Assignment 2
Likelihood Ratio Test

Jan Izquierdo, Eloi Vilella

2023-10-25

Setting libraries

```
#install.packages("HardyWeinberg")
library("HardyWeinberg")
#install.packages("readxl")
library("readxl")
```

Exercise 1

Likelihood ratio test for Hardy-Weinberg equilibrium. In a genetic association study, the genotypes of a single nucleotide polymorphism have been determined for a sample of individuals. The genotype data file snp.txt contains the genotyping results. .

Setting up variables

```
db<-read.table("./Assignment_data/A2-LRT.txt")
n<-length(db$V1)
```

***1.a)*Load the data in the R environment, and make a table of the different genotypes. Report the table. What is the sample size of the study?** .

```
gen_table<-table(db$V1)
```

```
## Sample size is: 186
```

```
## The different genotypes are
```

```
##
##  AT  TT
##  55 131
```

***1.b)* How many alleles does this SNP have? How many genotypes could it theoretically have? Estimate all relative genotype frequencies by maximum likelihood (ML). Report the values of the ML estimators.** .

```
## 2 alleles A and T, 3 genotypes TA, AT, TT
```

```r
#frequency=num(TT or AT)/total
gen_table/n
```

```
##
##        AT        TT
## 0.2956989 0.7043011
```

1.c) **Count the number of alleles of each type in the sample. Estimate the relative allele frequencies by ML. Report the values of the ML estimators.** .

```r
allele_num<-table(unlist(strsplit(as.character(db$V1),"")))
#frequency=num(A or B)/total alleles
#total alleles =2n
al_freq<-allele_num/(2*n)
```

```
## Number of alleles of each type in the sample
```

```
##
##   A   T
##  55 317
```

```
## Relative allele frequencies
```

```
##
##         A         T
## 0.1478495 0.8521505
```

1.d) **Which allele is the minor (least common) allele?** .

```
## The least common allele is A with 55 repetitions.
```

1.e) **Do a likelihood ratio test (LRT) for Hardy-Weinberg equilibrium using the HWLratio function of the R-package HardyWeinberg. Report the likelihood ratio statistic and the p-value.** .

```r
#for HWLratio to work we need 3 genotypes on the table, we had AT and TT, we add AA
gen_table["AA"]<-0
print(HWLratio(gen_table))
```

```
## Likelihood ratio test for Hardy-Weinberg equilibrium
## G2 = 9.591049 DF = 1 p-value = 0.001955282
## $Lambda
##          AT
## 0.008266661
##
## $G2
##       AT
## 9.591049
##
## $pval
##          AT
## 0.001955282
```

***1.f)* State your conclusion of the LRT.** .

```
## It is very unlikely thet these results are due to random chance, as lambda and the p-value have
## low values. We can state that the alleles are not associated with each other.
```

***1.g)* State the distribution the LR statistic for this problem.** .

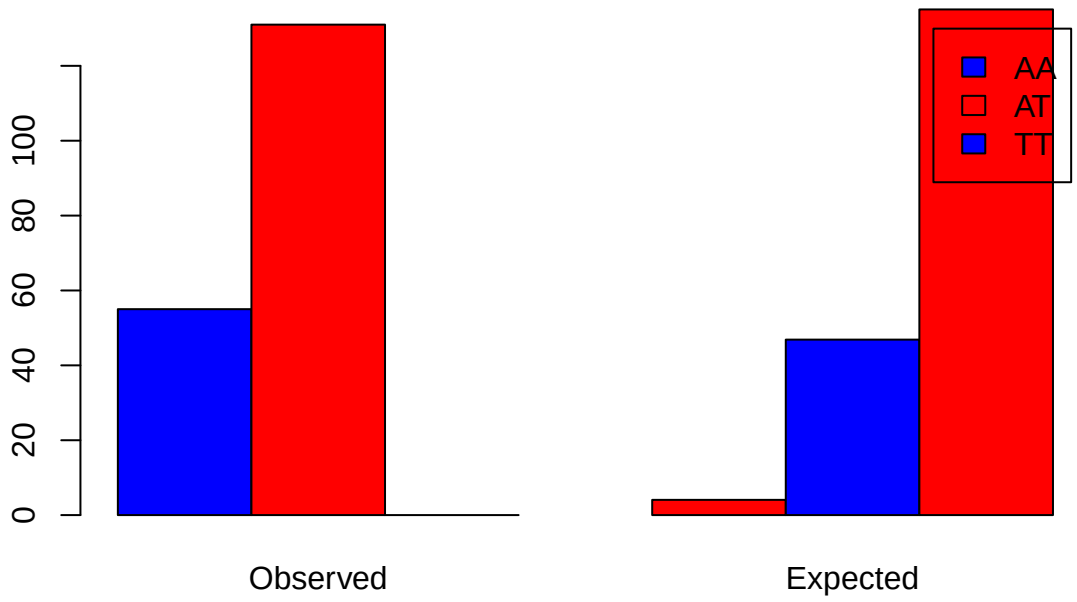
```r
#the data is skewed to the right, we check if its is chi-square
set.seed(1)
obs<-gen_table
total<-n
A_freq<- (2*obs["AA"]+obs["AT"]) / (2*total)
expected<-c(AA=A_freq^2 * total,
            AT=2 * A_freq * (1-A_freq)*total,
            TT=(1-A_freq)^2 * total)
chi_sq<-sum((obs-expected)^2 / expected)

p_value<-1-pchisq(chi_sq, df=1)

# Assuming 'observed' and 'expected' are vectors representing counts

# Combine observed and expected counts into a matrix
combined_data <- cbind(obs, expected)

# Barplot visualization
barplot(combined_data, beside = TRUE, col = c("blue", "red"),
        names.arg = c("Observed", "Expected"), legend.text = c("AA", "AT", "TT"))
```




```
## The observed distribution aligns closely with the expected chi-square distribution.
## Therefore, there is evidence to suggest that the observed data follows a chi-square distribution,
## supporting the hypothesis of a good fit.
```

1.h) **Calculate the p-value "by hand" using the value observed for the LR statistic and its distribution. Show your computations. Do you obtain the same result as the HWLratio function?** .

```
obs_LR<-9.591049
df<-2-1
p_value <- 1 - pchisq(obs_LR, df)
```

```
## The p-value of our distribution is 0.001955282 which is the same as the result of
## the HWLratio function
```

1.i) **Calculate the expected genotype counts under the assumption of Hardy-Weinberg equilibrium. Compare them with the observed counts. What do you observe?** .

```
total_alleles <- 2 * sum(gen_table)
alleles <- unlist(strsplit(db$V1, ""))
A_count <- sum(alleles == "A")
T_count <- total_alleles - A_count
freq_A <- A_count / total_alleles
```

```
freq_T <- T_count / total_alleles
expected_AA <- freq_A^2
expected_AT <- 2 * freq_A * freq_T
expected_TT <- freq_T^2
#calculate the expected genotypes
expected <- c(
  AA = expected_AA * total_alleles/2,
  AT = expected_AT * total_alleles/2,
  TT = expected_TT * total_alleles/2
)
```

```
## Observed:
```

```
##  AT  TT  AA
##  55 131   0
```

```
## Expected:
```

```
##        AA        AT        TT
##   4.06586  46.86828 135.06586
```

```
## Comparing both sets, it appears that the expected values closely resemble the observed values
```

Exercise 2

Comparison of regression models. The outcome or response variable is seize and the explanatory variables or predictors are trt, base and age. Subject contains an ID for every individual. The dataset seizures_visit4.xls contains the measures performed at visit 4 in a clinical trial. The Clinical trial was conducted in m = 59 subjects suffering from simple or partial seizures

- Patients were randomized to the anti-epileptic drug progabide or placebo (0= Placebo, 1=Progabide; variable trt).
- A baseline measure of each subject's propensity for seizures was recorded, namely, the number of seizures suffered in the 8 weeks leading up to the start of the study (variable base).
- Each subject's age at the start of assigned treatment was also recorded (variable age).
- After initiation of assigned treatment, the number of seizures experienced by each subject in n = 4 consecutive two-week periods was recorded, so that the response is a count measured at week 8 of follow up (variable seize).

The variable seize is used as the response variable in a multiple regression with the available variables as predictors.

***2.a)* Load the data into the R environment. Do a summary of the data set.** .

```
Edata<-read_excel("./Assignment_data/A2.2-LRT.xls", sheet = 1)
summary(Edata)
```

```
##     subject          seize             trt              base
##  Min.   :101.0   Min.   : 0.000   Min.   :0.0000   Min.   :  6.00
##  1st Qu.:119.5   1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.: 12.00
##  Median :147.0   Median : 4.000   Median :1.0000   Median : 22.00
##  Mean   :168.4   Mean   : 7.305   Mean   :0.5254   Mean   : 31.22
##  3rd Qu.:216.0   3rd Qu.: 8.000   3rd Qu.:1.0000   3rd Qu.: 41.00
##  Max.   :238.0   Max.   :63.000   Max.   :1.0000   Max.   :151.00
##       age
##  Min.   :18.00
##  1st Qu.:23.00
##  Median :28.00
##  Mean   :28.34
##  3rd Qu.:32.00
##  Max.   :42.00
```

2.b) **Fit a full model by the regression of seize on all predictors available in the data set. Report the adjusted R 2 statistic of this model. Which variables are not significant? (use alpha = 0.05).** .

```
reg_model<-lm(seize~base+trt+ age, Edata)
r2_model<-summary(reg_model)$adj.r.sq
res <- summary(reg_model)$coef[,"Pr(>|t|)"]
c=0
for (i in res){c=c+1;if (i<=0.05){significant<-res[c]}}
```

```
## The adjusted R² statistic of the model is 0.704585
```

```
## Base is the only significant variable, as it sthe only variable that has a value above
## the critical value in the model
```

```
##         base
## 1.090869e-16
```

2.c) **Fit a reduced model, eliminating all insignificant predictors from the regression equation in a stepwise fashion (use alpha = 0.05). Report the adjusted R2 statistic of this reduced model. Does this model have a better or worse fit, according to this statistic?** .

```
red_model<-lm(seize~base, Edata)
red_R2_model<-summary(red_model)$adj.r.sq
```

```
## The adjusted R² statistic of the reduced model is 0.7027655
```

```
## The reduced model has a better fit as 0.7027655 < 0.704585
```

2.d) **Do a likelihood ratio test (F-test) to see whether the full or reduced model fits the data better. Report the F statistic, its reference distribution and the p-value, and state your conclusion.** .

```
F_test<-var.test(reg_model, red_model)
F_stat<-F_test[1]
p_value<-F_test[3]
```

```
## F-test between reduced model and full model:
```

```
##
##  F test to compare two variances
##
## data:  reg_model and red_model
## F = 0.99388, num df = 55, denom df = 57, p-value = 0.983
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.586164 1.689450
## sample estimates:
## ratio of variances
##          0.9938786
```

```
## The F statistic is:
```

```
## $statistic
##         F
## 0.9938786
```

```
## The P-value is:
```

```
## $p.value
## [1] 0.9830279
```

2.e) **Do simple linear regressions of seize on the predictors that you eliminated from the model. Do these regressions confirm that the eliminated predictors do not explain seize? State your findings and conclusions.** .

```
attach(Edata)
model_pre1 <- lm(seize ~ trt)
summary(model_pre1)
```

```
##
## Call:
## lm(formula = seize ~ trt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.964 -4.837 -2.964  1.163 56.290
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.964      1.836   4.339 5.92e-05 ***
## trt           -1.255      2.532  -0.495    0.622
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.713 on 57 degrees of freedom
## Multiple R-squared:  0.004288,   Adjusted R-squared:  -0.01318
## F-statistic: 0.2455 on 1 and 57 DF,  p-value: 0.6222
```

```
model_pre2 <- lm(seize ~ base)
summary(model_pre2)
```

```
##
## Call:
## lm(formula = seize ~ base)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1543  -2.2385   0.2925   2.3028  19.5135
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.1255     1.0550  -2.015   0.0487 *
## base          0.3021     0.0257  11.753   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.261 on 57 degrees of freedom
## Multiple R-squared:  0.7079, Adjusted R-squared:  0.7028
## F-statistic: 138.1 on 1 and 57 DF,  p-value: < 2.2e-16
```

```
model_pre3 <- lm(seize ~ age)
summary(model_pre3)
```

```
##
## Call:
## lm(formula = seize ~ age)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.287 -4.434 -2.698  0.361 54.949
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.6407     5.8684   1.813   0.0751 .
## age          -0.1177     0.2022  -0.582   0.5628
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.705 on 57 degrees of freedom
## Multiple R-squared:  0.005909,   Adjusted R-squared:  -0.01153
## F-statistic: 0.3388 on 1 and 57 DF,  p-value: 0.5628
```

.

.

.

Conclusions

```
## These regressions confirm that the eliminated predictors do not explain seize because when we
## analyse every eliminated predictor independenly we can see that their p-value tells us that
## their affect on seize is probably random, but when we  take 'base' as a predictor,we can observe
## that it is nearly impossible that the effect on seize is random due to its minimum p-value
## of < 2.2e-16.
```

2.f) **Are regression coefficients you found in the different regressions consistent with each other? Comment on your findings.** .

```
## The results consistently show the main variable's significance, while trt and age individually lack
## significance. The adjusted R² is 0.7027655 . The difference in fit between the models is marginal,
## with being 0.7027655 lower than 0.704585 the reduced model has slightly better fit.
```