Project2: Survival data

Table of Contents

# Data

## Load

act\_data <- read.delim("../Data/actg320.txt", header = T, sep = "\t")  
log\_data <- read.delim("../Data/Logistic.txt", header = T, sep = "\t") %>%   
 bind\_rows(data.frame(AZT = "Total", AIDS\_yes = sum(.$AIDS\_yes), n= sum(.$n)))

## Summaries

### Logistic data

summary(log\_data)

## AZT AIDS\_yes n   
## Length:3 Min. :25.0 Min. :168.0   
## Class :character 1st Qu.:34.5 1st Qu.:169.0   
## Mode :character Median :44.0 Median :170.0   
## Mean :46.0 Mean :225.3   
## 3rd Qu.:56.5 3rd Qu.:254.0   
## Max. :69.0 Max. :338.0

### ACT data

summary(act\_data)

## id time event tx   
## Min. : 1.0 Min. : 1.0 Min. :0.00000 Min. :0.0000   
## 1st Qu.: 290.5 1st Qu.:174.0 1st Qu.:0.00000 1st Qu.:0.0000   
## Median : 579.0 Median :257.0 Median :0.00000 Median :0.0000   
## Mean : 579.1 Mean :230.2 Mean :0.08341 Mean :0.4987   
## 3rd Qu.: 868.5 3rd Qu.:300.0 3rd Qu.:0.00000 3rd Qu.:1.0000   
## Max. :1156.0 Max. :364.0 Max. :1.00000 Max. :1.0000   
## sex raceth karnof cd4   
## Min. :1.000 Min. :1.000 Min. : 70.0 Min. : 0.00   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.: 90.0 1st Qu.: 23.00   
## Median :1.000 Median :1.000 Median : 90.0 Median : 74.50   
## Mean :1.174 Mean :1.712 Mean : 91.3 Mean : 86.46   
## 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:100.0 3rd Qu.:136.50   
## Max. :2.000 Max. :5.000 Max. :100.0 Max. :392.00   
## age   
## Min. :15.00   
## 1st Qu.:33.00   
## Median :38.00   
## Mean :38.65   
## 3rd Qu.:44.00   
## Max. :73.00

## Plot

#pairplot\_act <- ggpairs(act\_data)  
#ggsave("Plots/pairplot\_act.png", pairplot\_act)

# Analysis of binomial data

## Binomial fit - all

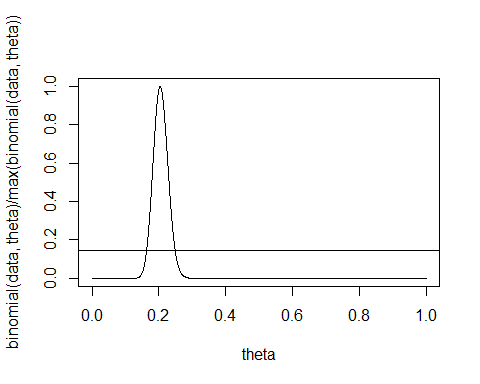
We summarise the data into one group - total number of AIDS cases and total number of patients. Afterwards we fit the binomial distribution to the full (total) dataset. The binomial distribution has the density:

With being the total number of patients and the total number of patients with aids. Hence we get the likelihood:

MLE for is estimated numerically, buut: Confidence intervals are estimated using the likelihood interval method

with $c = exp(- \_{1,(1-)}^2) $ - with being distribution with 1 degree of freedom.

data <- as.numeric(log\_data[3,2:3])  
binomial <- function(data, theta){  
 k <- data[1]  
 n <- data[2]  
 choose(n,k) \* theta^k \*(1-theta)^(n-k)  
}  
  
theta <- seq(0,1,0.001)  
MLE <- data[1]/data[2]  
c <- exp(-1/2\*qchisq(1-0.05,1))  
  
plot(theta, binomial(data, theta)/max(binomial(data, theta)), "l")  
abline(c,0)



a <- theta[binomial(data,theta)/max(binomial(data,theta)) > c]  
CI <- c(min(a), max(a))  
cat("Maximum Likelihood estimate: ", MLE, "\n")

## Maximum Likelihood estimate: 0.204142

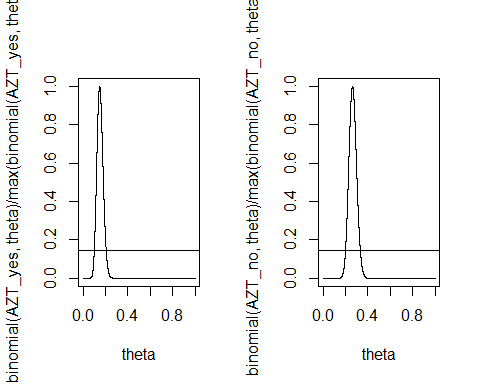
cat("Likelihood Confidence intervals: ", CI)

## Likelihood Confidence intervals: 0.164 0.249

## Binomial fit seperately

Data is divided into 2 groups: a control group (without AZT) and a test group (with AZT). A binomial distribution is fitted to each - as in previous question.

AZT\_yes <- as.numeric(log\_data[1,2:3])  
AZT\_no <- as.numeric(log\_data[2,2:3])  
  
theta <- seq(0,1,0.00001)  
par(mfrow=c(1,2))  
alpha <- 0.05  
c <- exp(-1/2\*qchisq(1-alpha,1))  
plot(theta, binomial(AZT\_yes, theta)/max(binomial(AZT\_yes, theta)), "l")  
abline(c, 0)  
plot(theta, binomial(AZT\_no, theta)/max(binomial(AZT\_no, theta)), "l")  
abline(c, 0)



MLE\_yes <- optimize(binomial, c(0,1), data = AZT\_yes, maximum = TRUE)$maximum  
MLE\_no <- optimize(binomial, c(0,1), data = AZT\_no, maximum = TRUE)$maximum  
  
a\_yes <- theta[binomial(AZT\_yes,theta)/max(binomial(AZT\_yes,theta)) > c]  
CI\_yes <- c(min(a\_yes), max(a\_yes))  
  
a\_no <- theta[binomial(AZT\_no,theta)/max(binomial(AZT\_no,theta)) > c]  
CI\_no <- c(min(a\_no), max(a\_no))  
  
CI\_yes

## [1] 0.09927 0.20540

CI\_no

## [1] 0.19936 0.33166

We notice a slight overlap in the 95 % confidence intervals, indicating no difference (on a 5 & significance level) in the succes probabilities (‘succes’ = you have AIDS).

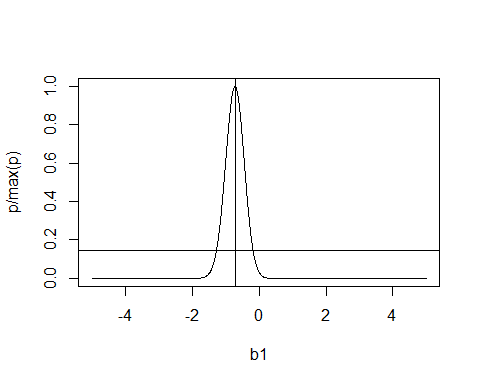
## Estimate the parameters in the model (logistic model) with parameters and

With a logistic model we get the succes probability for the control group and for the test group :

The parameter is the parameter describing the difference in succes probabilities ie. if no difference. We can write the likelihood function for the logsitic distribution:

with denoting succes in category i. control group, test group. The log likelihood for this is

logL\_partial <- function(b0, b1, data, AZT = TRUE){  
 y = if (AZT) {data[1,2]} else {data[2,2]} %>% as.numeric()  
 n = if (AZT) {data[1,3]} else {data[2,3]} %>% as.numeric()  
 false = n-y  
   
 y\*((b0 + b1 \* AZT) \* 1 - log(1+exp(b0+b1\*AZT))) + false\*((b0 + b1 \* AZT) \* 0 - log(1+exp(b0+b1\*AZT)))  
}  
  
logL <- function(theta, data){  
 - (logL\_partial(theta[1], theta[2],data, F) + logL\_partial(theta[1],theta[2],data, T))  
}  
  
  
opt <- nlminb(c(0, 0), logL, lower = c(-Inf,-Inf), data = log\_data)  
  
p0 <- exp(opt$par[1])/(1+exp(opt$par[1]))  
p1 <- exp(opt$par[1] + opt$par[2])/(1+exp(opt$par[1] + opt$par[2]))  
  
  
  
  
  
logL\_b1 <- function(theta, b1, data){  
 -(logL\_partial(theta[1], b1,data, F) + logL\_partial(theta[1],b1,data, T))  
}  
  
profile\_b1 <- function(b1, data){  
 opt <- nlminb(c(0), logL\_b1, lower = c(-Inf), data = log\_data, b1 = b1)  
 # optimize(logL)  
 b0 <- opt$par  
 -logL\_b1(c(b0),b1,data)  
}  
  
  
b1 <- seq(-5,5,0.01)  
alpha <- 0.05  
c <- exp(-1/2\*qchisq(1-alpha,1))  
p = c()  
for (i in 1:length(b1)){  
 p[i] = exp(profile\_b1(b1[i], log\_data))  
}  
  
MLE\_b1 <- optimize(profile\_b1, c(-5,5), data = log\_data, maximum = T)$maximum  
  
plot(b1, p/max(p), 'l')  
abline(c, 0)  
abline(v = MLE\_b1)



a\_b1 <- b1[p/max(p) > c]  
CI\_b1 <- c(min(a\_b1), max(a\_b1))  
  
  
cat("The maximum likelihood estimates of p0 and p1 are: p0 = ", p0, " p1 = ", p1, "\n")

## The maximum likelihood estimates of p0 and p1 are: p0 = 0.2619048 p1 = 0.1470588

cat("The maximum profile likelihood estimate of b1 is: ", MLE\_b1, "\n")

## The maximum profile likelihood estimate of b1 is: -0.7217656

cat("With confidence intervals: ", CI\_b1, "\n")

## With confidence intervals: -1.27 -0.19

The 95% confidence interval does not include 0. The difference is significant.

# Survival stuff

A little bit of summary statistics …. Proportion of death/AIDS in each group…

act\_data <- act\_data %>%   
 select(time, event, tx)

## Exponential distribution

data <- act\_data$time  
likelihood\_gamma <- function(theta, data){  
 -sum(dgamma(data, shape = theta[1], rate = theta[2], log = T))  
}  
likelihood\_pois <- function(theta, data)  
 -sum(dpois(data, lambda = theta[1], log = T))  
dpois

## function (x, lambda, log = FALSE)   
## .Call(C\_dpois, x, lambda, log)  
## <bytecode: 0x000000001cc056a8>  
## <environment: namespace:stats>

opt <- nlminb(c(5,5), likelihood\_gamma, lower = c(0,0), data = data)  
  
x <- seq(1,400, 1)  
plot(x,dgamma(x, opt$par[1], opt$par[2]), 'l')  
par(new = T)  
hist(data)

