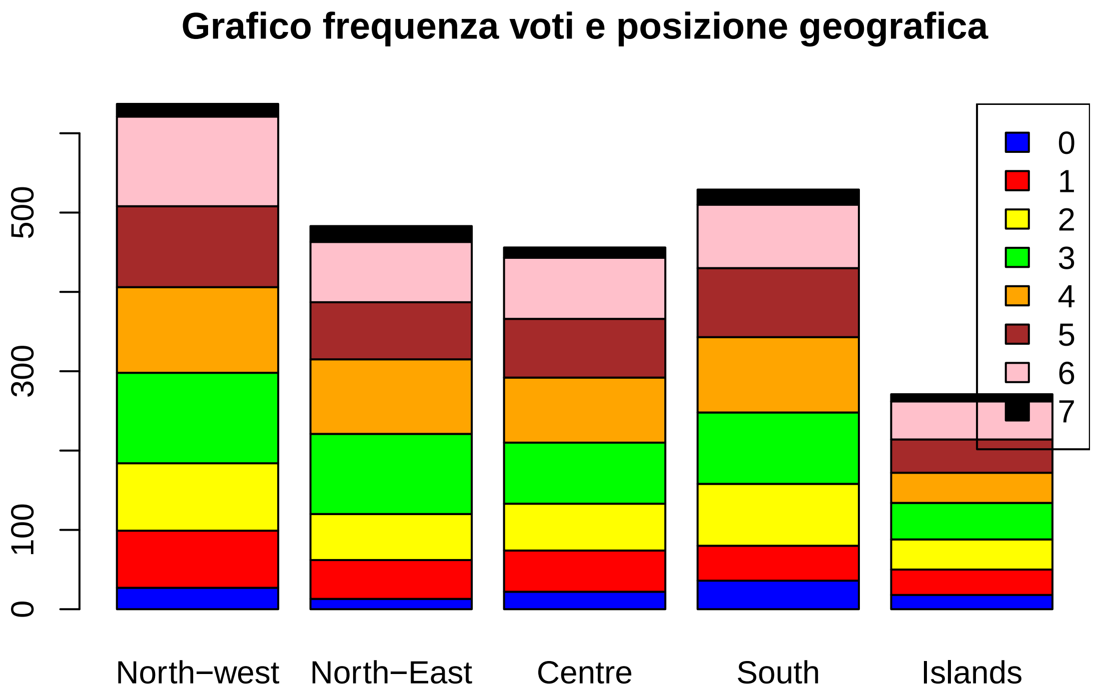



```
tab<-table(score = know$tot, area = ds$AREA5)
tab
```

```
##      area
## score   1   2   3   4   5
##     0  27  13  22  36  18
##     1  72  49  52  44  32
##     2  85  58  59  78  38
##     3 114 101  77  90  46
##     4 108  94  82  95  38
##     5 102  72  74  87  42
##     6 113  76  77  80  48
```

```
##     7  16  20  13  19   9
```

```
barplot(tab,col=c("blue","red","yellow","green","orange","brown","pink","black"),main="Grafico frequenza
legend("topright",                        # Posizione della legenda
       legend = c("0","1","2","3","4","5","6",7),  # Etichette della legenda (categorie)
       fill = c("blue","red","yellow","green","orange","brown","pink","black"))
```




```
#frequenze relative
prop.table(tab)
```

```
##      area
## score           1           2           3           4           5
##     0 0.011363636 0.005471380 0.009259259 0.015151515 0.007575758
##     1 0.030303030 0.020622896 0.021885522 0.018518519 0.013468013
##     2 0.035774411 0.024410774 0.024831650 0.032828283 0.015993266
##     3 0.047979798 0.042508418 0.032407407 0.037878788 0.019360269
##     4 0.045454545 0.039562290 0.034511785 0.039983165 0.015993266
##     5 0.042929293 0.030303030 0.031144781 0.036616162 0.017676768
##     6 0.047558923 0.031986532 0.032407407 0.033670034 0.020202020
##     7 0.006734007 0.008417508 0.005471380 0.007996633 0.003787879
```

```
#Eta media per ciascun livello
tapply(ds$Age, know$tot, mean)
```

```
##        0        1        2        3        4        5        6        7
## 50.25000 49.38153 51.06604 51.25467 51.49880 50.12732 48.62437 48.92208
```

2.2 Ordinal Regression Model

Observe the distribution of the financial knowledge score across socio-demographic variables

```
table(know$tot, ds$Gender)
```

```
##
##       0   1
```

```
##   0  61  55
##   1 137 112
##   2 171 147
##   3 229 199
##   4 222 195
##   5 178 199
##   6 179 215
##   7  35  42
```

```r
table(know$tot, ds$Aged)
```

```
##
##       1   2   3   4
##   0  25  36  22  33
##   1  63  60  62  64
##   2  64  95  66  93
##   3  82 121 115 110
##   4  72 123 118 104
##   5  78 109 102  88
##   6  82 122 112  78
##   7  14  22  24  17
```

```r
table(know$tot, ds$Education)
```

```
##
##       1   2   3
##   0  60  36  20
##   1 127  86  36
##   2 159 103  56
##   3 173 165  90
##   4 162 174  81
##   5 118 151 108
##   6  96 179 119
##   7  19  31  27
```

Visualize the distribution of financial knowledge scores

```r
table(know$tot)
```

```
##
##   0   1   2   3   4   5   6   7
## 116 249 318 428 417 377 394  77
```

Reduce the number of levels in the financial knowledge variable from 8 to 3 categories. Group low scores (0-2), medium scores (3-4), and high scores (5-7).

```r
know$tot1 <- know$tot
know$tot1[know$tot == 0 | know$tot == 1 | know$tot == 2] <- 1
know$tot1[know$tot == 3 | know$tot == 4] <- 2
know$tot1[know$tot == 5 |know$tot == 6 | know$tot == 7] <- 3
know$tot1 <- ordered(factor(know$tot1),levels=c(1:3))
```

Visualize the distribution of the new financial knowledge variable

```r
table(know$tot1)
```

```
##
##   1   2   3
## 683 845 848
```

2.2.1 Full Ordinal Regression Model

Fit a full ordinal regression model to predict financial knowledge levels based on demographic variables.

```
mod1_1<- polr(know$tot1 ~
                Gender +
                Household +
                Aged +
                Education +
                Employment1 +
                AREA5,
              ds)
summary(mod1_1)
```

```
## Call:
## polr(formula = know$tot1 ~ Gender + Household + Aged + Education +
##     Employment1 + AREA5, data = ds)
##
## Coefficients:
##                 Value Std. Error  t value
## Gender1      0.252231    0.07852  3.21214
## Household.L  0.126173    0.09512  1.32650
## Household.Q -0.197649    0.08219 -2.40491
## Household.C  0.036787    0.07734  0.47565
## Aged.L       0.219926    0.09522  2.30962
## Aged.Q      -0.158901    0.08968 -1.77181
## Aged.C      -0.102218    0.07405 -1.38037
## Education.L  0.710353    0.07757  9.15722
## Education.Q -0.115681    0.06582 -1.75766
## Employment1  0.074949    0.09448  0.79328
## AREA52       0.091789    0.11171  0.82164
## AREA53      -0.055400    0.11479 -0.48260
## AREA54      -0.042349    0.11141 -0.38012
## AREA55       0.004706    0.13856  0.03396
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -0.8246  0.0978    -8.4293
## 2|3  0.7394  0.0975     7.5843
##
## Residual Deviance: 5067.364
## AIC: 5099.364
```

2.2.2 Variable Selection with Stepwise Regression

Anova analysis of the full model to identify significant predictors.

```
Anova(mod1_1, type = "II", test.statistic = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: know$tot1
##             LR Chisq Df Pr(>Chisq)
## Gender        10.338  1   0.001304 **
## Household      6.248  3   0.100134
## Aged          12.029  3   0.007283 **
```

```
## Education     94.663  2  < 2.2e-16 ***
## Employment1    0.629  1   0.427685
## AREA5          1.850  4   0.763282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Perform stepwise regression to identify significant predictors for financial knowledge levels.

```r
step(mod1_1)
```

```
## Start:  AIC=5099.36
## know$tot1 ~ Gender + Household + Aged + Education + Employment1 +
##     AREA5
##
##                Df    AIC
## - AREA5         4 5093.2
## - Employment1   1 5098.0
## <none>            5099.4
## - Household     3 5099.6
## - Aged          3 5105.4
## - Gender        1 5107.7
## - Education     2 5190.0
##
## Step:  AIC=5093.21
## know$tot1 ~ Gender + Household + Aged + Education + Employment1
##
##                Df    AIC
## - Employment1   1 5091.9
## <none>            5093.2
## - Household     3 5093.5
## - Aged          3 5099.1
## - Gender        1 5101.5
## - Education     2 5183.3
##
## Step:  AIC=5091.94
## know$tot1 ~ Gender + Household + Aged + Education
##
##              Df    AIC
## <none>          5091.9
## - Household  3 5092.0
## - Aged       3 5098.9
## - Gender     1 5101.7
## - Education  2 5187.5

## Call:
## polr(formula = know$tot1 ~ Gender + Household + Aged + Education,
##     data = ds)
##
## Coefficients:
##     Gender1 Household.L Household.Q Household.C      Aged.L      Aged.Q
##  0.26412487  0.10602191 -0.19956058  0.03015021  0.20258318 -0.19099459
##      Aged.C Education.L Education.Q
## -0.10053613  0.71614175 -0.11598886
##
## Intercepts:
##        1|2        2|3
```

```
## -0.8564399  0.7062942
##
## Residual Deviance: 5069.944
## AIC: 5091.944
```

2.2.3 Reduced Ordinal Regression Model

Fit a reduced ordinal regression model with selected variables.

```
mod1_2<- polr(know$tot1 ~ Gender + Age1 + Education, ds)
summary(mod1_2)
```

```
## Call:
## polr(formula = know$tot1 ~ Gender + Age1 + Education, data = ds)
##
## Coefficients:
##                 Value Std. Error t value
## Gender1      0.268754   0.076412   3.517
## Age1         0.004613   0.002369   1.948
## Education.L  0.702029   0.076167   9.217
## Education.Q -0.141958   0.064932  -2.186
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -0.6486  0.1326    -4.8913
## 2|3  0.9058  0.1332     6.7985
##
## Residual Deviance: 5086.209
## AIC: 5098.209
```

Visualize the summary of the reduced model

```
summary_table <- coef(summary(mod1_2))
pval <- pt(abs(summary_table[, "t value"]),lower.tail = FALSE,nrow(ds)-4)
summary_table <- cbind(summary_table, "p value" = round(pval,5))
summary_table
```

```
##                    Value  Std. Error   t value p value
## Gender1      0.268753725 0.076412458  3.517145 0.00022
## Age1         0.004613213 0.002368657  1.947607 0.02579
## Education.L  0.702029111 0.076166564  9.217025 0.00000
## Education.Q -0.141957878 0.064932290 -2.186245 0.01445
## 1|2         -0.648594203 0.132600748 -4.891331 0.00000
## 2|3          0.905779342 0.133231440  6.798541 0.00000
```

The Brant test is used to check the proportional odds assumption in ordinal regression models.

```
library(brant)
brant(mod1_2)
```

```
## --------------------------------------------
## Test for X2  df  probability
## --------------------------------------------
## Omnibus      7.46    4   0.11
## Gender1      1.91    1   0.17
## Age1     3.89    1   0.05
## Education.L  0.2 1   0.66
```

```
## Education.Q  0.88    1   0.35
## --------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

The null hypothesis of the Brant test is that the proportional odds assumption holds. The assumption of parallel regression seems to be satisfied, indicating that the chosen ordinal regression model is appropriate.

2.2.4 Compare Full and Reduced Models

Compare the full and reduced models using a likelihood ratio test.

```
anova(mod1_1,mod1_2,  test = "Chisq")
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: know$tot1
##                                                      Model Resid. df
## 1                                   Gender + Age1 + Education      2370
## 2 Gender + Household + Aged + Education + Employment1 + AREA5      2360
##   Resid. Dev   Test    Df LR stat.    Pr(Chi)
## 1   5086.209
## 2   5067.364 1 vs 2    10 18.84507 0.04227524
```

The p-value of the likelihood ratio test is 0.0573, indicating that the difference between the two models is not highly significant. This implies that the null hypothesis (which states that the reduced model is sufficient to explain the data) should be accepted.

2.2.5 Compare Reduced Model and Null Model

Compare the reduced model to a null model to assess its explanatory power.

```
mod1_0 <- polr(know$tot1 ~ 1)
anova(mod1_2, mod1_0, test = "Chisq")
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: know$tot1
##                      Model Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
## 1                        1      2374   5197.525
## 2 Gender + Age1 + Education      2370   5086.209 1 vs 2     4 111.3162       0
```

Select the reduced model (mod1_2) as it is significantly better than the null model.

3 Financial Attitude Analysis

Analyze financial attitude using responses to questions QF10_2, QF10_3, and QF10_5.

Display frequency distributions for each question.

```
table(ds$qf10_2)
```

```
##
## -99 -97   1   2   3   4   5
##  19  34 339 424 612 438 510
```

```
table(ds$qf10_3)
```

```
##
## -99 -97   1   2   3   4   5
##  24  44 155 389 772 497 495
```

```r
table(ds$qf10_5)
```

```
##
##  -99  -97    1    2    3    4    5
##   34   54   79  193  421  441 1154
```

Analyze missing responses (-99) and “don’t know” answers (-97) in financial attitude questions.

```r
ds[which((ds$qf10_2 == -97 | ds$qf10_2 == -99) & (ds$qf10_3 == -97 | ds$qf10_3 == -99) & (ds$qf10_8 ==
```

```
## # A tibble: 5 x 4
##       id qf10_2 qf10_3 qf10_8
##    <dbl> <fct>  <fct>  <fct>
## 1    551 -99    -99    -99
## 2   1285 -99    -99    -99
## 3   1312 -97    -97    -97
## 4   1328 -97    -97    -97
## 5 686590 -99    -99    -99
```

```r
ds[which((ds$qf10_3 == -97 | ds$qf10_3 == -99) & (ds$qf10_8 == -97 | ds$qf10_8 == -99)),c(59,60,65)]
```

```
## # A tibble: 18 x 3
##    qf10_2 qf10_3 qf10_8
##    <fct>  <fct>  <fct>
##  1 2      -97    -99
##  2 1      -97    -97
##  3 1      -99    -99
##  4 -99    -99    -99
##  5 1      -97    -97
##  6 1      -97    -97
##  7 1      -99    -99
##  8 1      -97    -97
##  9 1      -99    -99
## 10 2      -97    -97
## 11 3      -97    -97
## 12 1      -97    -97
## 13 1      -97    -97
## 14 -99    -99    -99
## 15 -97    -97    -97
## 16 -97    -97    -97
## 17 -99    -99    -99
## 18 2      -97    -97
```

```r
ds[which((ds$qf10_2 == -97 | ds$qf10_2 == -99) & (ds$qf10_8 == -97 | ds$qf10_8 == -99)),c(59,60,65)]
```

```
## # A tibble: 8 x 3
##   qf10_2 qf10_3 qf10_8
##   <fct>  <fct>  <fct>
## 1 -99    4      -99
## 2 -99    -99    -99
## 3 -97    2      -97
## 4 -99    -99    -99
## 5 -97    -97    -97
```

```
## 6 -97    -97    -97
## 7 -99    -99    -99
## 8 -99    3      -99
```

```r
ds[which((ds$qf10_3 == -97 | ds$qf10_3 == -99) & (ds$qf10_8 == -97 | ds$qf10_8 == -99)),c(59,60,65)]
```

```
## # A tibble: 18 x 3
##    qf10_2 qf10_3 qf10_8
##    <fct>  <fct>  <fct>
##  1 2      -97    -99
##  2 1      -97    -97
##  3 1      -99    -99
##  4 -99    -99    -99
##  5 1      -97    -97
##  6 1      -97    -97
##  7 1      -99    -99
##  8 1      -97    -97
##  9 1      -99    -99
## 10 2      -97    -97
## 11 3      -97    -97
## 12 1      -97    -97
## 13 1      -97    -97
## 14 -99    -99    -99
## 15 -97    -97    -97
## 16 -97    -97    -97
## 17 -99    -99    -99
## 18 2      -97    -97
```

Delete observations with Don't Know (-99) and Not Answer (-97) for questions QF10_2, QF10_3, QF10_5, QF10_7, and QF10_8. Reverse scoring for selected risk-related questions to interpret higher scores as greater risk tolerance.

```r
ds_R <- ds[!(ds$qf10_2 %in% c(-99, -97) | ds$qf10_3 %in% c(-99, -97) | ds$qf10_5 %in% c(-99, -97)| ds$q

Attitude <- ds_R[1]

Attitude$qf10_2 <- ds_R$qf10_2
Attitude$qf10_3 <- ds_R$qf10_3
Attitude$qf10_5 <- ds_R$qf10_5

Attitude$qf10_7 <- ds_R$qf10_7
Attitude$qf10_8 <- ds_R$qf10_8

Attitude$qf10_7[ds_R$qf10_7 == 1] <- 5
Attitude$qf10_7[ds_R$qf10_7 == 2] <- 4
Attitude$qf10_7[ds_R$qf10_7 == 3] <- 3
Attitude$qf10_7[ds_R$qf10_7 == 4] <- 2
Attitude$qf10_7[ds_R$qf10_7 == 5] <- 1
```

```r
Attitude$qf10_2 <- ordered(Attitude$qf10_2, level=c(1:5))
Attitude$qf10_3 <- ordered(Attitude$qf10_3, level=c(1:5))
Attitude$qf10_5 <- ordered(Attitude$qf10_5, level=c(1:5))
Attitude$qf10_7 <- ordered(Attitude$qf10_5, level=c(1:5))
Attitude$qf10_8 <- ordered(Attitude$qf10_5, level=c(1:5))
```

```r
table(Attitude$qf10_2)
```

```
##
##   1   2   3   4   5
## 274 393 564 400 447
```

```r
table(Attitude$qf10_3)
```

```
##
##   1   2   3   4   5
## 132 354 720 451 421
```

```r
table(Attitude$qf10_5)
```

```
##
##    1    2    3    4    5
##   60  181  396  423 1018
```

```r
table(ds_R$qf10_2)
```

```
##
## -99 -97   1   2   3   4   5
##   0   0 274 393 564 400 447
```

```r
table(ds_R$qf10_3)
```

```
##
## -99 -97   1   2   3   4   5
##   0   0 132 354 720 451 421
```

3.1 Calculate Financial Attitude Score

Calculate an overall financial attitude score based on selected questions (QF10). The score is calculated as the average of the responses to questions QF10_2, QF10_3,, QF10_7, and QF10_8.

```r
Attitude$score <- round(unlist((as.numeric(as.character(Attitude$qf10_2)) + as.numeric(as.character(Att
Attitude$score <- ordered(Attitude$score, levels = c(1:5))
table(Attitude$score)
```

```
##
##    1    2    3    4    5
##   16  256  433 1124  249
```

Reduce financial attitude categories into three levels: - 1,2 low risk tolerance (1), - 3 neutral (2), - 4,5 high risk tolerance (3).

```r
Attitude$score2 <- Attitude$score
Attitude$score2[Attitude$score == 1 | Attitude$score == 2] <- 1
Attitude$score2[Attitude$score == 3] <- 2
Attitude$score2[Attitude$score == 4 | Attitude$score == 5] <- 3
Attitude$score2 <- ordered(Attitude$score2, levels=c(1:3))
```

4 Investment Attitudes

4.0.1 Full Ordinal Regression Model

Fit a full ordinal regression model to predict investment attitudes based on demographic variables.

```r
mod2_1<- polr(Attitude$score2 ~ Gender + Household + Aged + Education + Employment1  + AREA5 ,ds_R)
summary(mod2_1)
```

```
## Call:
## polr(formula = Attitude$score2 ~ Gender + Household + Aged +
##     Education + Employment1 + AREA5, data = ds_R)
##
## Coefficients:
##                 Value Std. Error t value
## Gender1     -0.23657    0.09461 -2.5005
## Household.L  0.09847    0.11712  0.8407
## Household.Q  0.06251    0.10154  0.6157
## Household.C  0.02121    0.09409  0.2255
## Aged.L       0.82358    0.11801  6.9791
## Aged.Q      -0.04557    0.10805 -0.4218
## Aged.C       0.11544    0.08855  1.3036
## Education.L  0.06454    0.09135  0.7065
## Education.Q -0.08170    0.07879 -1.0369
## Employment1 -0.04954    0.11181 -0.4431
## AREA52      -0.26165    0.13805 -1.8953
## AREA53      -0.09664    0.14241 -0.6786
## AREA54      -0.36010    0.13319 -2.7037
## AREA55      -0.19558    0.16725 -1.1693
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -2.2329   0.1285   -17.3735
## 2|3  -0.9685   0.1186    -8.1671
##
## Residual Deviance: 3519.526
## AIC: 3551.526
```

4.0.2 Variable Selection for Reduced Model

Perform ANOVA analysis of the full model to identify significant predictors.

```r
library(car)
Anova(mod2_1, type = "II", test.statistic = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Attitude$score2
##             LR Chisq Df Pr(>Chisq)
## Gender         6.263  1    0.01233 *
## Household      1.732  3    0.62993
## Aged          53.518  3  1.422e-11 ***
## Education      1.719  2    0.42329
## Employment1    0.196  1    0.65757
## AREA5          8.641  4    0.07072 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Perform stepwise selection to identify significant predictors for investment attitudes.

```r
step(mod2_1)
```

```
## Start:  AIC=3551.53
## Attitude$score2 ~ Gender + Household + Aged + Education + Employment1 +
##     AREA5
##
##               Df    AIC
## - Household    3 3547.3
## - Education    2 3549.2
## - Employment1  1 3549.7
## <none>           3551.5
## - AREA5        4 3552.2
## - Gender       1 3555.8
## - Aged         3 3599.0
##
## Step:  AIC=3547.26
## Attitude$score2 ~ Gender + Aged + Education + Employment1 + AREA5
##
##               Df    AIC
## - Education    2 3545.0
## - Employment1  1 3545.6
## <none>           3547.3
## - AREA5        4 3547.5
## - Gender       1 3551.1
## - Aged         3 3596.6
##
## Step:  AIC=3545.01
## Attitude$score2 ~ Gender + Aged + Employment1 + AREA5
##
##               Df    AIC
## - Employment1  1 3543.2
## <none>           3545.0
## - AREA5        4 3545.3
## - Gender       1 3549.1
## - Aged         3 3594.5
##
## Step:  AIC=3543.18
## Attitude$score2 ~ Gender + Aged + AREA5
##
##          Df    AIC
## <none>      3543.2
## - AREA5   4 3543.4
## - Gender  1 3548.0
## - Aged    3 3597.7

## Call:
## polr(formula = Attitude$score2 ~ Gender + Aged + AREA5, data = ds_R)
##
## Coefficients:
##     Gender1      Aged.L      Aged.Q      Aged.C      AREA52      AREA53
## -0.23986373  0.75986855 -0.04583058  0.11037070 -0.26919411 -0.08712781
##      AREA54      AREA55
## -0.33752607 -0.18939537
##
## Intercepts:
##       1|2       2|3
```

```
## -2.2159991 -0.9534354
##
## Residual Deviance: 3523.184
## AIC: 3543.184
```

4.0.3 Reduced Ordinal Regression Model

Fit a reduced ordinal regression model with selected variables.

```r
mod2_2 <- polr(Attitude$score2 ~ Gender + Aged + AREA5, ds_R)
summary_table <- coef(summary(mod2_2))
pval <- pnorm(abs(summary_table[, "t value"]),lower.tail = FALSE)* 2
summary_table <- cbind(summary_table, "p value" = round(pval,5))
summary_table
```

```
##                Value Std. Error     t value p value
## Gender1 -0.23986373 0.09220132  -2.6015217 0.00928
## Aged.L   0.75986855 0.09946387   7.6396438 0.00000
## Aged.Q  -0.04583058 0.09349029  -0.4902175 0.62398
## Aged.C   0.11037070 0.08793915   1.2550803 0.20945
## AREA52  -0.26919411 0.13768907  -1.9550869 0.05057
## AREA53  -0.08712781 0.14199756  -0.6135867 0.53949
## AREA54  -0.33752607 0.13165080  -2.5637981 0.01035
## AREA55  -0.18939537 0.16539259  -1.1451261 0.25216
## 1|2     -2.21599906 0.11642973 -19.0329313 0.00000
## 2|3     -0.95343540 0.10538181  -9.0474385 0.00000
```

```r
summary(mod2_2)
```

```
## Call:
## polr(formula = Attitude$score2 ~ Gender + Aged + AREA5, data = ds_R)
##
## Coefficients:
##             Value Std. Error t value
## Gender1 -0.23986    0.09220 -2.6015
## Aged.L   0.75987    0.09946  7.6396
## Aged.Q  -0.04583    0.09349 -0.4902
## Aged.C   0.11037    0.08794  1.2551
## AREA52  -0.26919    0.13769 -1.9551
## AREA53  -0.08713    0.14200 -0.6136
## AREA54  -0.33753    0.13165 -2.5638
## AREA55  -0.18940    0.16539 -1.1451
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -2.2160   0.1164   -19.0329
## 2|3  -0.9534   0.1054    -9.0474
##
## Residual Deviance: 3523.184
## AIC: 3543.184
```

```r
library(brant)
brant(mod2_2)
```

```
## --------------------------------------------
## Test for X2  df  probability
```

```
## --------------------------------------------
## Omnibus      7.39    8   0.5
## Gender1      0.09    1   0.76
## Aged.L       0.27    1   0.6
## Aged.Q       0.01    1   0.93
## Aged.C       0.4 1   0.52
## AREA52       1.82    1   0.18
## AREA53       0   1   0.98
## AREA54       0.09    1   0.76
## AREA55       2.91    1   0.09
## --------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

The assumption of parallel regression seems to be satisfied, indicating that the chosen ordinal regression model is appropriate.

4.0.4 Reduced Ordinal Regression Model 2

```
mod2_3 <- polr(Attitude$score2 ~ Gender + Aged, ds_R)
summary_table <- coef(summary(mod2_3))
pval <- pnorm(abs(summary_table[, "t value"]),lower.tail = FALSE)* 2
summary_table <- cbind(summary_table, "p value" = round(pval,5))
summary_table
```

```
##                Value Std. Error     t value p value
## Gender1 -0.23938687 0.09205849  -2.6003780 0.00931
## Aged.L   0.77411444 0.09891821   7.8258031 0.00000
## Aged.Q  -0.04855779 0.09334986  -0.5201699 0.60295
## Aged.C   0.12212490 0.08763230   1.3936059 0.16344
## 1|2     -2.03813261 0.08230421 -24.7634066 0.00000
## 2|3     -0.77942817 0.06723924 -11.5918643 0.00000
```

```
summary(mod2_3)
```

```
## Call:
## polr(formula = Attitude$score2 ~ Gender + Aged, data = ds_R)
##
## Coefficients:
##            Value Std. Error t value
## Gender1 -0.23939    0.09206 -2.6004
## Aged.L   0.77411    0.09892  7.8258
## Aged.Q  -0.04856    0.09335 -0.5202
## Aged.C   0.12212    0.08763  1.3936
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -2.0381   0.0823   -24.7634
## 2|3  -0.7794   0.0672   -11.5919
##
## Residual Deviance: 3531.361
## AIC: 3543.361
```

```
library(brant)
brant(mod2_3)
```

```
## ---------------------------------------------
## Test for X2  df  probability
## ---------------------------------------------
## Omnibus      0.94    4   0.92
## Gender1      0.1 1   0.76
## Aged.L       0.25    1   0.62
## Aged.Q       0.01    1   0.94
## Aged.C       0.48    1   0.49
## ---------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

The assumption of parallel regression seems to be satisfied, indicating that the chosen ordinal regression model is appropriate.

4.0.5 Compare Full and Reduced Models

Compare the full and reduced models using a likelihood ratio test.

```r
anova(mod2_1, mod2_2, test = "Chisq")
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: Attitude$score2
##                                                        Model Resid. df
## 1                                        Gender + Aged + AREA5      2068
## 2 Gender + Household + Aged + Education + Employment1 + AREA5      2062
##   Resid. Dev    Test    Df LR stat.  Pr(Chi)
## 1   3523.184
## 2   3519.526 1 vs 2     6 3.657147 0.722958
```

Anova test indicates that the reduced model (mod2_2) is not significantly different from the full model (mod2_1).

4.0.6 Compare Reduced Models

```r
anova(mod2_2, mod2_3, test = "Chisq")
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: Attitude$score2
##                     Model Resid. df Resid. Dev   Test    Df LR stat.    Pr(Chi)
## 1           Gender + Aged      2072   3531.361
## 2 Gender + Aged + AREA5      2068   3523.184 1 vs 2     4  8.17716 0.08529999
```

Likelihood ratio test indicates that the reduced model 2 (mod2_3) is the best fit for the data.

4.0.7 Compare Reduced Model and Null Model

```r
mod2_0 <- polr(Attitude$score2 ~ 1)
anova(mod2_3, mod2_0, test = "Chisq")
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: Attitude$score2
##             Model Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
## 1               1      2076   3602.366
```

```
## 2 Gender + Aged      2072   3531.361 1 vs 2     4 71.00498 1.387779e-14
```

The model mod2_3 is significantly better than the null model (mod2_0), indicating that the selected predictors explain a significant amount of variance in the financial attitude score.

5 Risk Attitudes

Analyze responses to QF10_5. Create a binary variable indicating risk tolerance (0 = low risk tolerance, 1 = high risk tolerance).

```r
table(Attitude$qf10_5)
```

```
##
##    1    2    3    4    5
##   60  181  396  423 1018
```

```r
Attitude$risk1 <- 1
Attitude$risk1[Attitude$qf10_5==5|Attitude$qf10_5==4] <- 0
Attitude$risk1 <- factor(Attitude$risk1)
table(Attitude$risk1)
```

```
##
##    0    1
## 1441  637
```

5.0.1 Full Logistic Regression Model for Risk Attitudes

Fit a logistic regression model to predict risk attitudes based on demographic variables.

```r
mod21_1 <- glm(Attitude$risk1 ~ Gender + Household + Aged + Education + Employment1  + AREA5, family =
summary(mod21_1)
```

```
##
## Call:
## glm(formula = Attitude$risk1 ~ Gender + Household + Aged + Education +
##     Employment1 + AREA5, family = "binomial", data = ds_R)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.06827    0.12354  -8.647  < 2e-16 ***
## Gender1      0.28921    0.09917   2.916  0.00354 **
## Household.L  0.10841    0.12606   0.860  0.38978
## Household.Q -0.13606    0.10826  -1.257  0.20884
## Household.C -0.03957    0.09809  -0.403  0.68661
## Aged.L      -0.46222    0.12192  -3.791  0.00015 ***
## Aged.Q       0.05620    0.11376   0.494  0.62128
## Aged.C      -0.11202    0.09260  -1.210  0.22641
## Education.L  0.17783    0.09481   1.876  0.06071 .
## Education.Q  0.10530    0.08233   1.279  0.20089
## Employment1  0.11099    0.11841   0.937  0.34860
## AREA52       0.13381    0.14263   0.938  0.34815
## AREA53      -0.04615    0.14822  -0.311  0.75554
## AREA54       0.20526    0.13906   1.476  0.13993
## AREA55      -0.06351    0.17559  -0.362  0.71759
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2561.4  on 2077  degrees of freedom
## Residual deviance: 2507.0  on 2063  degrees of freedom
## AIC: 2537
##
## Number of Fisher Scoring iterations: 4
```

5.0.2 Variable Selection

Variable Selection for Full Logistic Regression Model

```r
Anova(mod21_1, type = "II", test.statistic = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Attitude$risk1
##             LR Chisq Df Pr(>Chisq)
## Gender        8.5312  1  0.0034912 **
## Household     2.2520  3  0.5217697
## Aged         16.6795  3  0.0008225 ***
## Education     4.7659  2  0.0922763 .
## Employment1   0.8802  1  0.3481366
## AREA5         4.6298  4  0.3274367
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Perform stepwise selection to identify significant predictors for risk attitudes.

```r
step(mod21_1)
```

```
## Start:  AIC=2537.04
## Attitude$risk1 ~ Gender + Household + Aged + Education + Employment1 +
##     AREA5
##
##               Df Deviance    AIC
## - Household    3   2509.3 2533.3
## - AREA5        4   2511.7 2533.7
## - Employment1  1   2507.9 2535.9
## <none>             2507.0 2537.0
## - Education    2   2511.8 2537.8
## - Gender       1   2515.6 2543.6
## - Aged         3   2523.7 2547.7
##
## Step:  AIC=2533.3
## Attitude$risk1 ~ Gender + Aged + Education + Employment1 + AREA5
##
##               Df Deviance    AIC
## - AREA5        4   2514.1 2530.1
## - Employment1  1   2510.1 2532.1
## <none>             2509.3 2533.3
## - Education    2   2513.8 2533.8
## - Gender       1   2518.4 2540.4
## - Aged         3   2529.8 2547.8
##
## Step:  AIC=2530.05
```

```
## Attitude$risk1 ~ Gender + Aged + Education + Employment1
##
##               Df Deviance    AIC
## - Employment1  1   2514.6 2528.6
## <none>             2514.1 2530.1
## - Education    2   2518.6 2530.6
## - Gender       1   2523.2 2537.2
## - Aged         3   2535.8 2545.8
##
## Step:  AIC=2528.62
## Attitude$risk1 ~ Gender + Aged + Education
##
##             Df Deviance    AIC
## <none>           2514.6 2528.6
## - Education  2   2519.6 2529.6
## - Gender     1   2525.1 2537.1
## - Aged       3   2538.5 2546.5
```

```
##
## Call:  glm(formula = Attitude$risk1 ~ Gender + Aged + Education, family = "binomial",
##     data = ds_R)
##
## Coefficients:
## (Intercept)      Gender1       Aged.L       Aged.Q       Aged.C  Education.L
##   -0.943945     0.311591    -0.513127     0.005167    -0.134943     0.188134
## Education.Q
##    0.094150
##
## Degrees of Freedom: 2077 Total (i.e. Null);  2071 Residual
## Null Deviance:       2561
## Residual Deviance: 2515  AIC: 2529
```

5.0.3 Reduced Logistic Regression Model 1

```
mod21_2 <- glm(Attitude$risk1 ~  Gender + Aged + Education,family = "binomial", ds_R)
summary(mod21_2)
```

```
##
## Call:
## glm(formula = Attitude$risk1 ~ Gender + Aged + Education, family = "binomial",
##     data = ds_R)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.943945   0.070471 -13.395  < 2e-16 ***
## Gender1      0.311591   0.096537   3.228  0.00125 **
## Aged.L      -0.513127   0.108484  -4.730 2.25e-06 ***
## Aged.Q       0.005167   0.098364   0.053  0.95811
## Aged.C      -0.134943   0.091622  -1.473  0.14080
## Education.L  0.188134   0.093301   2.016  0.04375 *
## Education.Q  0.094150   0.081768   1.151  0.24955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2561.4  on 2077  degrees of freedom
## Residual deviance: 2514.6  on 2071  degrees of freedom
## AIC: 2528.6
##
## Number of Fisher Scoring iterations: 4
```

5.0.4 Reduced Logistic Regression Model 2

Further reduce the model by removing insignificant predictors.

```r
mod21_3 <- glm(Attitude$risk1 ~  Gender + Aged,family = "binomial", ds_R)
summary(mod21_3)
```

```
##
## Call:
## glm(formula = Attitude$risk1 ~ Gender + Aged, family = "binomial",
##     data = ds_R)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.96338    0.06993 -13.777  < 2e-16 ***
## Gender1      0.30200    0.09629   3.137  0.00171 **
## Aged.L      -0.56667    0.10295  -5.504  3.7e-08 ***
## Aged.Q       0.01099    0.09745   0.113  0.91025
## Aged.C      -0.13088    0.09147  -1.431  0.15249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2561.4  on 2077  degrees of freedom
## Residual deviance: 2519.6  on 2073  degrees of freedom
## AIC: 2529.6
##
## Number of Fisher Scoring iterations: 4
```

5.0.5 Compare Full and Reduced Models

Compare the full and reduced logistic regression models using a likelihood ratio test.

```r
anova(mod21_1, mod21_2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Attitude$risk1 ~ Gender + Household + Aged + Education + Employment1 +
##     AREA5
## Model 2: Attitude$risk1 ~ Gender + Aged + Education
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      2063     2507.0
## 2      2071     2514.6 -8  -7.5743   0.4761
```

Compare the reduced models to assess if further simplification is justified.

```r
anova(mod21_2, mod21_3, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Attitude$risk1 ~ Gender + Aged + Education
## Model 2: Attitude$risk1 ~ Gender + Aged
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      2071     2514.6
## 2      2073     2519.6 -2   -4.972  0.08324 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.0.6 Compare Reduced Model vs Null Model

Compare the final reduced model to a null model to evaluate its explanatory power.

```
mod21_0 <- glm(Attitude$risk1 ~ 1,family = "binomial", ds_R)
anova(mod21_3, mod21_0, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Attitude$risk1 ~ Gender + Aged
## Model 2: Attitude$risk1 ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2073     2519.6
## 2      2077     2561.4 -4  -41.786 1.848e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Selection of the final reduced model (mod21_3) is justified as it is significantly better than the null model (mod21_0).

6 Retirement Analysis

With the next models, we are going to tackle questions related to retirement savings (QF8 and QF9).

6.1 Retirement Planning

```
table(ds$qf8)
```

```
##
##  -99  -97    1    2    3    4    5    6
##  139   41   28   51  187  174   65 1691
```

There are 139 individuals that have not provided an answer for the question, we are going to create a subset that does not include these observations

```
dsR <- ds[!(ds$qf8 == -99),]
rknow <- know[!(ds$qf8 == -99),]
dsR$know <- rknow$tot
dsR <- dsR[!(dsR$qf8 == -97),]
dsR$qf8 <- ordered(dsR$qf8, levels = c(6:1))
```

6.1.1 Full Ordinal Regression Model for Retirement Planning

Fit a full ordinal regression model to predict retirement planning (QF8) based on demographic and knowledge variables.

```r
modRet1 <- polr(qf8 ~  Gender + Household + Age1 + Education + Employment1 + AREA5 + know, data = dsR, H
summary(modRet1)
```

```
## Call:
## polr(formula = qf8 ~ Gender + Household + Age1 + Education +
##     Employment1 + AREA5 + know, data = dsR, Hess = TRUE)
##
## Coefficients:
##                 Value Std. Error  t value
## Gender1     -0.110383   0.108020 -1.02188
## Household.L  0.061041   0.130575  0.46748
## Household.Q  0.007788   0.115295  0.06755
## Household.C  0.136809   0.106826  1.28067
## Age1         0.021300   0.004214  5.05434
## Education.L  0.543033   0.106850  5.08221
## Education.Q  0.004346   0.088412  0.04915
## Employment1  1.630993   0.129421 12.60226
## AREA52      -0.315639   0.147986 -2.13289
## AREA53      -0.523417   0.154691 -3.38364
## AREA54      -0.557556   0.153881 -3.62330
## AREA55      -0.735607   0.200856 -3.66235
## know.L       0.181945   0.224060  0.81204
## know.Q       0.387333   0.216948  1.78537
## know.C       0.098035   0.196615  0.49861
## know^4      -0.037885   0.167098 -0.22672
## know^5       0.031133   0.150799  0.20646
## know^6      -0.190698   0.144828 -1.31672
## know^7       0.051427   0.132707  0.38753
##
## Intercepts:
##     Value   Std. Error t value
## 6|5  2.7519  0.2827     9.7343
## 5|4  2.9530  0.2838    10.4058
## 4|3  3.6197  0.2882    12.5612
## 3|2  4.9956  0.3051    16.3723
## 2|1  6.0737  0.3411    17.8046
##
## Residual Deviance: 3482.744
## AIC: 3530.744
```

Significant predictors include Age1, Education, Employment1, and AREA5. Knowledge scores (know) were not significant.

6.1.2 Feature selection

Perform stepwise selection to identify significant predictors for retirement planning

```r
step(modRet1)
```

```
## Start:  AIC=3530.74
## qf8 ~ Gender + Household + Age1 + Education + Employment1 + AREA5 +
##     know
##
##               Df    AIC
## - know         7 3524.2
```

```
## - Household    3 3526.7
## - Gender       1 3529.8
## <none>           3530.7
## - AREA5        4 3546.4
## - Education    2 3552.7
## - Age1         1 3555.4
## - Employment1  1 3715.0
##
## Step:  AIC=3524.19
## qf8 ~ Gender + Household + Age1 + Education + Employment1 + AREA5
##
##               Df    AIC
## - Household    3 3520.7
## - Gender       1 3523.1
## <none>           3524.2
## - AREA5        4 3539.9
## - Age1         1 3548.5
## - Education    2 3549.7
## - Employment1  1 3708.3
##
## Step:  AIC=3520.65
## qf8 ~ Gender + Age1 + Education + Employment1 + AREA5
##
##               Df    AIC
## - Gender       1 3519.5
## <none>           3520.7
## - AREA5        4 3535.5
## - Education    2 3545.8
## - Age1         1 3547.8
## - Employment1  1 3703.4
##
## Step:  AIC=3519.49
## qf8 ~ Age1 + Education + Employment1 + AREA5
##
##               Df    AIC
## <none>           3519.5
## - AREA5        4 3534.7
## - Education    2 3545.9
## - Age1         1 3546.6
## - Employment1  1 3702.5

## Call:
## polr(formula = qf8 ~ Age1 + Education + Employment1 + AREA5,
##     data = dsR, Hess = TRUE)
##
## Coefficients:
##         Age1  Education.L  Education.Q  Employment1       AREA52       AREA53
##  0.021072744  0.569760901 -0.005606407  1.601270689 -0.329755020 -0.521411337
##       AREA54       AREA55
## -0.557530918 -0.711489483
##
## Intercepts:
##      6|5      5|4      4|3      3|2      2|1
## 2.802579 3.002348 3.665609 5.037067 6.113645
```

```
##
## Residual Deviance: 3493.489
## AIC: 3519.489
```

The final model includes Age1, Education, Employment1, and AREA5 as key predictors. Gender, Household, and Knowledge were excluded.

ANOVE test type II with Likelihood Ratio test (LRT): The ANOVA test evaluates the significance of each predictor in the context of the full model, comparing the deviance of the full model with that of a reduced model (without the predictor in question).

```
Anova(modRet1, type = "II", test.statistic = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: qf8
##             LR Chisq Df Pr(>Chisq)
## Gender         1.045  1     0.3066
## Household      1.931  3     0.5869
## Age1          26.669  1  2.415e-07 ***
## Education     25.910  2  2.364e-06 ***
## Employment1  186.271  1  < 2.2e-16 ***
## AREA5         23.646  4  9.405e-05 ***
## know           7.451  7     0.3835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.1.3 Reduced Model

Fit a reduced ordinal regression model using variables selected through AIC.

```
modRet2 <- polr(formula = qf8 ~ Age1 + Education + Employment1 + AREA5, data = dsR, Hess = TRUE)
summary(modRet2)
```

```
## Call:
## polr(formula = qf8 ~ Age1 + Education + Employment1 + AREA5,
##     data = dsR, Hess = TRUE)
##
## Coefficients:
##                 Value Std. Error  t value
## Age1         0.021073    0.00401  5.25566
## Education.L  0.569761    0.10356  5.50185
## Education.Q -0.005606    0.08785 -0.06381
## Employment1  1.601271    0.12734 12.57495
## AREA52      -0.329755    0.14709 -2.24183
## AREA53      -0.521411    0.15404 -3.38486
## AREA54      -0.557531    0.15225 -3.66194
## AREA55      -0.711489    0.19997 -3.55803
##
## Intercepts:
##     Value   Std. Error t value
## 6|5  2.8026  0.2682    10.4486
## 5|4  3.0023  0.2693    11.1477
## 4|3  3.6656  0.2739    13.3810
## 3|2  5.0371  0.2918    17.2598
## 2|1  6.1136  0.3294    18.5616
##
```

```
## Residual Deviance: 3493.489
## AIC: 3519.489
```

The reduced model confirms the significance of Age1, Education (linear term), Employment1, and AREA5 in predicting retirement planning.

We compute the p-values:

```
summary_table <- coef(summary(modRet2))
pval <- pnorm(abs(summary_table[, "t value"]),lower.tail = FALSE)* 2
summary_table <- cbind(summary_table, "p value" = round(pval,5))
summary_table
```

```
##                     Value Std. Error     t value p value
## Age1         0.021072744 0.00400953  5.25566414 0.00000
## Education.L  0.569760901 0.10355809  5.50184837 0.00000
## Education.Q -0.005606407 0.08785407 -0.06381499 0.94912
## Employment1  1.601270689 0.12733817 12.57494649 0.00000
## AREA52      -0.329755020 0.14709202 -2.24182810 0.02497
## AREA53      -0.521411337 0.15404225 -3.38485930 0.00071
## AREA54      -0.557530918 0.15225030 -3.66193651 0.00025
## AREA55      -0.711489483 0.19996739 -3.55802750 0.00037
## 6|5          2.802578711 0.26822578 10.44858079 0.00000
## 5|4          3.002348384 0.26932384 11.14772594 0.00000
## 4|3          3.665609384 0.27394208 13.38096477 0.00000
## 3|2          5.037066996 0.29183817 17.25979481 0.00000
## 2|1          6.113644878 0.32937135 18.56155637 0.00000
```

We are now going to estimate a regression only with knowledge score, in order to understand the association that this variable has with the answer qf8.

The p-values are computed here:

```
summary_table <- coef(summary(modRet2))
pval <- pnorm(abs(summary_table[, "t value"]),lower.tail = FALSE)* 2
summary_table <- cbind(summary_table, "p value" = round(pval,5))
summary_table
```

```
##                     Value Std. Error     t value p value
## Age1         0.021072744 0.00400953  5.25566414 0.00000
## Education.L  0.569760901 0.10355809  5.50184837 0.00000
## Education.Q -0.005606407 0.08785407 -0.06381499 0.94912
## Employment1  1.601270689 0.12733817 12.57494649 0.00000
## AREA52      -0.329755020 0.14709202 -2.24182810 0.02497
## AREA53      -0.521411337 0.15404225 -3.38485930 0.00071
## AREA54      -0.557530918 0.15225030 -3.66193651 0.00025
## AREA55      -0.711489483 0.19996739 -3.55802750 0.00037
## 6|5          2.802578711 0.26822578 10.44858079 0.00000
## 5|4          3.002348384 0.26932384 11.14772594 0.00000
## 4|3          3.665609384 0.27394208 13.38096477 0.00000
## 3|2          5.037066996 0.29183817 17.25979481 0.00000
## 2|1          6.113644878 0.32937135 18.56155637 0.00000
```

```
library(brant)
brant(modRet2)
```

```
## --------------------------------------------
## Test for X2  df  probability
```

```
## ---------------------------------------------
## Omnibus      52.49   32  0.01
## Age1     12.65   4   0.01
## Education.L  5.17    4   0.27
## Education.Q  5.74    4   0.22
## Employment1  3.03    4   0.55
## AREA52       6.08    4   0.19
## AREA53       1.62    4   0.81
## AREA54       3.02    4   0.56
## AREA55       7.93    4   0.09
## ---------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

```
## Warning in brant(modRet2): 4 combinations in table(dv,ivs) do not occur.
## Because of that, the test results might be invalid.
```

The assumption of parallel regression seems to be satisfied, indicating that the chosen ordinal regression model is appropriate.

6.1.4 Compare Full vs Reduced Models

Compare the full and reduced models using a likelihood ratio test.

```
anova(modRet1, modRet2, test = "Chisq")
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: qf8
##                                                                  Model Resid. df
## 1                              Age1 + Education + Employment1 + AREA5      2183
## 2 Gender + Household + Age1 + Education + Employment1 + AREA5 + know      2172
##   Resid. Dev   Test    Df LR stat.   Pr(Chi)
## 1   3493.489
## 2   3482.744 1 vs 2    11 10.74496 0.4648703
```

The test shows no significant difference between the full and reduced models ($p > 0.05$), indicating that the reduced model is sufficient.

6.1.5 Compare Reduced Model vs Null Model

Compare the reduced model to a null model to assess its explanatory power.

```
modRet0 <- polr(formula = qf8 ~ 1, data = dsR, Hess = TRUE)
anova(modRet2, modRet0, test = "Chisq")
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: qf8
##                                    Model Resid. df Resid. Dev   Test    Df
## 1                                      1      2191   3773.006
## 2 Age1 + Education + Employment1 + AREA5      2183   3493.489 1 vs 2     8
##   LR stat. Pr(Chi)
## 1
## 2  279.518       0
```

The reduced model significantly improves over the null model ($p < 0.001$), confirming its validity.

6.2 Retirement Tools

Classify answers to QF9 into secure (1) or unsecure (0) retirement plans. We classify answer a-f and i as a stable/secure retirement plan (1), while all the other answer are considered unsecure (0). We are interested in identifying those variables that are related to the choice of an unsecure retirement plan.

```
table(ds$qf9_99)
```

```
##
##    0    1
## 2028  348
```

```
dsR2 <- ds[!(ds$qf9_99==1),]
```

```
# We create a new column that contain the sum of the columns related to secure retirement plans
dsR2$sum <- as.numeric(as.character(dsR2$qf9_1)) + as.numeric(as.character(dsR2$qf9_2)) + as.numeric(as
# We transform the observation that have any value different from 0 in this new column to 1.
# In this way any observation that have at least one secure tool for building their
# retirement plan will be classified as 1.
# While all the other observation will remain equal to zero.
dsR2$sum[dsR2$sum != 0] <- 1
```

```
dsR2$sum <- factor(dsR2$sum, levels = c(0,1))
```

Secure retirement plans are classified as 1 (e.g., answers a-f and i), while unsecure plans are classified as 0.

6.2.1 Full Logistic Regression Model

Fit a full logistic regression model to predict secure retirement plan usage based on demographic variables and apply the Akaike Information Criterion

```
mod_qf9_1 <- glm(sum ~ Gender + Household + Aged + Education + Employment1  + AREA5, data = dsR2, famil
summary(mod_qf9_1)
```

```
##
## Call:
## glm(formula = sum ~ Gender + Household + Aged + Education + Employment1 +
##     AREA5, family = "binomial", data = dsR2)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.325781   0.140222   2.323 0.020162 *
## Gender1      1.042872   0.123469   8.446  < 2e-16 ***
## Household.L  0.003521   0.152352   0.023 0.981562
## Household.Q  0.086705   0.125326   0.692 0.489041
## Household.C -0.329777   0.116181  -2.838 0.004533 **
## Aged.L       0.454615   0.142647   3.187 0.001438 **
## Aged.Q      -0.061894   0.134200  -0.461 0.644649
## Aged.C      -0.134164   0.121991  -1.100 0.271427
## Education.L  0.753043   0.124929   6.028 1.66e-09 ***
## Education.Q  0.061536   0.106267   0.579 0.562542
## Employment1  1.867067   0.153211  12.186  < 2e-16 ***
## AREA52      -0.072236   0.175722  -0.411 0.681014
## AREA53      -0.044032   0.182972  -0.241 0.809826
## AREA54      -0.515992   0.168562  -3.061 0.002205 **
## AREA55      -0.723403   0.198015  -3.653 0.000259 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2304.7  on 2027  degrees of freedom
## Residual deviance: 1841.1  on 2013  degrees of freedom
## AIC: 1871.1
##
## Number of Fisher Scoring iterations: 5
```

Significant predictors include Gender, Household (C), Aged (linear term), Education (linear term), Employment1, AREA54, and AREA55.

6.2.2 Variable Selection

Perform stepwise selection using AIC to identify significant predictors for secure retirement plans.

```r
step(mod_qf9_1)
```

```
## Start:  AIC=1871.11
## sum ~ Gender + Household + Aged + Education + Employment1 + AREA5
##
##               Df Deviance    AIC
## <none>             1841.1 1871.1
## - Household    3   1849.7 1873.7
## - Aged         3   1852.7 1876.7
## - AREA5        4   1862.8 1884.8
## - Education    2   1881.3 1907.3
## - Gender       1   1916.1 1944.1
## - Employment1  1   2008.9 2036.9

##
## Call:  glm(formula = sum ~ Gender + Household + Aged + Education + Employment1 +
##     AREA5, family = "binomial", data = dsR2)
##
## Coefficients:
## (Intercept)      Gender1  Household.L  Household.Q  Household.C       Aged.L
##    0.325781     1.042872     0.003521     0.086705    -0.329777     0.454615
##      Aged.Q       Aged.C  Education.L  Education.Q  Employment1       AREA52
##   -0.061894    -0.134164     0.753043     0.061536     1.867067    -0.072236
##      AREA53       AREA54       AREA55
##   -0.044032    -0.515992    -0.723403
##
## Degrees of Freedom: 2027 Total (i.e. Null);  2013 Residual
## Null Deviance:      2305
## Residual Deviance: 1841  AIC: 1871
```

The final model includes Age1, Education (linear term), Employment1, and AREA5 as significant predictors.

The ANOVA test type II with Likelihood Ratio test (LRT) evaluates the significance of each predictor in the context of the full model, comparing the deviance of the full model with that of a reduced model (without the predictor in question).

```r
Anova(modRet1, type = "II", test.statistic = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: qf8
##             LR Chisq Df Pr(>Chisq)
```

```
## Gender          1.045  1     0.3066
## Household       1.931  3     0.5869
## Age1           26.669  1  2.415e-07 ***
## Education      25.910  2  2.364e-06 ***
## Employment1   186.271  1  < 2.2e-16 ***
## AREA5          23.646  4  9.405e-05 ***
## know            7.451  7     0.3835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.2.3 Reduced Logistic Regression Model

Now we re-estimate the model with the variables identified by the AIC.

```r
mod_qf9_2 <- glm(sum ~ Age1 + Education + Employment1 + AREA5, data = dsR2, family = "binomial")
summary(mod_qf9_2)
```

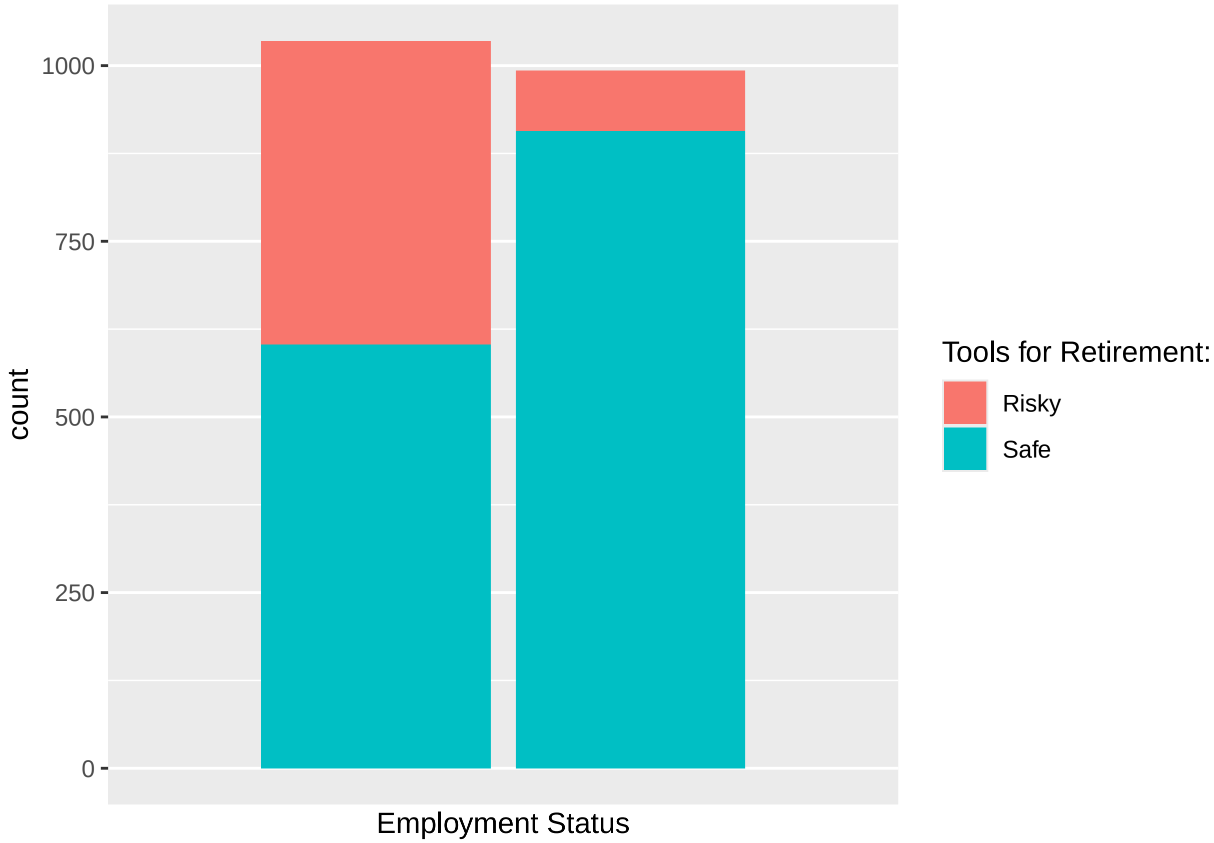
```
##
## Call:
## glm(formula = sum ~ Age1 + Education + Employment1 + AREA5, family = "binomial",
##     data = dsR2)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.212681   0.232129   0.916 0.359551
## Age1         0.009697   0.003406   2.847 0.004413 **
## Education.L  0.694413   0.121454   5.717 1.08e-08 ***
## Education.Q  0.055504   0.102496   0.542 0.588148
## Employment1  1.984408   0.136185  14.571  < 2e-16 ***
## AREA52      -0.094911   0.171253  -0.554 0.579432
## AREA53      -0.094313   0.178182  -0.529 0.596593
## AREA54      -0.495075   0.161748  -3.061 0.002208 **
## AREA55      -0.695197   0.191060  -3.639 0.000274 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2304.7  on 2027  degrees of freedom
## Residual deviance: 1931.0  on 2019  degrees of freedom
## AIC: 1949
##
## Number of Fisher Scoring iterations: 5
```

The reduced model confirms that Age1, Education (linear term), Employment1, AREA54, and AREA55 are significant predictors of secure retirement plan usage.

We are now going to build a stacked bar-plot to further investigate the relationship between employment status and the answer to QF9

```r
# Stacked
ggplot(dsR2, aes(fill=factor(sum, levels=c(0,1)), y = after_stat(count), x=Employment1)) +
    geom_bar(position="stack", stat="count") +
    xlab("Employment Status") +
#    legend("topleft", legend = c("Unsecure tools for retirement", "Secure tools for retirement"))
    scale_fill_discrete(labels=c('Risky', 'Safe')) +
```

```r
    guides(fill=guide_legend(title="Tools for Retirement:")) +
    scale_x_discrete(labels= c("Unemployed", "Employed"))
```



6.2.4 Compare Full vs Reduced Models

```r
anova(mod_qf9_1, mod_qf9_2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sum ~ Gender + Household + Aged + Education + Employment1 + AREA5
## Model 2: sum ~ Age1 + Education + Employment1 + AREA5
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2013     1841.1
## 2      2019     1931.0 -6  -89.924 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The very small p-value indicates that the difference in deviance between the two models is highly statistically significant. This means that the additional predictors included in Model 1 (mod_qf9_1) significantly improve the model's fit compared to Model 2 (mod_qf9_2).

7 Personal Finance

This section analyzes personal finance questions related to savings (QF3) and the ability to handle unexpected expenses (QF4).

7.1 Savings Behavior Analysis

We classify answer b, d, e as a secure way of saving money (1), while all the other answer are considered unsecure (0). We are interested in identifying those variables that are related to the choice of an unsecure plan for personal savings.

We remove the observation that have not given an answer for this question (155)

```
dsPF3 <- ds[!(ds$qf3_99==1),]
```

```
dsPF3$sum <- as.numeric(as.character(dsPF3$qf3_3)) + as.numeric(as.character(dsPF3$qf3_6)) + as.numeric
dsPF3$sum[dsPF3$sum != 0] <- 1
```

```
dsPF3$sum <- factor(dsPF3$sum, levels = c(0,1))
```

7.1.1 Full Logistic Regression Model for Savings Plans

Fit a full logistic regression model to predict secure savings plan usage based on demographic variables.

```
mod_PF3 <- glm(sum ~ Gender + Household + Aged + Education + Employment1  + AREA5, data = dsPF3, family
summary(mod_PF3)
```

```
##
## Call:
## glm(formula = sum ~ Gender + Household + Aged + Education + Employment1 +
##     AREA5, family = "binomial", data = dsPF3)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.33044    0.11106  -2.975 0.002928 **
## Gender1     -0.05311    0.09212  -0.577 0.564235
## Household.L  0.07295    0.11244   0.649 0.516489
## Household.Q -0.34503    0.09697  -3.558 0.000374 ***
## Household.C  0.10307    0.09000   1.145 0.252119
## Aged.L       0.72078    0.11309   6.373 1.85e-10 ***
## Aged.Q       0.29179    0.10739   2.717 0.006585 **
## Aged.C       0.21439    0.08733   2.455 0.014096 *
## Education.L  0.58151    0.09031   6.439 1.20e-10 ***
## Education.Q  0.04500    0.07700   0.584 0.558937
## Employment1  0.81945    0.11371   7.206 5.74e-13 ***
## AREA52      -0.03113    0.12985  -0.240 0.810512
## AREA53      -0.11800    0.13178  -0.895 0.370561
## AREA54      -0.49318    0.12988  -3.797 0.000146 ***
## AREA55      -0.88090    0.16973  -5.190 2.10e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3052.3  on 2220  degrees of freedom
## Residual deviance: 2846.4  on 2206  degrees of freedom
## AIC: 2876.4
##
## Number of Fisher Scoring iterations: 4
```

Significant predictors include Household (Q), Aged (Q and C), Education (L), Employment1, AREA54, and AREA55. Gender and Country were not significant.

7.1.2 Variable selections

Perform stepwise selection using AIC to identify significant predictors for unsecure savings plans.

```r
step(mod_PF3)
```

```
## Start:  AIC=2876.45
## sum ~ Gender + Household + Aged + Education + Employment1 + AREA5
##
##               Df Deviance    AIC
## - Gender       1   2846.8 2874.8
## <none>             2846.4 2876.4
## - Household    3   2859.9 2883.9
## - AREA5        4   2886.9 2908.9
## - Education    2   2888.8 2914.8
## - Aged         3   2895.1 2919.1
## - Employment1  1   2899.9 2927.9
##
## Step:  AIC=2874.78
## sum ~ Household + Aged + Education + Employment1 + AREA5
##
##               Df Deviance    AIC
## <none>             2846.8 2874.8
## - Household    3   2860.1 2882.1
## - AREA5        4   2887.5 2907.5
## - Education    2   2889.8 2913.8
## - Aged         3   2895.1 2917.1
## - Employment1  1   2900.8 2926.8

##
## Call:  glm(formula = sum ~ Household + Aged + Education + Employment1 +
##     AREA5, family = "binomial", data = dsPF3)
##
## Coefficients:
## (Intercept)  Household.L  Household.Q  Household.C       Aged.L       Aged.Q
##    -0.34819      0.06522     -0.34207      0.10261      0.71675      0.28365
##      Aged.C  Education.L  Education.Q  Employment1       AREA52       AREA53
##     0.21442      0.58507      0.04552      0.80611     -0.03087     -0.11814
##      AREA54       AREA55
##    -0.49453     -0.88402
##
## Degrees of Freedom: 2220 Total (i.e. Null);  2207 Residual
## Null Deviance:      3052
## Residual Deviance: 2847  AIC: 2875
```

```r
Anova(mod_PF3, type = "II", test.statistic = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: sum
##             LR Chisq Df Pr(>Chisq)
## Gender         0.333  1   0.564153
## Household     13.422  3   0.003808 **
## Aged          48.672  3  1.532e-10 ***
## Education     42.315  2  6.477e-10 ***
## Employment1   53.411  1  2.706e-13 ***
```

```
## AREA5         40.418  4  3.546e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The final model includes Household, Aged, Education, Employment1, and AREA5. Gender and Country were excluded.

7.1.3 Reduced Logistic Regression Model

Fit a reduced logistic regression model with selected variables.

```r
mod_PF3_1 <- glm(sum ~  Household + Aged + Education + Employment1 + AREA5, data = dsPF3, family = "bin
summary(mod_PF3_1)
```

```
##
## Call:
## glm(formula = sum ~ Household + Aged + Education + Employment1 +
##     AREA5, family = "binomial", data = dsPF3)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.34819    0.10671  -3.263 0.001102 **
## Household.L  0.06522    0.11162   0.584 0.558990
## Household.Q -0.34207    0.09682  -3.533 0.000411 ***
## Household.C  0.10261    0.08998   1.140 0.254175
## Aged.L       0.71675    0.11284   6.352 2.12e-10 ***
## Aged.Q       0.28365    0.10647   2.664 0.007718 **
## Aged.C       0.21442    0.08733   2.455 0.014079 *
## Education.L  0.58507    0.09009   6.494 8.36e-11 ***
## Education.Q  0.04552    0.07698   0.591 0.554293
## Employment1  0.80611    0.11130   7.243 4.40e-13 ***
## AREA52      -0.03087    0.12983  -0.238 0.812060
## AREA53      -0.11814    0.13175  -0.897 0.369864
## AREA54      -0.49453    0.12984  -3.809 0.000140 ***
## AREA55      -0.88402    0.16968  -5.210 1.89e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3052.3  on 2220  degrees of freedom
## Residual deviance: 2846.8  on 2207  degrees of freedom
## AIC: 2874.8
##
## Number of Fisher Scoring iterations: 4
```

The reduced model confirms that Household (Q), Aged (L, Q, C), Education (L), Employment1, AREA54, and AREA55 are significant predictors of unsecure savings plan usage.

7.1.4 Compare Full vs Reduced Models

Compare the full and reduced logistic regression models using a likelihood ratio test.

```r
anova(mod_PF3, mod_PF3_1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: sum ~ Gender + Household + Aged + Education + Employment1 + AREA5
## Model 2: sum ~ Household + Aged + Education + Employment1 + AREA5
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      2206     2846.4
## 2      2207     2846.8 -1 -0.33256   0.5642
```

The test shows no significant difference between the full and reduced models, indicating that the reduced model is sufficient.

7.1.5 Compare Reduced Model vs Null Model

Compare the reduced model to a null model to assess its explanatory power.

```r
mod_PF3_0 <- glm(sum ~  1, data = dsPF3, family = "binomial")
anova(mod_PF3_0, mod_PF3_1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sum ~ 1
## Model 2: sum ~ Household + Aged + Education + Employment1 + AREA5
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2220     3052.3
## 2      2207     2846.8 13   205.54 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The reduced model significantly improves over the null model ($p < 0.001$), confirming its validity.

7.2 Handling Unexpected Expenses

Prepare data for analysis by removing missing responses (-99) and those with no personal income (-98).

```r
dsPF4 <- ds[!(ds$qf4 == -99),]
dsPF4 <- dsPF4[!(dsPF4$qf4 == -98),]
```

Transform all answers other than "Yes" (1) into 0 (negative category) because "not knowing" is considered a negative response to the question

```r
dsPF4$qf4[dsPF4$qf4 != 1] <- 0
```

```r
dsPF4$qf4 <- factor(dsPF4$qf4, levels = c(0,1))
```

7.2.1 Full Logistic Regression Model

Fit a full logistic regression model to predict inability to handle an improvised expense based on demographic variables

```r
mod_PF4 <- glm(qf4 ~ Gender + Age1 + Education + Employment1 + AREA5 + Household, data = dsPF4, family =
summary(mod_PF4)
```

```
##
## Call:
## glm(formula = qf4 ~ Gender + Age1 + Education + Employment1 +
##     AREA5 + Household, family = "binomial", data = dsPF4)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.608321   0.235093  -6.841 7.85e-12 ***
```

```
## Gender1      0.116392   0.094795   1.228  0.21951
## Age1         0.033725   0.003563   9.465  < 2e-16 ***
## Education.L  0.620979   0.095427   6.507 7.65e-11 ***
## Education.Q -0.140914   0.080068  -1.760  0.07842 .
## Employment1  0.486052   0.106552   4.562 5.08e-06 ***
## AREA52      -0.057689   0.135522  -0.426  0.67034
## AREA53      -0.211151   0.137339  -1.537  0.12418
## AREA54      -0.398045   0.134973  -2.949  0.00319 **
## AREA55      -0.365559   0.168527  -2.169  0.03007 *
## Household.L -0.002975   0.113072  -0.026  0.97901
## Household.Q -0.212942   0.099273  -2.145  0.03195 *
## Household.C  0.058646   0.092649   0.633  0.52673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2809.1  on 2042  degrees of freedom
## Residual deviance: 2638.7  on 2030  degrees of freedom
## AIC: 2664.7
##
## Number of Fisher Scoring iterations: 4
```

Significant predictors include Age1, Education (L), Employment1, AREA54, and AREA55. Gender and Household were not significant.

7.2.2 Variable selection

```
step(mod_PF4)
```

```
## Start:  AIC=2664.65
## qf4 ~ Gender + Age1 + Education + Employment1 + AREA5 + Household
##
##               Df Deviance    AIC
## - Household    3   2643.9 2663.9
## - Gender       1   2640.2 2664.2
## <none>             2638.7 2664.7
## - AREA5        4   2650.4 2668.4
## - Employment1  1   2659.8 2683.8
## - Education    2   2689.6 2711.6
## - Age1         1   2734.8 2758.8
##
## Step:  AIC=2663.87
## qf4 ~ Gender + Age1 + Education + Employment1 + AREA5
##
##               Df Deviance    AIC
## - Gender       1   2645.6 2663.6
## <none>             2643.9 2663.9
## - AREA5        4   2657.1 2669.1
## - Employment1  1   2664.7 2682.7
## - Education    2   2694.1 2710.1
## - Age1         1   2759.8 2777.8
##
## Step:  AIC=2663.55
```

```
## qf4 ~ Age1 + Education + Employment1 + AREA5
##
##               Df Deviance    AIC
## <none>             2645.6 2663.6
## - AREA5        4   2658.4 2668.4
## - Employment1  1   2668.7 2684.7
## - Education    2   2694.7 2708.7
## - Age1         1   2762.2 2778.2
```

```
##
## Call:  glm(formula = qf4 ~ Age1 + Education + Employment1 + AREA5, family = "binomial",
##     data = dsPF4)
##
## Coefficients:
## (Intercept)         Age1  Education.L  Education.Q  Employment1       AREA52
##    -1.61238      0.03488      0.60308     -0.14542      0.50163     -0.04587
##      AREA53       AREA54       AREA55
##    -0.19800     -0.40747     -0.37052
##
## Degrees of Freedom: 2042 Total (i.e. Null);  2034 Residual
## Null Deviance:      2809
## Residual Deviance: 2646  AIC: 2664
```

```r
Anova(mod_PF4, type = "II", test.statistic = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: qf4
##             LR Chisq Df Pr(>Chisq)
## Gender         1.508  1    0.21943
## Age1          96.155  1  < 2.2e-16 ***
## Education     50.947  2  8.650e-12 ***
## Employment1   21.147  1  4.253e-06 ***
## AREA5         11.767  4    0.01917 *
## Household      5.218  3    0.15650
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The final model includes Age1, Education, Employment1, and AREA5. Gender and Household were excluded.

7.2.3 Reduced Logistic Regression Model

```r
mod_PF4_1 <- glm(qf4 ~ Age1 + Education + Employment1 + AREA5, family = "binomial", data = dsPF4)
summary(mod_PF4_1)
```

```
##
## Call:
## glm(formula = qf4 ~ Age1 + Education + Employment1 + AREA5, family = "binomial",
##     data = dsPF4)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.612379   0.223654  -7.209 5.63e-13 ***
## Age1         0.034884   0.003374  10.338  < 2e-16 ***
## Education.L  0.603079   0.094804   6.361 2.00e-10 ***
```

```
## Education.Q -0.145418   0.079751  -1.823  0.06824 .
## Employment1  0.501626   0.105090   4.773 1.81e-06 ***
## AREA52      -0.045868   0.135071  -0.340  0.73417
## AREA53      -0.197997   0.137067  -1.445  0.14859
## AREA54      -0.407466   0.133771  -3.046  0.00232 **
## AREA55      -0.370515   0.167400  -2.213  0.02687 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2809.1  on 2042  degrees of freedom
## Residual deviance: 2645.5  on 2034  degrees of freedom
## AIC: 2663.5
##
## Number of Fisher Scoring iterations: 4
```

The reduced model confirms that Age1, Education (L), Employment1, AREA54, and AREA55 are significant predictors of inability to handle an improvised expense.

7.2.4 Compare Full vs Reduced Models

```r
anova(mod_PF4, mod_PF4_1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: qf4 ~ Gender + Age1 + Education + Employment1 + AREA5 + Household
## Model 2: qf4 ~ Age1 + Education + Employment1 + AREA5
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      2030     2638.7
## 2      2034     2645.6 -4   -6.896   0.1415
```

The test shows no significant difference between the full and reduced models ($p > 0.05$), indicating that the reduced model is sufficient.

7.2.5 Compare Reduced Model vs Null Model

```r
mod_PF4_0 <- glm(qf4 ~ 1, family = "binomial", data = dsPF4)
anova(mod_PF4_1, mod_PF4_0, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: qf4 ~ Age1 + Education + Employment1 + AREA5
## Model 2: qf4 ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2034     2645.6
## 2      2042     2809.1 -8  -163.56 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The reduced model significantly improves over the null model ($p < 0.001$), confirming its validity.