Project.Rmd

Caggiano Paolo & Giardini Davide

2023-06-23

Import the data:

```
df <- read.csv("Dataset/ProjectData.csv")
```

Preliminary Operation to the Explanatory Variables

First of all, we are going to do some preliminary operation on the "instruction" variables. This is because the variable's levels are deeply unbalanced: class 7 ("No formal education") has only one observation, while class 6 ("Some primary school") has 25.

```
table(df$instruction)
```

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

This is going to be a problem for the classification model. For this reason, we decide to aggregate classes 5,6,7 in one class, that is going to represent people with a primary instruction or lower.

```
df$instruction[df$instruction == 6 | df$instruction == 7 ] <- 5
table(df$instruction)
```

```
##
##   1   3   4   5
## 537 925 717 197
```

Now, we are going to change the labels of the classes into 1-4 with 1 being the lower level of instruction and 4 being the highest. We are doing this because the regression function we are going to utilize later does not accept variables with missing levels (in this case 2).

```
df$instruction[df$instruction == 1] <- 2
df$instruction <-  df$instruction - 1
df$instruction <-  5- df$instruction
table(df$instruction)
```

```
##
##   1   2   3   4
## 197 717 925 537
```

Now we are going to convert all the variables into ordered and unordered categorical variables. Instruction, Household members and Knowledge Score have to be converted into ordered categorical variables:

```
df$instruction <- ordered(df$instruction, levels = c(1:4))
df$household_members <- ordered(df$household_members, levels = c(1:6))
df$know_score <- ordered(df$know_score, levels = c(0:7))
```

Employment status and area are instead converted to unordered categorical variables:

```
df$employment_status <- factor(df$employment_status, levels = c(1,2,4,5,6,9,10))
df$area <- factor(df$area, levels = c(1:5))
```

Literacy Model

With this first model we want to understand which are the socio-economic factors that help to explain financial literacy among people. In other words, we are going to build a model that tries to explain the "Knowledge Score" that we built in the preprocessing phase. To do so, we are going to use a Proportional Odds Logistic Regression Model.

```
modKnow <- polr(know_score ~ sex + area+ household_members + age + instruction + employment_status,
             data = df, Hess=TRUE)
step(modKnow)
```

```
## Start:  AIC=9525.66
## know_score ~ sex + area + household_members + age + instruction +
##     employment_status
##
##                     Df    AIC
## - area               4 9522.5
## - employment_status  6 9523.4
## <none>                 9525.7
## - household_members  5 9526.6
## - sex                1 9527.7
## - age                1 9533.6
## - instruction        3 9638.1
##
## Step:  AIC=9522.49
## know_score ~ sex + household_members + age + instruction + employment_status
##
##                     Df    AIC
## - employment_status  6 9521.5
## <none>                 9522.5
## - household_members  5 9523.2
## - sex                1 9524.1
## - age                1 9530.4
## - instruction        3 9633.9
##
## Step:  AIC=9521.46
## know_score ~ sex + household_members + age + instruction
##
##                     Df    AIC
```

```
## - household_members  5 9520.8
## <none>                9521.5
## - sex                1 9528.7
## - age                1 9532.2
## - instruction        3 9646.2
##
## Step:  AIC=9520.82
## know_score ~ sex + age + instruction
##
##               Df    AIC
## <none>           9520.8
## - sex          1 9529.4
## - age          1 9532.4
## - instruction  3 9643.6
```

```
## Call:
## polr(formula = know_score ~ sex + age + instruction, data = df,
##     Hess = TRUE)
##
## Coefficients:
##           sex           age instruction.L instruction.Q instruction.C
##   0.234367708   0.008678315   1.132914067  -0.118861983  -0.071318047
##
## Intercepts:
##        0|1        1|2        2|3        3|4        4|5        5|6        6|7
## -2.2529404 -1.0728016 -0.2465529  0.5588980  1.2594029  2.0400163  3.0004618
##
## Residual Deviance: 9496.823
## AIC: 9520.823
```

The Akaike Information Criterion suggests that we should remove the variables related to area, employment status and household members. We therefore re-estimate the model:

```
modKnow <- polr(formula = know_score ~ sex + age + instruction, data = df,
    Hess = TRUE)
summary(modKnow)
```

```
## Call:
## polr(formula = know_score ~ sex + age + instruction, data = df,
##     Hess = TRUE)
##
## Coefficients:
##                   Value Std. Error t value
## sex            0.234368   0.072177   3.247
## age            0.008678   0.002358   3.681
## instruction.L  1.132914   0.113029  10.023
## instruction.Q -0.118862   0.087604  -1.357
## instruction.C -0.071318   0.067188  -1.061
##
## Intercepts:
##     Value    Std. Error t value
## 0|1  -2.2529   0.1597   -14.1078
## 1|2  -1.0728   0.1440    -7.4505
```

```
## 2|3  -0.2466   0.1413    -1.7445
## 3|4   0.5589   0.1416     3.9461
## 4|5   1.2594   0.1432     8.7917
## 5|6   2.0400   0.1468    13.8953
## 6|7   3.0005   0.1556    19.2861
##
## Residual Deviance: 9496.823
## AIC: 9520.823
```

Since the polr function does not automatically give us the p-value, we are going to compute them separately:

```
summary_table <- coef(summary(modKnow))
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
## sex             0.234367708 0.072177249   3.247113 0.00117
## age             0.008678315 0.002357533   3.681100 0.00023
## instruction.L   1.132914067 0.113029033  10.023213 0.00000
## instruction.Q  -0.118861983 0.087604058  -1.356809 0.17484
## instruction.C  -0.071318047 0.067188411  -1.061464 0.28848
## 0|1            -2.252940402 0.159694721 -14.107795 0.00000
## 1|2            -1.072801592 0.143989905  -7.450533 0.00000
## 2|3            -0.246552854 0.141327956  -1.744544 0.08106
## 3|4             0.558897968 0.141633105   3.946097 0.00008
## 4|5             1.259402922 0.143249260   8.791689 0.00000
## 5|6             2.040016276 0.146813078  13.895331 0.00000
## 6|7             3.000461830 0.155576501  19.286086 0.00000
```

Since the regression is logistic, we have to compute the exponential transformation to fully interpret them:

```
exp(coef(summary(modKnow)))
```

```
##                    Value Std. Error      t value
## sex             1.2641092   1.074846 2.571600e+01
## age             1.0087161   1.002360 3.969004e+01
## instruction.L   3.1046906   1.119664 2.254374e+04
## instruction.Q   0.8879303   1.091556 2.574811e-01
## instruction.C   0.9311657   1.069497 3.459491e-01
## 0|1             0.1050898   1.173153 7.465561e-07
## 1|2             0.3420489   1.154872 5.811316e-04
## 2|3             0.7814901   1.151802 1.747246e-01
## 3|4             1.7487443   1.152154 5.173306e+01
## 4|5             3.5233172   1.154017 6.579334e+03
## 5|6             7.6907344   1.158137 1.083092e+06
## 6|7            20.0948152   1.168331 2.375969e+08
```

Retirement

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

QF8 - Who does not think to have a good retirement plan?

```
table(df$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:

```
dfR <- df[!(df$qf8==-99),]
dfR <- dfR[!(df$qf8==-97),]
dfR$qf8 <- ordered(dfR$qf8, levels = c(6:1))
```

Once again, we are going to use the "polr" function to estimate the Proportional Odds Logistic Regression Model:

```
modRet1 <- polr(qf8 ~ sex + area+ household_members + age + instruction + employment_status +
                 know_score, data = dfR, Hess = TRUE)
step(modRet1)
```

```
## Start:  AIC=3447.11
## qf8 ~ sex + area + household_members + age + instruction + employment_status +
##     know_score
##
##                     Df    AIC
## - household_members  5 3440.6
## - know_score         7 3440.6
## - sex                1 3445.3
## <none>                 3447.1
## - area               4 3462.0
## - age                1 3466.2
## - instruction        3 3469.0
## - employment_status  6 3639.4
##
## Step:  AIC=3440.57
## qf8 ~ sex + area + age + instruction + employment_status + know_score
##
##                     Df    AIC
## - know_score         7 3434.5
## - sex                1 3438.8
## <none>                 3440.6
## - area               4 3454.6
## - age                1 3461.6
## - instruction        3 3462.2
## - employment_status  6 3634.5
##
## Step:  AIC=3434.46
## qf8 ~ sex + area + age + instruction + employment_status
##
```

```
##                     Df    AIC
## - sex                1 3432.6
## <none>                 3434.5
## - area               4 3449.9
## - age                1 3456.5
## - instruction        3 3459.8
## - employment_status  6 3628.7
##
## Step:  AIC=3432.59
## qf8 ~ area + age + instruction + employment_status
##
##                     Df    AIC
## <none>                 3432.6
## - area               4 3448.4
## - age                1 3454.6
## - instruction        3 3458.9
## - employment_status  6 3627.3
```

```
## Call:
## polr(formula = qf8 ~ area + age + instruction + employment_status,
##     data = dfR, Hess = TRUE)
##
## Coefficients:
##               area2               area3               area4               area5
##         -0.32213894         -0.52795802         -0.59315839         -0.72721696
##                 age       instruction.L       instruction.Q       instruction.C
##          0.02589529          0.41610522          0.41103033         -0.20405685
##  employment_status2  employment_status4  employment_status5  employment_status6
##         -0.06323186         -1.23368482         -1.18560236         -2.15641481
##  employment_status9 employment_status10
##         -2.54452446         -0.95518520
##
## Intercepts:
##      6|5      5|4      4|3      3|2      2|1
## 1.385893 1.585304 2.258881 3.631325 4.759011
##
## Residual Deviance: 3394.593
## AIC: 3432.593
## (180 osservazioni eliminate a causa di valori mancanti)
```

We re-estimate the model utilizing only the variables that have been selected through the AIC criterion:

```r
modRet1 <- polr(formula = qf8 ~ area + age + instruction + employment_status,
    data = dfR, Hess = TRUE)
summary(modRet1)
```

```
## Call:
## polr(formula = qf8 ~ area + age + instruction + employment_status,
##     data = dfR, Hess = TRUE)
##
## Coefficients:
##                         Value Std. Error t value
## area2                -0.32214   0.150311 -2.1432
```

```
## area3                -0.52796   0.157252 -3.3574
## area4                -0.59316   0.155247 -3.8207
## area5                -0.72722   0.202052 -3.5992
## age                   0.02590   0.005326  4.8622
## instruction.L         0.41611   0.188226  2.2107
## instruction.Q         0.41103   0.147124  2.7938
## instruction.C        -0.20406   0.108160 -1.8866
## employment_status2   -0.06323   0.151389 -0.4177
## employment_status4   -1.23368   0.233036 -5.2940
## employment_status5   -1.18560   0.252152 -4.7019
## employment_status6   -2.15641   0.239795 -8.9927
## employment_status9   -2.54452   0.542458 -4.6907
## employment_status10 -0.95519   0.581149 -1.6436
##
## Intercepts:
##     Value   Std. Error t value
## 6|5  1.3859  0.3168     4.3753
## 5|4  1.5853  0.3173     4.9958
## 4|3  2.2589  0.3206     7.0464
## 3|2  3.6313  0.3357    10.8188
## 2|1  4.7590  0.3717    12.8043
##
## Residual Deviance: 3394.593
## AIC: 3432.593
## (180 osservazioni eliminate a causa di valori mancanti)
```

We compute the p-values:

```
summary_table <- coef(summary(modRet1))
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
## area2               -0.32213894 0.150310894 -2.1431510 0.03210
## area3               -0.52795802 0.157252486 -3.3573906 0.00079
## area4               -0.59315839 0.155246911 -3.8207420 0.00013
## area5               -0.72721696 0.202051715 -3.5991625 0.00032
## age                  0.02589529 0.005325833  4.8622046 0.00000
## instruction.L        0.41610522 0.188225628  2.2106725 0.02706
## instruction.Q        0.41103033 0.147124381  2.7937608 0.00521
## instruction.C       -0.20405685 0.108159767 -1.8866243 0.05921
## employment_status2  -0.06323186 0.151388814 -0.4176786 0.67618
## employment_status4  -1.23368482 0.233036322 -5.2939594 0.00000
## employment_status5  -1.18560236 0.252151644 -4.7019418 0.00000
## employment_status6  -2.15641481 0.239795202 -8.9927355 0.00000
## employment_status9  -2.54452446 0.542458147 -4.6907295 0.00000
## employment_status10 -0.95518520 0.581149267 -1.6436142 0.10026
## 6|5                  1.38589345 0.316755977  4.3752717 0.00001
## 5|4                  1.58530403 0.317325658  4.9958268 0.00000
## 4|3                  2.25888073 0.320571505  7.0464177 0.00000
## 3|2                  3.63132494 0.335650420 10.8187707 0.00000
## 2|1                  4.75901104 0.371673101 12.8042923 0.00000
```

We compute the exponential transformation of the estimates for better interpretability:

```
exp(coef(summary(modRet1)))
```

```
##                            Value Std. Error      t value
## area2                 0.72459751   1.162196 1.172847e-01
## area3                 0.58980812   1.170291 3.482602e-02
## area4                 0.55257926   1.167946 2.191154e-02
## area5                 0.48325203   1.223911 2.734662e-02
## age                   1.02623349   1.005340 1.293090e+02
## instruction.L         1.51604538   1.207106 9.121849e+00
## instruction.Q         1.50837110   1.158498 1.634236e+01
## instruction.C         0.81541602   1.114226 1.515826e-01
## employment_status2    0.93872579   1.163449 6.585739e-01
## employment_status4    0.29121751   1.262427 5.021838e-03
## employment_status5    0.30556207   1.286791 9.077633e-03
## employment_status6    0.11573933   1.270989 1.243096e-04
## employment_status9    0.07851038   1.720230 9.179986e-03
## employment_status10   0.38474089   1.788092 1.932802e-01
## 6|5                   3.99839668   1.372668 7.946142e+01
## 5|4                   4.88077504   1.373450 1.477951e+02
## 4|3                   9.57236913   1.377915 1.148736e+03
## 3|2                  37.76281713   1.398850 4.994965e+04
## 2|1                 116.63052586   1.450159 3.637755e+05
```

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:

```
modRet1 <- polr(qf8 ~ know_score, data = dfR, Hess = TRUE)
summary(modRet1)
```

```
## Call:
## polr(formula = qf8 ~ know_score, data = dfR, Hess = TRUE)
##
## Coefficients:
##                 Value Std. Error t value
## know_score.L  0.61590     0.1875  3.2839
## know_score.Q  0.17131     0.1800  0.9520
## know_score.C  0.05273     0.1713  0.3079
## know_score^4  0.08497     0.1590  0.5346
## know_score^5  0.18274     0.1500  1.2182
## know_score^6 -0.09368     0.1421 -0.6594
## know_score^7  0.16486     0.1296  1.2722
##
## Intercepts:
##     Value   Std. Error t value
## 6|5  1.2416  0.0574    21.6263
## 5|4  1.4170  0.0598    23.6852
## 4|3  2.0260  0.0713    28.4133
## 3|2  3.3349  0.1190    28.0152
## 2|1  4.4458  0.1991    22.3303
##
## Residual Deviance: 3671.106
```

```
## AIC: 3695.106
## (180 osservazioni eliminate a causa di valori mancanti)
```

Once again, the p-values are computed here:

```
summary_table <- coef(summary(modRet1))
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
## know_score.L  0.61589687 0.18754910  3.2839234 0.00102
## know_score.Q  0.17131129 0.17995014  0.9519931 0.34110
## know_score.C  0.05272759 0.17125239  0.3078940 0.75816
## know_score^4  0.08497346 0.15896153  0.5345536 0.59296
## know_score^5  0.18274276 0.15000668  1.2182309 0.22314
## know_score^6 -0.09368198 0.14206750 -0.6594188 0.50963
## know_score^7  0.16485878 0.12958420  1.2722135 0.20330
## 6|5           1.24163678 0.05741327 21.6263021 0.00000
## 5|4           1.41702553 0.05982758 23.6851538 0.00000
## 4|3           2.02595522 0.07130318 28.4132517 0.00000
## 3|2           3.33487475 0.11903786 28.0152439 0.00000
## 2|1           4.44576064 0.19909107 22.3302867 0.00000
```

QF9 - Who does utilize more secure tools for building their retirement fund?

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.

```
# We create a new subset without the observation that have not given an answer for this question
dfR2 <- df[!(df$qf9_99==1),]
```

```
# We create a new column that contain the sum of the columns related to secure retirement plans
dfR2$sum <- dfR2$qf9_1 + dfR2$qf9_2 + dfR2$qf9_3 +
  dfR2$qf9_4 + dfR2$qf9_5 + dfR2$qf9_6 + dfR2$qf9_9
# 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.
dfR2$sum[dfR2$sum != 0] <- 1
```

```
dfR2$sum <- factor(dfR2$sum, levels = c(0,1))
```

We now estimate the model and apply the Akaike Information Criterion:

```
modRet2 <- glm(sum ~ sex + area + household_members + age + instruction +
                 employment_status + know_score, data = dfR2, family = "binomial")
step(modRet2)
```

```
## Start:  AIC=1800.04
## sum ~ sex + area + household_members + age + instruction + employment_status +
##     know_score
##
##                     Df Deviance    AIC
## - know_score         7   1745.4 1791.4
## <none>                   1740.0 1800.0
## - household_members  5   1751.3 1801.3
## - area               4   1750.9 1802.9
## - age                1   1748.2 1806.2
## - instruction        5   1759.1 1809.1
## - sex                1   1751.7 1809.7
## - employment_status  6   1958.7 2006.7
##
## Step:  AIC=1791.42
## sum ~ sex + area + household_members + age + instruction + employment_status
##
##                     Df Deviance    AIC
## <none>                   1745.4 1791.4
## - household_members  5   1756.4 1792.4
## - area               4   1755.8 1793.8
## - age                1   1753.4 1797.4
## - sex                1   1758.0 1802.0
## - instruction        5   1766.8 1802.8
## - employment_status  6   1962.8 1996.8
```

```
##
## Call:  glm(formula = sum ~ sex + area + household_members + age + instruction +
##     employment_status, family = "binomial", data = dfR2)
##
## Coefficients:
##        (Intercept)                  sex                area2
##          -1.029877             0.478149             0.001433
##              area3                area4                area5
##           0.050283            -0.259401            -0.537479
## household_members.L  household_members.Q  household_members.C
##          -0.080300            -0.150613            -0.195031
## household_members^4  household_members^5                  age
##           0.152890            -0.334887             0.018428
##      instruction.L        instruction.Q        instruction.C
##           7.678353            -6.280116             4.750665
##      instruction^4        instruction^5   employment_status2
##          -2.466127             0.655084             0.581748
##  employment_status4   employment_status5   employment_status6
##          -2.135715            -1.213032            -1.441470
##  employment_status9  employment_status10
##          -0.799438             0.128523
##
## Degrees of Freedom: 2027 Total (i.e. Null);  2005 Residual
## Null Deviance:       2134
## Residual Deviance: 1745  AIC: 1791
```

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

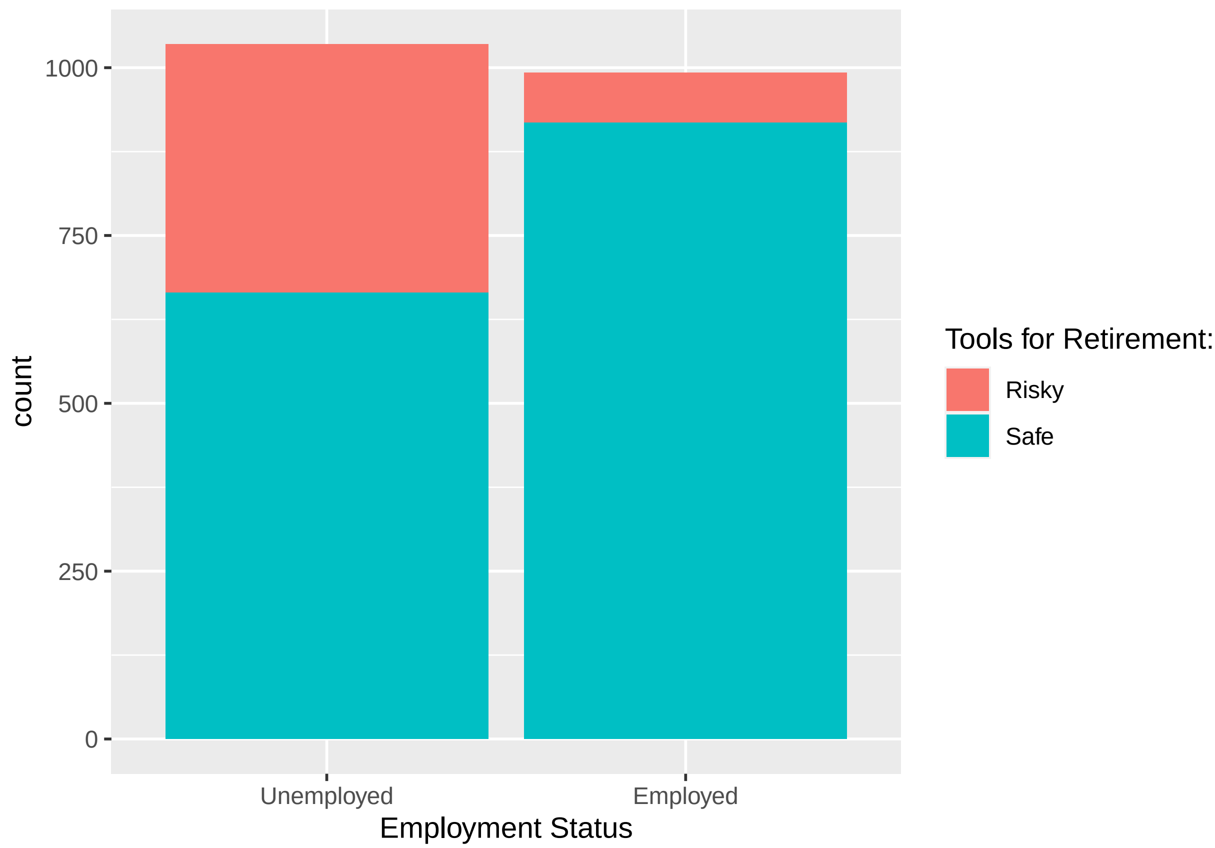
```r
modRet2 <- glm(sum ~ sex + area + household_members + age + instruction
               + employment_status, data = dfR2, family = "binomial")
summary(modRet2)
```

```
##
## Call:
## glm(formula = sum ~ sex + area + household_members + age + instruction +
##     employment_status, family = "binomial", data = dfR2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6025   0.2638   0.3903   0.7061   1.8254
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.029877  54.125583  -0.019 0.984819
## sex                   0.478149   0.135026   3.541 0.000398 ***
## area2                 0.001433   0.183776   0.008 0.993780
## area3                 0.050283   0.191075   0.263 0.792427
## area4                -0.259401   0.176311  -1.471 0.141218
## area5                -0.537479   0.204235  -2.632 0.008497 **
## household_members.L  -0.080300   0.279890  -0.287 0.774190
## household_members.Q  -0.150613   0.240866  -0.625 0.531775
## household_members.C  -0.195031   0.215308  -0.906 0.365031
## household_members^4   0.152890   0.183694   0.832 0.405235
## household_members^5  -0.334887   0.143063  -2.341 0.019240 *
## age                   0.018428   0.006532   2.821 0.004787 **
## instruction.L         7.678353 194.071741   0.040 0.968440
## instruction.Q        -6.280116 177.162433  -0.035 0.971722
## instruction.C         4.750665 121.025298   0.039 0.968688
## instruction^4        -2.466127  61.371601  -0.040 0.967947
## instruction^5         0.655084  20.458033   0.032 0.974455
## employment_status2    0.581748   0.267678   2.173 0.029757 *
## employment_status4   -2.135715   0.282426  -7.562 3.97e-14 ***
## employment_status5   -1.213032   0.285803  -4.244 2.19e-05 ***
## employment_status6   -1.441470   0.286290  -5.035 4.78e-07 ***
## employment_status9   -0.799438   0.358149  -2.232 0.025606 *
## employment_status10   0.128523   0.789847   0.163 0.870740
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2134.2  on 2027  degrees of freedom
## Residual deviance: 1745.4  on 2005  degrees of freedom
## AIC: 1791.4
##
## Number of Fisher Scoring iterations: 11
```

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

```
dfR2$employment_status <- as.integer(dfR2$employment_status)
dfR2$employment_status[dfR2$employment_status == 2] <- 1
dfR2$employment_status[dfR2$employment_status != 1] <- 0
library(ggplot2)
```

```
# Stacked
ggplot(dfR2, aes(fill=factor(sum, levels=c(0,1)), y = after_stat(count), x=factor(employment_status, le
    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')) +
    guides(fill=guide_legend(title="Tools for Retirement:")) +
    scale_x_discrete(labels= c("Unemployed", "Employed"))
```



Personal Finance

In this section we are going to tackle questions related to Personal Finance (savings): QF2, QF3, QF4, QF13

QF3 - Who uses non-smart ways to save money?

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.

```r
# We remove the observation that have not given an answer for this question (155)
dfPF2 <- df[!(df$qf3_99==1),]
```

```r
# The method is equal to the one in QF9
dfPF2$sum <- dfPF2$qf3_3 + dfPF2$qf3_6 + dfPF2$qf3_7
dfPF2$sum[dfPF2$sum != 0] <- 1
```

```r
dfPF2$sum <- factor(dfPF2$sum, levels = c(0,1))
```

```r
modPF2 <- glm(sum ~ sex + area + household_members + age + instruction +
                employment_status + know_score, data = dfPF2, family = "binomial")
step(modPF2)
```

```
## Start:  AIC=2850.21
## sum ~ sex + area + household_members + age + instruction + employment_status +
##     know_score
##
##                     Df Deviance    AIC
## - age                1   2790.7 2848.7
## - sex                1   2791.0 2849.0
## <none>                   2790.2 2850.2
## - household_members  5   2802.8 2852.8
## - know_score         7   2824.3 2870.3
## - instruction        5   2822.5 2872.5
## - area               4   2827.1 2879.1
## - employment_status  6   2868.8 2916.8
##
## Step:  AIC=2848.69
## sum ~ sex + area + household_members + instruction + employment_status +
##     know_score
##
##                     Df Deviance    AIC
## - sex                1   2791.4 2847.4
## <none>                   2790.7 2848.7
## - household_members  5   2803.6 2851.6
## - know_score         7   2825.7 2869.7
## - instruction        5   2822.5 2870.5
## - area               4   2827.8 2877.8
## - employment_status  6   2884.4 2930.4
##
## Step:  AIC=2847.44
## sum ~ area + household_members + instruction + employment_status +
##     know_score
##
##                     Df Deviance    AIC
## <none>                   2791.4 2847.4
## - household_members  5   2804.2 2850.2
## - know_score         7   2825.9 2867.9
## - instruction        5   2824.9 2870.9
## - area               4   2829.3 2877.3
## - employment_status  6   2884.7 2928.7
```

```
##
```

```
## Call:  glm(formula = sum ~ area + household_members + instruction +
##     employment_status + know_score, family = "binomial", data = dfPF2)
##
## Coefficients:
##         (Intercept)                area2                area3
##            -2.10899             -0.01589             -0.05959
##               area4                area5  household_members.L
##            -0.43400             -0.88052             -0.16456
## household_members.Q  household_members.C  household_members^4
##            -0.14047              0.36907              0.02578
## household_members^5        instruction.L        instruction.Q
##             0.13510              6.80998             -5.34902
##       instruction.C        instruction^4        instruction^5
##             4.05691             -1.96977              0.60474
##  employment_status2   employment_status4   employment_status5
##             0.46718             -0.31829             -0.62716
##  employment_status6   employment_status9  employment_status10
##             0.41359             -0.89740             -0.67605
##        know_score.L         know_score.Q         know_score.C
##             0.87293              0.03589              0.06862
##        know_score^4         know_score^5         know_score^6
##             0.07780             -0.13964              0.14476
##        know_score^7
##             0.04313
##
## Degrees of Freedom: 2220 Total (i.e. Null);  2193 Residual
## Null Deviance:      3052
## Residual Deviance: 2791  AIC: 2847
```

```
modPF2 <- glm(formula = sum ~ area + household_members + instruction +
                employment_status + know_score, family = "binomial", data = dfPF2)
summary(modPF2)
```

```
##
## Call:
## glm(formula = sum ~ area + household_members + instruction +
##     employment_status + know_score, family = "binomial", data = dfPF2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8242  -1.0451  -0.6344   1.1018   2.2533
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.10899   54.12427  -0.039 0.968918
## area2               -0.01589    0.13250  -0.120 0.904549
## area3               -0.05959    0.13431  -0.444 0.657294
## area4               -0.43400    0.13299  -3.263 0.001101 **
## area5               -0.88052    0.17218  -5.114 3.16e-07 ***
## household_members.L -0.16456    0.22434  -0.734 0.463248
## household_members.Q -0.14047    0.19748  -0.711 0.476879
## household_members.C  0.36907    0.17237   2.141 0.032260 *
## household_members^4  0.02578    0.14164   0.182 0.855547
## household_members^5  0.13510    0.10489   1.288 0.197740
```

```
## instruction.L         6.80998  194.07171   0.035 0.972008
## instruction.Q        -5.34902  177.16238  -0.030 0.975913
## instruction.C         4.05691  121.02519   0.034 0.973259
## instruction^4        -1.96977   61.37142  -0.032 0.974396
## instruction^5         0.60474   20.45780   0.030 0.976418
## employment_status2    0.46718    0.15382   3.037 0.002388 **
## employment_status4   -0.31829    0.20033  -1.589 0.112096
## employment_status5   -0.62716    0.20970  -2.991 0.002782 **
## employment_status6    0.41359    0.17393   2.378 0.017412 *
## employment_status9   -0.89740    0.23906  -3.754 0.000174 ***
## employment_status10  -0.67605    0.51512  -1.312 0.189378
## know_score.L          0.87293    0.16628   5.250 1.52e-07 ***
## know_score.Q          0.03589    0.15628   0.230 0.818380
## know_score.C          0.06862    0.14886   0.461 0.644836
## know_score^4          0.07780    0.14026   0.555 0.579119
## know_score^5         -0.13964    0.13245  -1.054 0.291755
## know_score^6          0.14476    0.12309   1.176 0.239583
## know_score^7          0.04313    0.11307   0.381 0.702866
## ---
## 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: 2791.4  on 2193  degrees of freedom
## AIC: 2847.4
##
## Number of Fisher Scoring iterations: 11
```

QF4 - Who is not capable of sustaining an imporvise expense?

```
# We remove the observation that have not given an answer for this question,
# and those who have not a personal income (78 + 255)
dfPF3 <- df[!(df$qf4 == -99),]
dfPF3 <- dfPF3[!(df$qf4 == -98),]
```

```
# We transform all the observation that have not answered with 1 ("Yes") as 0 (negative category)
# This is because "not knowing" is considered a negative response to the question
dfPF3$qf4[dfPF3$qf4 != 1] <- 0
```

```
dfPF3$qf4 <- factor(dfPF3$qf4, levels = c(0,1))
```

```
modPF3 <- glm(qf4 ~ sex + area + household_members + age + instruction +
                employment_status + know_score, data = dfPF3, family = "binomial")
step(modPF3)
```

```
## Start:  AIC=2456.49
## qf4 ~ sex + area + household_members + age + instruction + employment_status +
##     know_score
##
##                      Df Deviance    AIC
```

```
## - household_members  5   2400.7 2450.7
## - sex                1   2396.8 2454.8
## - area               4   2403.8 2455.8
## <none>                   2396.5 2456.5
## - age                1   2426.1 2484.1
## - instruction        5   2447.7 2497.7
## - know_score         7   2482.1 2528.1
## - employment_status  6   2497.8 2545.8
##
## Step:  AIC=2450.71
## qf4 ~ sex + area + age + instruction + employment_status + know_score
##
##                     Df Deviance    AIC
## - sex                1   2401.1 2449.1
## - area               4   2408.7 2450.7
## <none>                   2400.7 2450.7
## - age                1   2432.5 2480.5
## - instruction        5   2452.0 2492.0
## - know_score         7   2487.1 2523.1
## - employment_status  6   2505.1 2543.1
##
## Step:  AIC=2449.06
## qf4 ~ area + age + instruction + employment_status + know_score
##
##                     Df Deviance    AIC
## - area               4   2408.8 2448.8
## <none>                   2401.1 2449.1
## - age                1   2433.0 2479.0
## - instruction        5   2452.1 2490.1
## - know_score         7   2488.0 2522.0
## - employment_status  6   2509.0 2545.0
##
## Step:  AIC=2448.84
## qf4 ~ age + instruction + employment_status + know_score
##
##                     Df Deviance    AIC
## <none>                   2408.8 2448.8
## - age                1   2442.6 2480.6
## - instruction        5   2463.7 2493.7
## - know_score         7   2497.4 2523.4
## - employment_status  6   2527.5 2555.5
```

```
##
## Call:  glm(formula = qf4 ~ age + instruction + employment_status + know_score,
##     family = "binomial", data = dfPF3)
##
## Coefficients:
##         (Intercept)                  age        instruction.L
##            -3.51307              0.02857              7.32786
##       instruction.Q        instruction.C        instruction^4
##            -4.53731              2.29325             -0.88546
##       instruction^5   employment_status2   employment_status4
##             0.30394              0.17011             -0.72290
##  employment_status5   employment_status6   employment_status9
```

```
##            -1.25216             0.14588            -1.76989
## employment_status10         know_score.L         know_score.Q
##            -0.98187             1.57056             0.16562
##        know_score.C         know_score^4         know_score^5
##             0.19444             0.16591             0.09706
##        know_score^6         know_score^7
##             0.01251             0.11732
##
## Degrees of Freedom: 2050 Total (i.e. Null);  2031 Residual
##   (70 osservazioni eliminate a causa di valori mancanti)
## Null Deviance:       2843
## Residual Deviance: 2409  AIC: 2449
```

```r
modPF3 <- glm(qf4 ~ age + instruction + employment_status + know_score,
              family = "binomial", data = dfPF3)
summary(modPF3)
```

```
##
## Call:
## glm(formula = qf4 ~ age + instruction + employment_status + know_score,
##     family = "binomial", data = dfPF3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0563  -1.0283  -0.3066   0.9725   2.5838
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.513074  54.124786  -0.065 0.948248
## age                   0.028567   0.004976   5.741 9.44e-09 ***
## instruction.L         7.327863 194.071685   0.038 0.969880
## instruction.Q        -4.537314 177.162328  -0.026 0.979568
## instruction.C         2.293249 121.025382   0.019 0.984882
## instruction^4        -0.885461  61.371927  -0.014 0.988489
## instruction^5         0.303939  20.458288   0.015 0.988147
## employment_status2    0.170105   0.162235   1.049 0.294403
## employment_status4   -0.722901   0.206996  -3.492 0.000479 ***
## employment_status5   -1.252162   0.228417  -5.482 4.21e-08 ***
## employment_status6    0.145882   0.214351   0.681 0.496140
## employment_status9   -1.769886   0.325667  -5.435 5.49e-08 ***
## employment_status10  -0.981866   0.549038  -1.788 0.073721 .
## know_score.L          1.570564   0.184174   8.528  < 2e-16 ***
## know_score.Q          0.165622   0.173008   0.957 0.338410
## know_score.C          0.194443   0.164113   1.185 0.236092
## know_score^4          0.165905   0.151603   1.094 0.273804
## know_score^5          0.097058   0.142675   0.680 0.496333
## know_score^6          0.012506   0.133394   0.094 0.925307
## know_score^7          0.117323   0.120138   0.977 0.328781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2842.9  on 2050  degrees of freedom
```

```
## Residual deviance: 2408.8  on 2031  degrees of freedom
##   (70 osservazioni eliminate a causa di valori mancanti)
## AIC: 2448.8
##
## Number of Fisher Scoring iterations: 11
```

We plot the distribution of knowledge score between people that responded positively and negatively to answer QF4, in order to further investigate the relationship between this two variables:

```
library(tidyverse)
dfPF3 <- dfPF3 %>% drop_na(know_score)
dfPF3_0 <- dfPF3[dfPF3['qf4'] == 0,]
dfPF3_1 <- dfPF3[dfPF3['qf4'] == 1,]
```

```
require(gridExtra)
```

```
## Caricamento del pacchetto richiesto: gridExtra
```

```
##
## Caricamento pacchetto: 'gridExtra'
```

```
## Il seguente oggetto è mascherato da 'package:dplyr':
##
##     combine
```

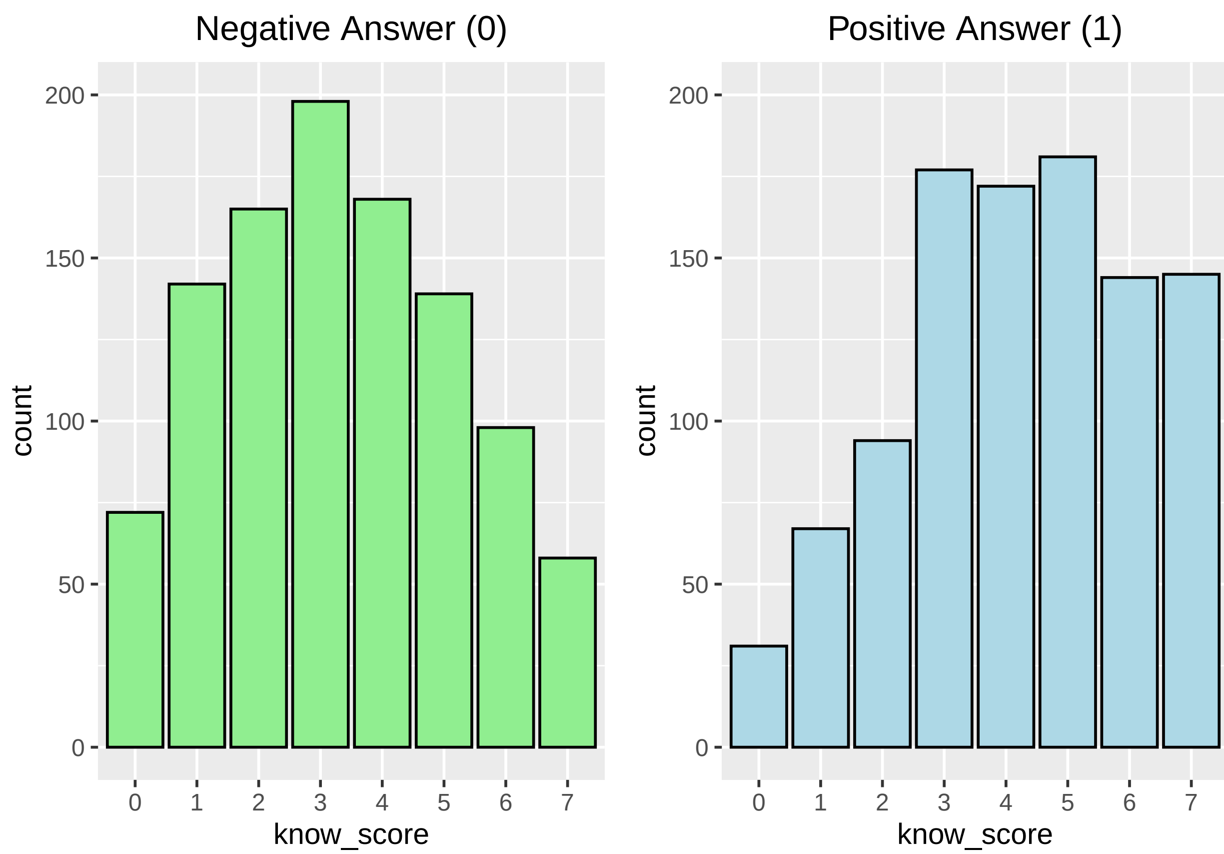
```
plot1 <- ggplot(dfPF3_0, aes(x=know_score)) +
    geom_histogram(binwidth=.5, colour="black", fill="light green", stat="count") +
    ggtitle("Negative Answer (0)")  +
    theme(plot.title = element_text(hjust = 0.5)) +
    ylim(c(0,200))
```

```
## Warning in geom_histogram(binwidth = 0.5, colour = "black", fill = "light
## green", : Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```

```
plot2 <- ggplot(dfPF3_1, aes(x=know_score)) +
    geom_histogram(binwidth=.5, colour="black", fill="light blue", stat="count") +
    ggtitle("Positive Answer (1)") +
    theme(plot.title = element_text(hjust = 0.5)) +
    ylim(c(0,200))
```

```
## Warning in geom_histogram(binwidth = 0.5, colour = "black", fill = "light
## blue", : Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```

```
grid.arrange(plot1, plot2, ncol=2)
```



QF13 - Who does not have an emergency fund?

This question evaluates if the person have an "Emergency Fund". According to the popular opinion, an emergency fund should cover at least 3-6 months of expenses. For this reason, answer d and e are considered positive (1), while all the other are considered negative (0).

```
# We remove the observation that have not given an answer for this question
dfPF4 <- df[!(df$qf13 == -99),]
```

```
dfPF4$qf13[dfPF4$qf13 == 1 | dfPF4$qf13 == 2 | dfPF4$qf13 == 3 | dfPF4$qf13 == -97] <- 0
dfPF4$qf13[dfPF4$qf13 != 0] <- 1
```

```
dfPF4$qf13 <- factor(dfPF4$qf13, levels = c(0,1))
```

```
modPF4 <- glm(qf13 ~ sex + area + household_members + age + instruction
              + employment_status + know_score, data = dfPF4, family = "binomial")
step(modPF4)
```

```
## Start:  AIC=2770.84
## qf13 ~ sex + area + household_members + age + instruction + employment_status +
##     know_score
##
##                     Df Deviance    AIC
```

```
## - household_members  5   2712.8 2762.8
## - sex                1   2710.9 2768.9
## - instruction        5   2719.6 2769.6
## <none>                   2710.8 2770.8
## - age                1   2717.8 2775.8
## - area               4   2725.3 2777.3
## - employment_status  6   2731.4 2779.4
## - know_score         7   2849.2 2895.2
##
## Step:  AIC=2762.82
## qf13 ~ sex + area + age + instruction + employment_status + know_score
##
##                     Df Deviance    AIC
## - sex                1   2712.8 2760.8
## - instruction        5   2721.8 2761.8
## <none>                   2712.8 2762.8
## - age                1   2719.6 2767.6
## - area               4   2727.4 2769.4
## - employment_status  6   2734.0 2772.0
## - know_score         7   2852.6 2888.6
##
## Step:  AIC=2760.84
## qf13 ~ area + age + instruction + employment_status + know_score
##
##                     Df Deviance    AIC
## - instruction        5   2721.9 2759.9
## <none>                   2712.8 2760.8
## - age                1   2719.7 2765.7
## - area               4   2727.5 2767.5
## - employment_status  6   2734.0 2770.0
## - know_score         7   2852.9 2886.9
##
## Step:  AIC=2759.94
## qf13 ~ area + age + employment_status + know_score
##
##                     Df Deviance    AIC
## <none>                   2721.9 2759.9
## - age                1   2726.8 2762.8
## - area               4   2738.5 2768.5
## - employment_status  6   2745.9 2771.9
## - know_score         7   2883.6 2907.6

##
## Call:  glm(formula = qf13 ~ area + age + employment_status + know_score,
##     family = "binomial", data = dfPF4)
##
## Coefficients:
##        (Intercept)                area2                area3
##           -0.65647             -0.22961             -0.11883
##              area4                area5                  age
##           -0.30883             -0.64594              0.01013
##  employment_status2   employment_status4   employment_status5
##           -0.08960             -0.37794             -0.83078
##  employment_status6   employment_status9  employment_status10
```

```
##            -0.26303            -0.45946            -0.86135
##        know_score.L        know_score.Q        know_score.C
##             2.02691            -0.19715             0.06382
##        know_score^4        know_score^5        know_score^6
##            -0.07980            -0.03036             0.05032
##        know_score^7
##             0.08991
##
## Degrees of Freedom: 2210 Total (i.e. Null);  2192 Residual
## Null Deviance:       2957
## Residual Deviance: 2722  AIC: 2760
```

```
modPF4 <- glm(formula = qf13 ~ area + age + employment_status + know_score,
    family = "binomial", data = dfPF4)
summary(modPF4)
```

```
##
## Call:
## glm(formula = qf13 ~ area + age + employment_status + know_score,
##     family = "binomial", data = dfPF4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6047  -0.9819  -0.6574   1.1346   2.4823
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.656472   0.276007  -2.378 0.017385 *
## area2               -0.229612   0.134471  -1.708 0.087725 .
## area3               -0.118833   0.136647  -0.870 0.384502
## area4               -0.308830   0.133787  -2.308 0.020978 *
## area5               -0.645936   0.170750  -3.783 0.000155 ***
## age                  0.010125   0.004618   2.192 0.028344 *
## employment_status2  -0.089602   0.155544  -0.576 0.564574
## employment_status4  -0.377937   0.199756  -1.892 0.058493 .
## employment_status5  -0.830778   0.219084  -3.792 0.000149 ***
## employment_status6  -0.263035   0.195594  -1.345 0.178690
## employment_status9  -0.459464   0.259111  -1.773 0.076191 .
## employment_status10 -0.861347   0.550172  -1.566 0.117444
## know_score.L         2.026913   0.190119  10.661  < 2e-16 ***
## know_score.Q        -0.197151   0.180620  -1.092 0.275044
## know_score.C         0.063816   0.167597   0.381 0.703372
## know_score^4        -0.079804   0.153837  -0.519 0.603931
## know_score^5        -0.030358   0.140385  -0.216 0.828792
## know_score^6         0.050319   0.126782   0.397 0.691447
## know_score^7         0.089910   0.114249   0.787 0.431301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2956.9  on 2210  degrees of freedom
## Residual deviance: 2721.9  on 2192  degrees of freedom
## AIC: 2759.9
```

```
##
## Number of Fisher Scoring iterations: 4
```