Practical 4

Eudald Romo Grau y Laura Santulario Verdú

March 15, 2018

1. Load the contingency table in the R environment.

Contingency table

```
setwd("C:\\Users\\TOSHIBA\\Documents\\GitHub\\practical1_multivariate_data_analysis\\practica _4")
N <- read.table("Mother_child.txt", header = TRUE, sep = " ", dec = ".", row.names = 1)
colnames(N) <- c("<18", "18-25", "25-30)", "[30-35)", ">35")
N
```

```
##         <18 18-25 25-30) [30-35) >35
## <1kg      2     0      3       3   2
## 1-2kg     3    27     18       6   0
## 2-3kg    42   105     63      18   6
## 3-3.5kg  15    84     12      27   3
## 3.5-4kg  15    57     24      12   3
## +4kg      2    12      6       3   6
```

2. (2p) Is there association between birth weight and age of the mother? Perform a chisquare test for independence between row and columns. Report the chi-square statistic, degrees of freedom, and p-value.

Chisquare test

```
chi <- chisq.test(N)
chi
```

```
##
##  Pearson's Chi-squared test
##
## data:  N
## X-squared = 84.029, df = 20, p-value = 8.046e-10
```

Applaying the independece test to our data:

H_0 : The mother's age and the weight of the child at birth are independent

H_1 : The mother's age and the weight of the child at birth are dependent

We have obtained a *p-value = 8.046e-10*, what means that the probability of rejecting H_0 being true is *8.046e-10* (nearly 0). So taking into account an α level equal to 0.05 (0.05 >> 8.046e-10) we can reject with enough statistical significance the null hypothesis. So there is some grade of association between the *age of the mum* and the *weight of the babies*.

In this case, the statistic that has been obtained is 84.029 and comes from a distribution χ^2 with 20 degrees of freedom. The df is the result of $(rownumber - 1) \times (colnumber - 1)$, in our specific case $5 \times 4 = 20$

3. (2p) Compute the correspondence matrix. Which category is the most frequent category? Which the less frequent?

Correspondance matrix

```
P <- round((1/sum(N))*N, 3)
P
```

```
##           <18 18-25 25-30) [30-35)   >35
## <1kg    0.003 0.000  0.005   0.005 0.003
## 1-2kg   0.005 0.047  0.031   0.010 0.000
## 2-3kg   0.073 0.181  0.109   0.031 0.010
## 3-3.5kg 0.026 0.145  0.021   0.047 0.005
## 3.5-4kg 0.026 0.098  0.041   0.021 0.005
## +4kg    0.003 0.021  0.010   0.005 0.010
```

```
P <- as.matrix(P)
P.most.freq <- which(P == max(P), arr.ind = TRUE)
P.less.freq <- which(P == min(P), arr.ind = TRUE)
```

The correspondence matrix is a probability matrix which displays the probabilities of all possible combinations of categories between the variables *woman's age* and *baby's weight* of the sample. From this matrix we can obtain the most and the less frequent categories. In our case the most frequent is that which the mothers are between 18 and 25 years and the babies have born with a weight between 2 and 3 kg (with a probability of 0.181). On the other hand, the less frequent categories, both with probability 0, are those where the mothers are between 18 and 25 years and have given birth babies with a weight less than 1 kg. The other category with probability 0 is that which corresponds with the mothers older than 35 years which babies have born with a weight between 1 and 2 kg.

4. Compute the expected counts under the assumption of independence. What is the sum of all expected counts?

Expected counts matrix

Expected matrix represents the frequency that would be expected in a cell if the variables were independent.

```
EXPECTED <- chi$expected
round(EXPECTED, 3)
```

```
##            <18   18-25 25-30) [30-35)   >35
## <1kg     1.364   4.922  2.176   1.192 0.345
## 1-2kg    7.368  26.580 11.751   6.435 1.865
## 2-3kg   31.927 115.181 50.922  27.886 8.083
## 3-3.5kg 19.238  69.404 30.684  16.803 4.870
## 3.5-4kg 15.145  54.637 24.155  13.228 3.834
## +4kg     3.957  14.275  6.311   3.456 1.002
```

```
sum.exp <- sum(EXPECTED)
```

When the expected matrix is computed the number of observations does not change, but only the distribution of the observations with the objective to ensure that the categories of the variable baby's weight are independent to the categories of the mother's age. So, as we were expecting, we have gotten 579 expected counts, the same as observed ones.

Comparing contingency table to the expected matrix it can be noticed that the higher difference for *baby's weight* is found in those who born with a weight between 2 and 3.5 Kg from women with ages between 18 and 35 years. And these cases are corresponded with the most commom context of developed countries (healthy babies born with a weight between 2.5 and 4 Kg and, usually, women plan to have descendents when they are around 20-35 years)

5. Compute for each cell of the contingency table its contribution to the chi-square statistic. Which cell(s) do contribute most to the chi-square statistic?

Cells - contribution matrix

The chi-square statistic is the sum of the contributions from each of the individual cells. These cell - contributions are given by the following formula:

$$\frac{(observed_{i,j} - expected_{i,j})}{\sqrt{expected_{i,j}}}$$

The next matrix represents the cell-contribution:

```
cell.contr <- round(((N - EXPECTED))/sqrt(EXPECTED), 3)
cell.contr <- as.matrix(cell.contr)
cell.contr
```

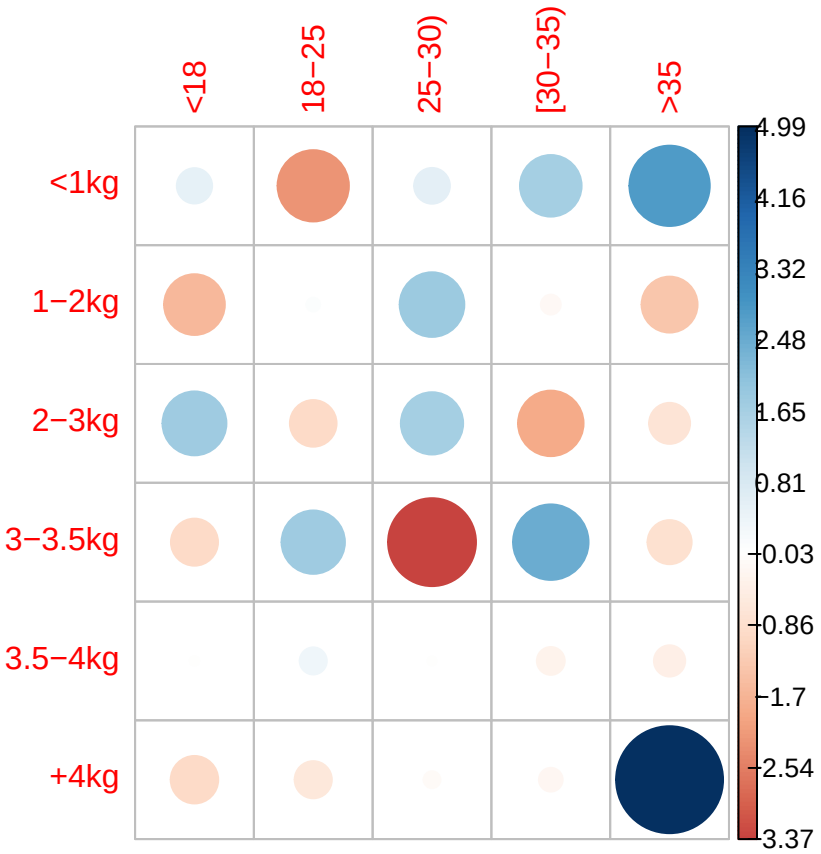
```
##            <18  18-25 25-30) [30-35)    >35
## <1kg     0.544 -2.219  0.558   1.656  2.815
## 1-2kg   -1.609  0.081  1.823  -0.172 -1.366
## 2-3kg    1.783 -0.949  1.693  -1.872 -0.733
## 3-3.5kg -0.966  1.752 -3.373   2.488 -0.848
## 3.5-4kg -0.037  0.320 -0.032  -0.338 -0.426
## +4kg    -0.984 -0.602 -0.124  -0.245  4.994
```

For representing these contributions we will use the comand **corrplot**.

```
which(cell.contr == max(cell.contr), arr.ind = TRUE)
```

```
##      row col
## +4kg   6   5
```

```
library(corrplot)
corrplot(cell.contr, is.cor = FALSE)
```



From this plot it can be said that the higher contributions are given by:

- Mothers who are older than 35 and have given birth babies heavier than 4 Kg. This is positive association, what means that for women elder than 35 years is more commom having overweight babies (it could be related with gestational diabetes, because being elder than 35 years is a risk factor for this kind of diabetes)
- Mothers between 25 and 30 years with babies that weigh between 3 and 3.5 Kg. In this case there is no association, what implies that is not common that women between 25 and 30 years have babies of 3 - 3.5 Kg (this could be related with the fact that women begin to be less fertile from 25 years and also because, usually, at these ages women are not completely compromise to change the way of life to another healthier)
- In a lesser extend mothers elder than 35 years that have delivered babies under weight or premature (less than 1 Kg). This association is positive, as in the first point, but in this case in a fewer grade (this could happend because from 35 years on the risk of having premature babies is higher)

6. Compute the row profies of the table. Compute also the weighted average of the row profiles, using the row masses as weights. Are the profiles homogeneous?

Row profiles

```
#Create a function that computes profiles and masses:
profile.func <- function(M, x){
```

```
  if (x == 1) {
    rows <- nrow(M)
    prof <- NULL
    mass <- rowSums(M)
    for (i in 1:rows) {
      prof <- rbind(prof, M[i,]/sum(M[i,]))
    }
  } else {
      cols <- ncol(M)
      prof <- NULL
      mass <- colSums(M)
      for (i in 1:cols) {
      prof <- cbind(prof, M[, i]/sum(M[, i]))
    }
  }

  solution <- list("prof" = round(prof, 3), "mass" = round(mass, 3))
  return(solution)
}

profile.func(P, 1)$prof
```

```
##        <18 18-25 25-30) [30-35)   >35
## [1,] 0.188 0.000  0.312   0.312 0.188
## [2,] 0.054 0.505  0.333   0.108 0.000
## [3,] 0.181 0.448  0.270   0.077 0.025
## [4,] 0.107 0.594  0.086   0.193 0.020
## [5,] 0.136 0.513  0.215   0.110 0.026
## [6,] 0.061 0.429  0.204   0.102 0.204
```

The matrix above expresses conditioned probabilities on the categories of the variable *baby's weight at birth* (we can realize that the sum of each row is 1). For instance, of the babies with 3 - 3.5 Kg at birth, most are children of women who are 18-25 years (59.4%), and the fewest were delivered by women older than 35 years (2.6%). The others are splitted in 8.6% are children of women with ages between 25-30, 10.7% belong to women younger than 18 and the spare children, 19.3% to women of 30-35 years.

Weighted average of the row profiles

In our case, the weighted average row profile represents the marginal distribution of the variable *mother's age at giving birth*. In this case, women between 18 and 25 years are the most likely to become pregnant (49.2%), while the women older than 35 the least common (3.3%).

```
row.average.prof <- round(as.vector(profile.func(P, 1)$mass)%*%profile.func(P, 1)$prof, 3)
row.average.prof
```

```
##        <18 18-25 25-30) [30-35)   >35
## [1,] 0.136 0.492  0.217   0.119 0.033
```

As we can check, weighted average row profile is the same as column masses.

```
profile.func(P, 2)$mass
```

```
##     <18   18-25  25-30) [30-35)     >35
##   0.136   0.492   0.217   0.119   0.033
```

From this, it can be said that our row profiles are not homogeneous, because if they were, the conditional probabilities, to the categories of *baby's weight*, would have the same distribution as the marginal distribution of *mother's age at giving birth*.

7. Compute the column profiles of the table. Compute also the weighted average of the column profiles, using the column masses as weights. Are the profiles homogeneous?

Column profiles

```
profile.func(P, 2)$prof
```

```
##          [,1]  [,2]  [,3]  [,4]  [,5]
## <1kg    0.022 0.000 0.023 0.042 0.091
## 1-2kg   0.037 0.096 0.143 0.084 0.000
## 2-3kg   0.537 0.368 0.502 0.261 0.303
## 3-3.5kg 0.191 0.295 0.097 0.395 0.152
## 3.5-4kg 0.191 0.199 0.189 0.176 0.152
## +4kg    0.022 0.043 0.046 0.042 0.303
```

The matrix above expresses conditioned probabilities on the categories of the variable *mother's age at giving birth* (we can realize that the sum of each column is 1)

Weighted average of the column profiles

In this study, the weighted average column profile represents the marginal distribution of the variable *baby's weight at birth*. In this case, the most common is having babies that weigh between 3 and 3.5 Kg (24.4%), while the most unusual is finding babies that born with less than 1 Kg (1.6%). These results make us think that this sample has been taken from a population of a developed country.

```
col.average.prof <- round(as.vector(profile.func(P, 2)$mass)%*%t(profile.func(P, 2)$prof), 3)
col.average.prof
```

```
##       <1kg 1-2kg 2-3kg 3-3.5kg 3.5-4kg  +4kg
## [1,] 0.016 0.093 0.404   0.244   0.191 0.049
```

As we can check, weighted average column profile is the same as row masses.

```
profile.func(P, 1)$mass
```

```
##    <1kg   1-2kg   2-3kg 3-3.5kg 3.5-4kg    +4kg
##   0.016   0.093   0.404   0.244   0.191   0.049
```

From this, it can be said that our column profiles are not homogeneous, because if they were, the conditional probabilities, to the categories of *mother's age at giving birth*, would have the same distribution as the marginal distribution of *baby's weight at birth*.

8. Compute the total intertia of the table.

Total inertia

Total inertia measures profile's dispersion with respect to the average (expected values for independence). The total inertia can be computed as:

$$\frac{\chi^2}{n}$$

```
tot.inercia <- round(chi$statistic/sum(N), 3)
```

In this study, total inertia is 0.145. As this value is not equal to 0 we cannot say that the two variables (*mother's age* and *baby's weight*) are independent, but it is true that the value of inertia is little (maximum inertia would be gotten at 4, $min(rownumber - 1, colnumber - 1)$) so we can deem that the association between these two categorical variables is weak, and, as we checked in the point **5.**, it is, mainly, induced by the babies from mothers older than 35 years (due to the condition of being older than 35 is a risk factor for the pregnancy and may cause problems like gestational diabetes or giving birth before due date)

9. Install the package ca. Perform correspondence analysis of the contingency table with the function ca. Use summary to obtain the numerical output of the corresponence analysis. How many dimensions does the solution of a correspondence analysis of this table have? How many dimensions are needed to obtain a good approximation to this table?

```
library(ca)
ca.data <- ca(N)
summary(ca.data)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.069210  47.7  47.7  ************
##  2      0.054206  37.4  85.0  *********
##  3      0.012328   8.5  93.5  **
##  4      0.009383   6.5 100.0  **
##         -------- -----
##  Total: 0.145127 100.0
##
##
## Rows:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |  1kg |   17  758  193 | 1105 754 305 |  -84   4   2 |
## 2 | 12kg |   93  326   93 | -147 149  29 |  160 177  44 |
## 3 | 23kg |  404  900  131 |  -10   2   1 |  205 898 314 |
## 4 | 335k |  244  994  265 | -151 144  80 | -367 850 604 |
## 5 | 354k |  192  721    5 |  -49 670   7 |   14  51   1 |
## 6 |  4kg |   50  922  314 |  894 881 579 | -194  42  35 |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |   18 |  136  310   95 |    7   0   0 |  177 309  78 |
## 2 | 1825 |  492  806  111 | -130 515 120 |  -98 291  87 |
## 3 | 2530 |  218  909  213 |   73  38  17 |  352 871 497 |
## 4 | 3035 |  119  742  150 |   15   1   0 | -368 741 298 |
## 5 |   35 |   35  990  430 | 1315 956 862 | -250  34  40 |
```

As we can see above, the command *summary(ca.data)* returns 3 tables:

- Principal inertias
- Row contributions
- Column contributions

The principal inertia table shows the decomposition of total inertia of a 6 x 5 contingency table into 4 components (axes). The total inertia explained by the four components is 0.145 (the same that has been computed in the previous point **8.**). Of the total inertia, the first component accounts for 47.7% of the inertia and the second component accounts for 37.4% of the inertia, so cumulative percentage of inertia of these two components account for 85% of the total inertia. Therefore, specifying 2 components for the analysis may be enough.

From the **row contributions** table we get some useful information:

- The column **mass** plays the role, as we have already seen in point **7.**, of marginal distribution of *baby's weight at birth*. So from this we can say that, from our sample, the probability that a child with appropriate weight is born is higher that a child born with underweight or overweight (because the higher probabilities are for the categories **"2 - 3Kg"** (40.4%), ___" 3 - 3.5Kg"___ (24.4%) and ___" 3.5 - 4Kg"___(19.2%))
- The column **qlt** represents how well define is the category by the 2 first components. In our case, the best representation for *baby's weight* categories are for **"3 - 3.5Kg"** (99.4%), **"> 4Kg"** (92.2%) and **"2 - 3Kg"** (90%), while the poorest is for **"1 - 2Kg"** (32.6%). From this we could say that the heavier the baby at birth the better is the representation by the two first components.
- The column **inr** represents how each category of the variable *baby's weight* contributes to the proportion of the total inertia. The **"> 4Kg"** category is the one who contributes the most (31.4%) to the computation of the χ^2 statistic, what means that this category is the one which deviates more from its expected value for the independence (the same as we obtained from point **5.**)
- The columns $\mathbf{k = 1}$ and $\mathbf{k = 2}$ represents the principal coordinates for the first two components.
- The column **cor** represents, for each component, the contribution of the component to the row's inertia. The most inertia showed by component 1 is given by the categories **"> 4Kg"** (88.1%) and **"< 1Kg"** (75.4%), so it could be said that the most inertia for component 1 is given by babies who born with underweight or overweight. The inercia that component 1 explains about **"2 - 3Kg"** category very a little (2%). On the other hand, the most inertia showed by component 2 is given by the categories **"2 - 3Kg"** (89.8%) and **"3 - 3.5Kg"** (85%), the other categories do not give a representative inertia for component 2.
- The column **ctr** represents the contribution of each row category to the inertia of each component (these values let us interpret the components). The categories **"> 4Kg"** (57.9%) and **"< 1Kg"** (30.5%) contribute, the most, to the inertia of component 1. Otherwise, the categories **"3 - 3.5kg"** (60.4%) and **"2 - 3kg"** (31.4%) contribute, the most, to the inertia of the component 2.

From the **column contributions** table we get some useful information:

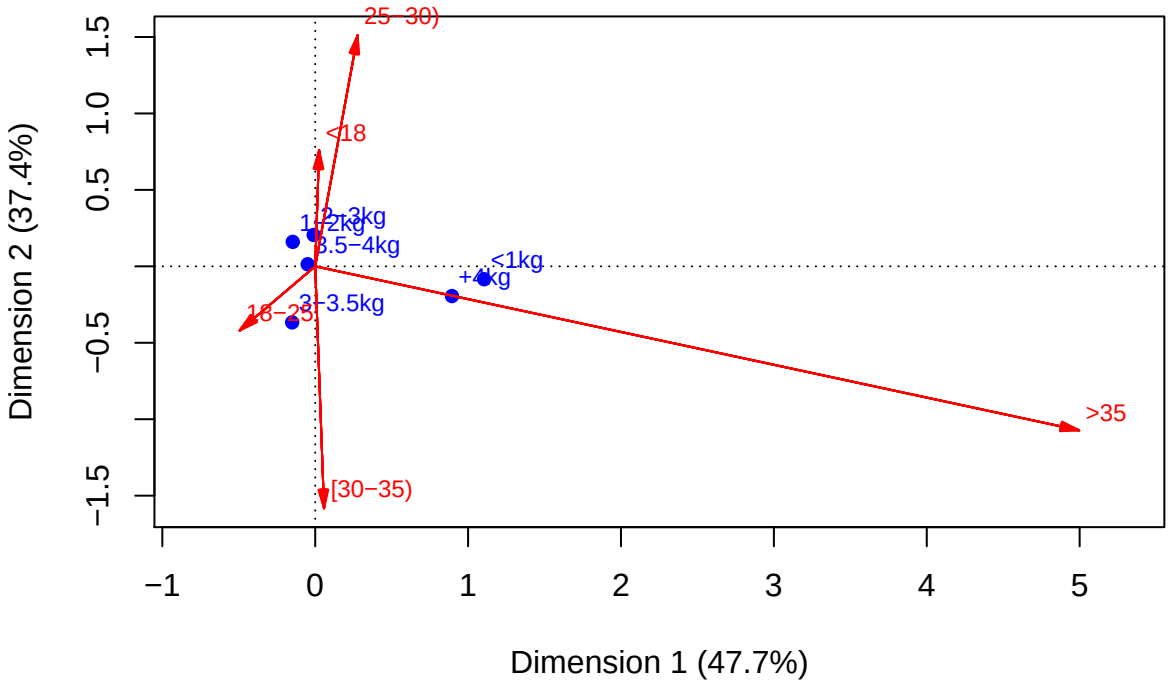
- The column **mass** plays the role, as we have already seen in point **6.**, of marginal distribution of *mother's age at giving birth*. So from this we can say that, from our sample, the highest probability for a woman of getting pregnant is at the range of ages 18-25, with probability 0.492. And the lowest is for women older than 35 years (0.035).

- The column **qlt** represents how well define is the category by the 2 first components. In our case, the best representation for *mother's age at giving birth* categories are for **"> 35"** (99%), **"25 - 30"** (90.9%), while the poorest is for **"< 18"** (31%). From this we could say that, except pregnant women younger than 18, all column categories are well represented by the two first components.
- The column **inr** represents how each category of the variable *mother's age at giving birth* contributes to the proportion of the total inertia. The **"> 35"** category is the one who contributes the most (43%) to the computation of the χ^2 statistic, what means that this category is the one which deviates more from its expected value for the independence (the same as we obtained from point **5.**)
- The columns $\mathbf{k = 1}$ and $\mathbf{k = 2}$ represents the principal coordinates for the first two components.
- The column **cor** represents, for each component, the contribution of the component to the column's inertia. The most inertia showed by component 1 is given by the category **"> 35"** (95.6%), so it could be said that the most inertia for component 1 is given by women who get pregnant after 35 years. The component 1 does not explain any inercia of the column category **"< 18"** and only a little of the category **"30 - 35"** (1%). On the other hand, the most inertia showed by component 2 is given by the categories **"25 - 30"** (87.1%) and **"30 - 35"** (74.1%).
- The column **ctr** represents the contribution of each column category to the inertia of each component (these values let us interpret the components). The only category which has a representative contribution to the inertia of the component 1 is **"> 35"** (86.2%). Otherwise, the categories **"25 - 30"** (49.7%) and **"30 - 35"** (29.8%) contribute, the most, to the inertia of the component 2.

10. Make a biplot of the row profiles by using *plot(out, map = "rowprincipal")*, where out contains the results of routine *ca*. Can you interpret the first dimension of the plot?

Biplot (row profiles)

```
plot(ca.data, arrow = c(FALSE, TRUE), map = "rowprincipal")
```



From this row profile biplot we can observe that the categories **"< 1Kg"** and **"> 4Kg"** of the variable *baby's weight at birth* are the further from the origin along the horizontal axis. This means that these two categories are the ones that most influence on the dispersion of the row profile with respect to the average (expected value for independence). Furthermore, the other categories are quite near from the origin (along the horizontal axis) but on opposite side of the origin. From this we can say that first dimension contrasts healthy babies at birth with problematic babies (overweight and underweight).

11. Which age category has the poorest quality of representation in the map?

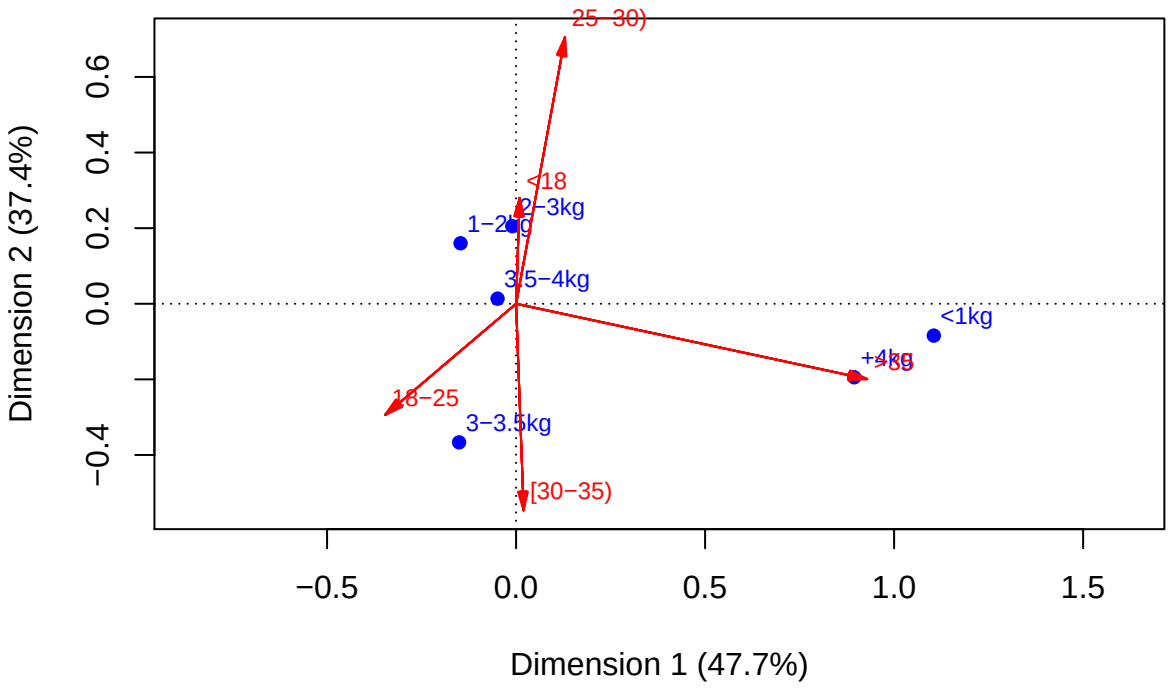
The category **"< 18"**, because is the closest to the two axes, and from the point **10.** is the category which has less value for both **qlt** and **inr**.

12. Which age and weight categories are the main contributors to the first dimension of the solution?

As we said in the point **9.** the *baby's weight* categories that are more influencial for the first dimension are **"< 1Kg"** and **"> 4Kg"**. And the age category that has more contribution to the first component is **"> 35 years"** (we can conclude this due to the category **"> 35 years"** of the variable *mother's age at giving birth* and the categories **"< 1Kg"** and **"> 4Kg"** are those which have longer distance to the origin along horizontal axis, and this implies that thess categories are the categories with more contribution on the first dimension)

13. Try the scaling option *map = "rowgreen"*. What changes in the biplot? Does it affect your interpretations?

```
plot(ca.data, arrow = c(FALSE, TRUE), map = "rowgreen")
```



Modifying the parameter *map* in the biplot by "rowgreen" let us identify more easily the most contributing points. As we can see the interpretation is the same as in **12.**, because the categories **"< 1Kg"** , **"> 4Kg"** and **"> 35 years"** are the furthest to the origin along the horizontal axis.