Aula 10 - Ficha de trabalho 8

Tiago A. Marques

November 2, 2018

Exercício 1

Determine os coeficientes de correlação de Pearson e de Spearman entre as variáveis R1 e R2 (DataTP8raios.csv). O que pode concluir quanto à relação entre as duas variáveis? Como poderá saber qual das duas abordagens será a adequada à situação em análise?

Read in the data

```
raios <- read.csv2("DataTP8raios.csv")
```

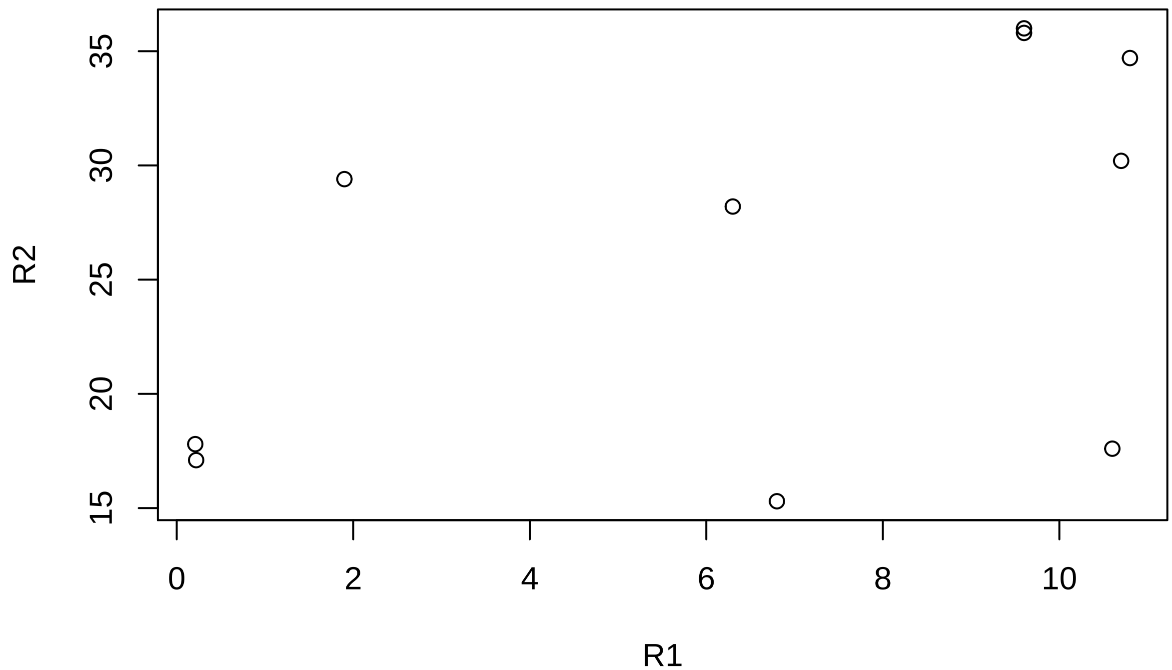
and as usual, check all is OK

```
str(raios)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ R1: num  0.21 1.9 0.22 10.7 6.8 10.6 9.6 6.3 10.8 9.6
##  $ R2: num  17.8 29.4 17.1 30.2 15.3 17.6 36 28.2 34.7 35.8
##  $ R3: num  17.6 28.9 17.2 30.9 15.3 17.6 36.2 29 35 36
```

We have 10 observations for 3 variables. Given that we were asked to look at `R1` and `R2` in particular, we look at these

```
with(raios,plot(R1,R2))
```



and now we calculate the correlations required. First the Pearson correlation

```
with(raios,cor(R1,R2,method="pearson"))
```

```
## [1] 0.5128447
```

and then the Spearman correlation

```
with(raios,cor(R1,R2,method="spearman"))
```

```
## [1] 0.4741663
```

To be fair, it is hard to choose between the two measures. The first uses the actual observations, the second the ranks, but then again, without knowing what the data are or aren't, it's hard choose.

Just as a curiosity, you can check that the Spearman correlation is really just the Pearson correlation applied to the ranks of the observations!

```
with(raios,cor(rank(R1),rank(R2),method="pearson"))
```

```
## [1] 0.4741663
```

Exercício 2

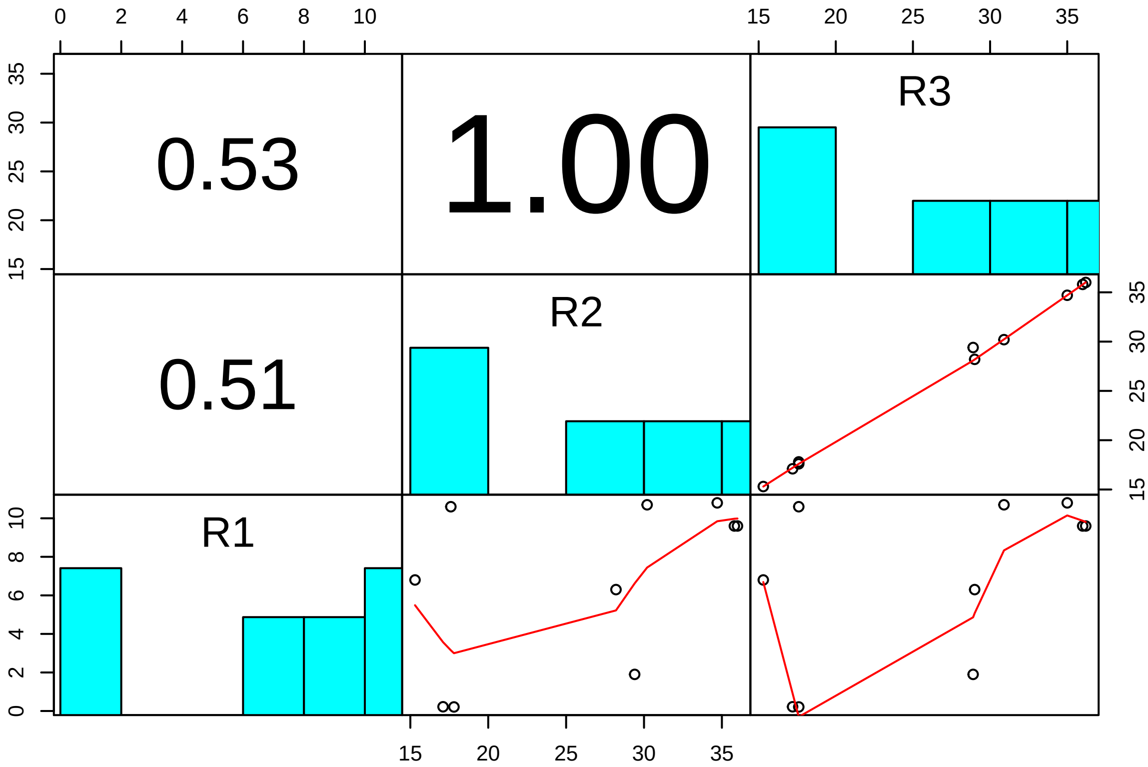
Determine os coeficientes de correlação múltipla (explicando R2) e o coeficiente de concordância de Kendall para todas as variáveis indicadas em DataTP8raios.csv. Efectue os respectivos testes (que têm por hipótese nula que não existe correlação) e indique qual a abordagem mais adequada.

We can calculate the pairwise correlations between all variables, and to do so we actually use the nice code in the example code for function `pairs` (note minor tweak to function `panel.cor` plot either Spearman or Pearson).

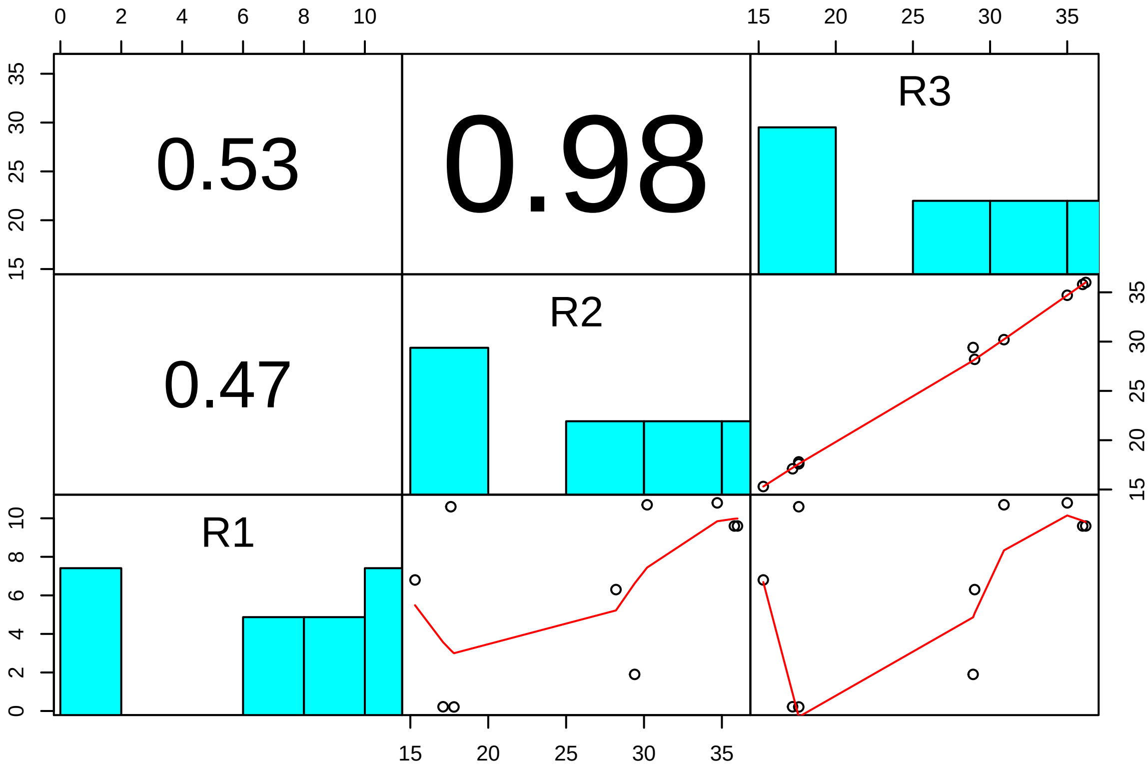
```
panel.hist <- function(x, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}
panel.corP <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}
panel.corS <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y,method="spearman"))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}
```

Pearson (top) and Spearman (bottom) correlation

```
pairs(raios, lower.panel = panel.smooth, upper.panel = panel.corP,
      gap=0, row1attop=FALSE,diag.panel = panel.hist)
```




```r
pairs(raios, lower.panel = panel.smooth, upper.panel = panel.corS,
      gap=0, row1attop=FALSE,diag.panel = panel.hist)
```



Nonetheless, these are not the multiple correlation that we are asked about!

We could compute Kendal's concordance coeficient and test it (using a permutation test) if it is different from 0 (i.e. no concordance, or correlation)

```r
library(vegan)
```

```
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.5-3
```

```r
kendall.global(raios)
```

```
## $Concordance_analysis
##                Group.1
## W         7.755240e-01
## F         6.909639e+00
## Prob.F    4.303572e-04
## Chi2      2.093915e+01
## Prob.perm 1.000000e-03
##
## attr(,"class")
## [1] "kendall.global"
```

We can see that we reject H0, so there is some concordance between the variables (in fact, it's the almost perfect correlation between `R2` and `R3` that induces the overall highly significant correlation.

Perhaps the most interesting part of the question is to acess if the multiple correlation is significant when we try to explain R2 as a function of the other two variables

```
lmR2=lm(R2~R1+R3,data=raios)
summary(lmR2)
```

```
##
## Call:
## lm(formula = R2 ~ R1 + R3, data = raios)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63753 -0.17934  0.08603  0.13961  0.48802
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.27612    0.39758   0.694    0.510
## R1          -0.03950    0.03316  -1.191    0.272
## R3           0.99346    0.01696  58.583 1.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3694 on 7 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9981
## F-statistic:  2330 on 2 and 7 DF,  p-value: 1.308e-10
```

```
summary(lmR2)$r.squared
```

```
## [1] 0.9984999
```

Since `R2` and `R3` are higly correlated, any model including them has a high Multiple R-Squared. Note that a formal test would be significant for the multiple correlation of `lmR2` (the F test statistic tests the multiple correlation). Notice that the value of the multiple correaltion is given by the correlation between the prediction from the model and the observed values

```
cor(predict(lmR2),raios$R2)
```

```
## [1] 0.9992496
```

and the corresponding multiple R-squared is just

```
cor(predict(lmR2),raios$R2)^2
```

```
## [1] 0.9984999
```

Exercício 3

Imagine que pretende seleccionar variáveis para incluir num modelo de regressão linear. Partindo dos dados DataTP8texugo.csv, e com recurso a técnicas de análise exploratória, faça uma proposta de um conjunto de variáveis a incluir no modelo.

As usual, we begin by reading the data

```
texugo <- read.csv2("DataTP8texugo.csv")
```

and making sure all went fine

```
str(texugo)
```

```
## 'data.frame':	14 obs. of  15 variables:
##  $ Densidade          : num  51.4 72.1 53.2 83.2 57.4 66.5 98.3 74.8 92.2 97.9 ...
##  $ Percentagem.floresta: num  0.2 1.9 0.2 10.7 6.8 10.6 9.6 6.3 10.8 9.6 ...
##  $ Coelho             : num  17.8 29.4 17.2 30.2 15.3 17.6 35.6 28.2 34.7 35.8 ...
##  $ Raposa             : num  24.6 20.7 18.5 10.6 8.9 11.1 10.6 8.8 11.9 10.8 ...
##  $ Precipitação       : num  18.9 8 22.6 7.1 27.3 20.8 5.6 13.1 5.9 5.5 ...
##  $ Humidade           : num  2.2 0.5 1.2 4.7 2.7 1.5 0 1 2.4 0 ...
##  $ Declive            : num  0 0 2.5 0 0 1.9 0 0 0 0 ...
##  $ Insectos           : num  41.1 35.4 20.9 5.1 1.6 2 2.5 0.9 1.2 0 ...
##  $ Agricola           : num  0 0.7 0 1.8 0 0 1.5 1.7 6.5 5.1 ...
##  $ Ratos              : num  16.4 14.8 44.1 23.6 21.9 23.7 24 27.4 42 59.5 ...
##  $ Temperatura        : num  0.5 9.3 11.5 20.8 38.2 26.2 31.5 13.7 9.3 2.6 ...
##  $ Estradas           : num  19.9 8.5 22.8 7.2 28 21.6 5.7 13.8 6.5 5.7 ...
##  $ Prox.casas         : num  18.1 29.4 18.1 30.8 16.2 18.1 36.5 28.6 35.6 36.3 ...
##  $ Tipo.solo          : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ Agua               : num  0 0 0 1 0 0 1 0 1 1 ...
```

We have 14 observations for 15 variables. If we had our brain turned off. . . :) we would try a global model

```r
lmtexall=lm(Densidade~.,data=texugo)
summary(lmtexall)
```

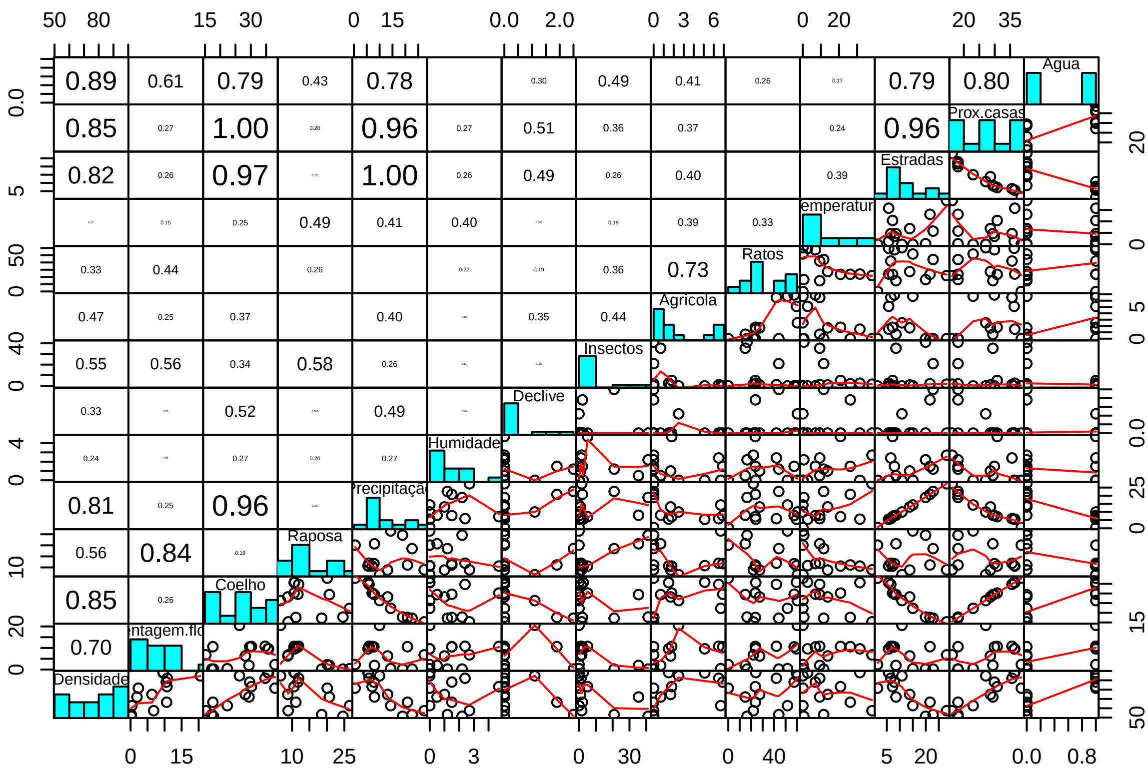
```
##
## Call:
## lm(formula = Densidade ~ ., data = texugo)
##
## Residuals:
## ALL 14 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (1 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -434.7363         NA      NA       NA
## Percentagem.floresta   20.8375         NA      NA       NA
## Coelho                -23.0347         NA      NA       NA
## Raposa                  7.3613         NA      NA       NA
## Precipitação          256.0860         NA      NA       NA
## Humidade              -12.5104         NA      NA       NA
## Declive               -10.8060         NA      NA       NA
## Insectos               -0.8929         NA      NA       NA
## Agricola               13.4181         NA      NA       NA
## Ratos                  -5.9867         NA      NA       NA
## Temperatura            -6.8469         NA      NA       NA
## Estradas             -233.4550         NA      NA       NA
## Prox.casas             37.7046         NA      NA       NA
## Tipo.solo                   NA         NA      NA       NA
## Agua                  -63.1427         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 13 and 0 DF,  p-value: NA
```

but that fails to fit. Why is that? Because we cannot fit a model with more variables than observations, in fact, there are often rules of thumb saying that you need at least X observations for each covariate you want to fit (where X varies by author, but say 10 seems reasonable). Therefore, with 15 variables, we would need at least 150 observations to fit a full model!

Therefore, we need to find a way top reduce the number of variables we include in our modelling exercise. The best way to start is to consider only variables that a priori might be useful to explain the response. In theory however, we only collect those, so we assume those in the table would a priori be potentially useful. Therefore, here we need some other way to reduce the number (well, in fact, one should never try to fit a multiple regression model to sucha a small number of observations, asusing the rule above,with less than 20 onverations you should not use more than 1 variable).

We look at all the pairwise correlations (no need to look at `Tipo.solo` as it is a factor covariate, so the correlation will be meaningless)

```
pairs(texugo[,-14], lower.panel = panel.smooth, upper.panel = panel.corP,
      gap=0, row1attop=FALSE,diag.panel = panel.hist)
```



The variables that seem morerelated to the response are `Coelho`, `Precipitação`, `Estradas`, `Prox.casas` and `Agua`.However, `Coelho` and `Precipitação` are higly correlated, so are `Estradas`, `Prox.casas` and `Precipitação`. Therefore, maybe using just `Precipitação` and `Agua` might be a good way to fit a parsimonious model.

We try that

```
lmtex=lm(Densidade~Precipitação+Agua,data=texugo)
summary(lmtex)
```

```
##
## Call:
## lm(formula = Densidade ~ Precipitação + Agua, data = texugo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -12.909  -5.254   1.103   6.339   9.086
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   74.2890     8.3078   8.942 2.23e-06 ***
## Precipitação  -0.6546     0.4364  -1.500  0.16175
## Agua          20.5473     6.5656   3.130  0.00959 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.62 on 11 degrees of freedom
## Multiple R-squared:  0.8209, Adjusted R-squared:  0.7884
## F-statistic: 25.21 on 2 and 11 DF,  p-value: 7.796e-05
```

and note that `Precipitação` is not relevant once `Água` is included in the model.

Exercício 4

Pretende saber se a proporção de sexos é idêntica em dois habitats distintos. Efectue um teste do qui-quadrado para avaliar esta hipótese, usando os dados DataTP8propsexos.csv.

Reading the data in

```
DataTP8propsexos <- read.csv("DataTP8propsexos.csv", sep=";")
```

The test is simple (note the test is only over the counts, present in the 2nd and 3rd columns).

```
mychi=chisq.test(DataTP8propsexos[,2:3])
mychi
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  DataTP8propsexos[, 2:3]
## X-squared = 0.24696, df = 1, p-value = 0.6192
```

At the usual significance levels, there is no reason to believe that the sex ratio is different across habitats.

The test statistic has a chi-squared distribution with 1 degree of freedom. Therefore, just for "fun", the P-value could be calculated manually given the value of the test statistic:

```
pchisq(mychi$statistic,1,lower.tail = FALSE)
```

```
## X-squared
## 0.6192236
```

Exercício 5

Efectue um teste G para avaliar se o consumo de determinadas presas é independente do género (macho ou fêmea) (DataTP8dieta.csv).

Read the data in

```
dieta <- read.csv("DataTP8dieta.csv", sep=";")
```

and we look at the data

```r
dieta
```

```
##   individuos presa.A presa.B presa.C
## 1     machos      41      25      12
## 2     femeas      50      27      87
```

The G test is in library `DescTools`, so we need to load it (and first install it if that has not been doen yet!). Then we implement the test

```r
library(DescTools)
GTest(dieta[,2:4],correct="none")
```

```
##
##  Log likelihood ratio (G-test) test of independence without
##  correction
##
## data:  dieta[, 2:4]
## G = 33.844, X-squared df = 2, p-value = 4.477e-08
```

concluding there are stong reasons to reject the null hypothesis that the prey consuption is gender independent

note that the same result would be obtained with a chi-suqared test. These two tests are assyntotically equivalent.

```r
mychidieta=chisq.test(dieta[,2:4])
```

We can look at the parcels building the test statistic

```r
mychidieta$expected
```

```
##       presa.A  presa.B  presa.C
## [1,] 29.33058 16.76033 31.90909
## [2,] 61.66942 35.23967 67.09091
```

```r
(mychidieta$expected-mychidieta$observed)^2/mychidieta$expected
```

```
##       presa.A  presa.B   presa.C
## [1,] 4.642779 4.050765 12.421911
## [2,] 2.208151 1.926583  5.907982
```

to conclude that males tend to ingest much less prey C than expected, unlike the females that show the opposite pattern.

Exercício 6

Efectue um teste do qui-quadrado para avaliar se a incidência de determinados grupos de parasitas é idêntica em juvenis e adultos de uma determinada espécie (DataTP8parasitas.csv).

Read in the data

```r
parasitas <- read.csv("DataTP8parasitas.csv", sep=";")
```

Look at the data

```r
parasitas
```

```
##   individuos Cestoda Monogenea Isopoda
## 1    juvenis      12        32      20
## 2    adultos      40        12      52
```

Implement the test

```
mychi2=chisq.test(parasitas[,2:4])
mychi2
```

```
##
##  Pearson's Chi-squared test
##
## data:  parasitas[, 2:4]
## X-squared = 30.601, df = 2, p-value = 2.265e-07
```

At the usual significance levels, there is strong reason to reject the H0 that the adults and juveniles might have the same amount of parasites.

Given the expected and observed values

```
mychi2$expected
```

```
##       Cestoda Monogenea  Isopoda
## [1,] 19.80952   16.7619 27.42857
## [2,] 32.19048   27.2381 44.57143
```

```
mychi2$observed
```

```
##      Cestoda Monogenea Isopoda
## [1,]      12        32      20
## [2,]      40        12      52
```

```
(mychi2$expected-mychi2$observed)^2/mychi2$expected
```

```
##       Cestoda Monogenea  Isopoda
## [1,] 3.078755 13.852814 2.011905
## [2,] 1.894618  8.524809 1.238095
```

we can see that there are far more `Monogea` in juveniles than what would be expected, and less in the adults that it would be expected, if both age classeshad similar infestation levels.

The test statistic has a chi-squared distribution with 2 degrees of freedom. Therefore, just for "fun", and as before the P-value could be calculated manually given the value of the test statistic:

```
pchisq(mychi2$statistic,2,lower.tail = FALSE)
```

```
##    X-squared
## 2.265053e-07
```