Toenail Infection [toenail]

Assignment

Dataset

The dataset considers information from a longitudinal clinical trial in dermatology which was set up to compare the efficacy of two oral treatments (*testing* and *standard*) for toenail infection (De Backer et al., 1998). One of the end points of the study was the degree of onycholysis which expresses the degree of separation of the nail plate from the nail-bed (0, *absent*; 1, *mild*; 2, *moderate*; 3, *severe*) and was evaluated at seven visits (approximately on weeks 0, 4, 8, 12, 24, 36 and 48). In total, 1 908 measurements on 294 patients are available. In this dataset, only a dichotomized onycholysis (0, *absent or mild*; 1, *moderate or severe*) is given.

The data have kindly been made available for statistical research by Novartis, Belgium. The source of the data must be acknowledged in any publication which uses them, see Lesaffre and Spiessens (2001) for more details.

Problem

Compare efficacy in treatment of onycholisis of the *testing* treatment in comparison to the *control* one.

Requirements

Perform exploratory analyzis as described in this document from page 3 and find solutions to each **TASK FOR YOU:** mentioned there. Summarize your solutions in a report (prepared by LaTeX, LibreOffice, MS Word, . . .).

Mail the report in the pdf format (file named as `Surname_Firstname_5.pdf`) and related R script (file named as `Surname_Firstname_5.R`) to `komarek@karlin.mff.cuni.cz`.

Deadline: ***Thursday May 13, 2021 [06:59 CEST]***.

Dataset

The dataset (in ASCII format, space separated values) can be downloaded from
`http://msekce.karlin.mff.cuni.cz/~komarek/vyuka/2020_21/nmst432/Problem_5/toenail.txt`

The dataset contains 1 908 rows (visits) conducted on 294 patients and 5 variables.

Variable list: See Table 1.

Table 1: Variable coding table

Variable Name	Variable Label	Variable Coding
`idnr`	identification number of the patient	integer
`infect`	dichotomized onycholysis	0: *absent or mild*; 1: *moderate or severe*
`trt`	treatment group	0: *control*; 1: *testing*
`time`	time of measurement (in months)	numeric $\in [0,\ 18.5]$
`visit`	visit number	integer $\in \{1,\ \ldots,\ 7\}$

Instructions, hints

This document was prepared using Sweave (Leisch, 2002) in R (R Core Team, 2021), version 4.0.5 (2021-03-31).

The rest of the document provides commented R code that provides some steps of the analyzis which finally leads to a solution to the Problem. Note that not all output is shown in the document below. It is assumed that you run the code by yourself, supplement it by additional commands if needed and use this document only as a guidance through the code and output (that you create).

Initial operations

```
> setwd("/home/komarek/teach/mff_2020/nmst432_AdvRegr/Problem_5/")
> #
> toenail <- read.table("./Data/toenail.txt", header = TRUE)
> dim(toenail)
> head(toenail)
> summary(toenail)
```

```
>   ### Derived variables
> toenail <- transform(toenail,
  +         ftrt = factor(trt, levels = 0:1, labels = c("Control", "Testing")),
  +         fvisit = factor(visit))
> summary(toenail)
      idnr           infect            trt              time            visit      
 Min.   :  1.0   Min.   :0.0000   Min.   :0.0000   Min.   : 0.000   Min.   :1.000  
 1st Qu.:101.8   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 1.000   1st Qu.:2.000  
 Median :192.0   Median :0.0000   Median :1.0000   Median : 3.000   Median :4.000  
 Mean   :189.8   Mean   :0.2138   Mean   :0.5089   Mean   : 4.691   Mean   :3.896  
 3rd Qu.:276.2   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.: 8.893   3rd Qu.:6.000  
 Max.   :383.0   Max.   :1.0000   Max.   :1.0000   Max.   :18.500   Max.   :7.000  

      ftrt      fvisit 
 Control:937   1:294  
 Testing:971   2:288  
               3:283  
               4:272  
               5:263  
               6:244  
               7:264  
```

```
> length(unique(toenail[, "idnr"]))  ### number of patients
[1] 294
```

Exploration

Let $Y_{i,j} \in \{0,\,1\}$ denote dichotomized onycholysis of patient i at his/her visit j performed at time $T_{i,j}$ [months], $i = 1, \dots, N = 294$, $j = 1, \dots, n_i \leq 7$. Further, let $Z_i \in \{0,\,1\}$ denote the treatment group of patient i. To be able to compare the two treatments, we must/may first model probability of infection in two groups of patients over time (and then somehow compare the two evolutions over time). That is, we need to model two functions of time $t \in [0,\,18.5]$:

$$\begin{aligned} \pi_0(t) &:= \mathsf{P}(Y_{i,j} = 1 \,|\, T_{i,j} = t,\, X_i = 0), \\ \pi_1(t) &:= \mathsf{P}(Y_{i,j} = 1 \,|\, T_{i,j} = t,\, X_i = 1). \end{aligned} \tag{1}$$

If it can be assumed that the observations are indepedent, standard logistic regression could be used (with just two covariates – time and treatment group). The fact that we have repeated (and hence not necessarily independent) observations per subject will “only” be a property of data that should be taken into account in final statistical inference (comparison of the two groups by a proper statistical test). Nevertheless, for exploratory part of the analyzis, dependencies can be largely ignored. Moreover, if we restrict our attention to observations from a single visit, independence can again be assumed.

In the following, several plots, also including the response variable, will be created. Since the response is dichotomous, we can somehow increase information value of created plots by “jittering” the response, i.e., by replacing, on plots only!, the observed value $y \in \{0,\,1\}$ by $y + \varepsilon$, where ε is a random variable with mean zero and a symmetric distribution (uniform distribution on interval $(-0.1,\,0.1)$ will be used here).

```
>   ### Jittered version of infect variable
>   ### (useful for plotting)
> set.seed(951913282)
> toenail <- transform(toenail, jinfect = infect + runif(nrow(toenail), -0.1, 0.1))
```

Scatterplot of (jittered) observed values against time while distinguishing the two groups can now be produced (see Figure 1):

```
>   ### Scatterplot
> COL <- c("red3", "darkgreen")
> BG <- c("pink", "aquamarine")
> PCH <- c(21, 23)
> names(COL) <- names(BG) <- names(PCH) <- levels(toenail[, "ftrt"])
> #
> par(mar = c(4, 4, 1, 1) + 0.1)
> with(toenail, plot(time, jinfect, pch = PCH[ftrt], col = COL[ftrt], bg = BG[ftrt],
  +                    xlab = "Time [months]", ylab = "Infection"))
> abline(h = c(0, 1), col = "grey40", lty = 2)
> legend(13, 0.75, legend = names(PCH), pch = PCH, col = COL, pt.bg = BG)
```

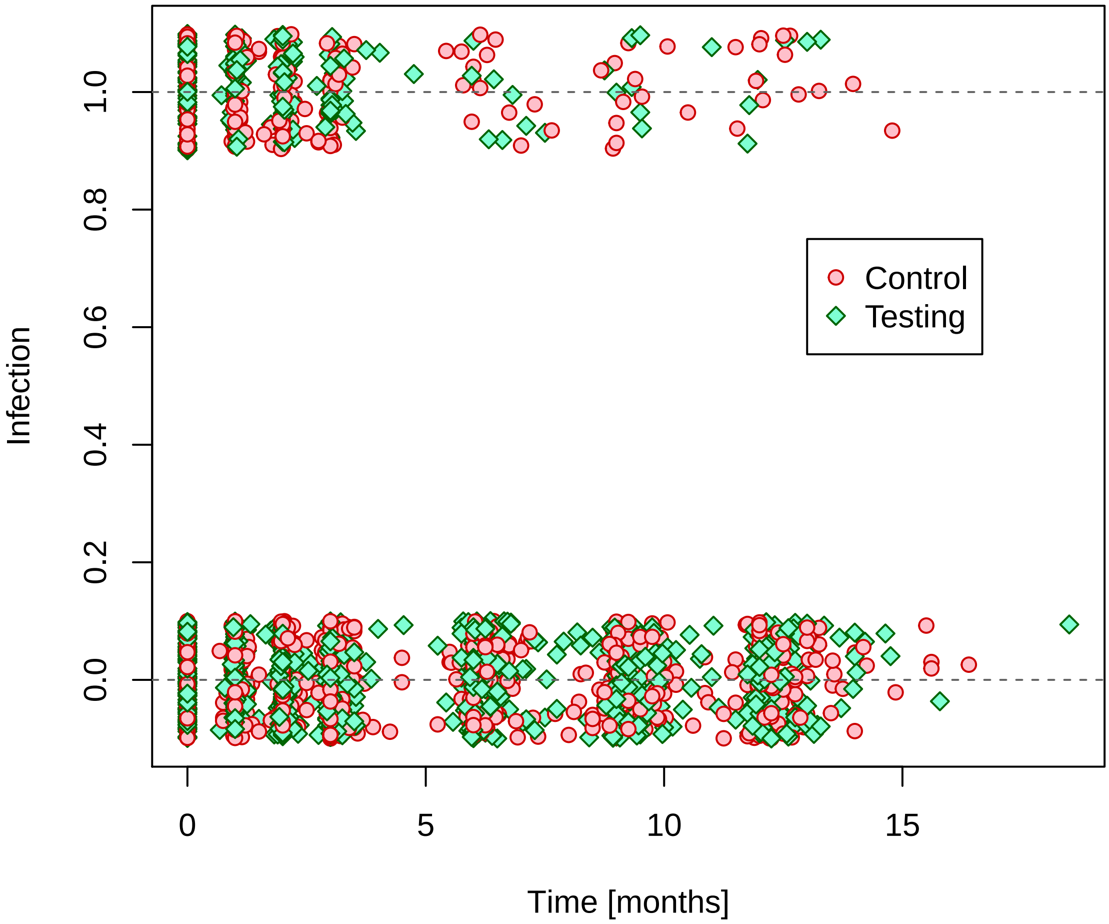

Figure 1: Observed data in the two groups.

Even though exact visit times differ among patients (except the first visit at time 0), it is possible to group observations by visit number and then calculate reasonable empirical estimates of probabilities (1) relevant for neighborhoods of mean visit times:

```
>   ### Frequency tables
> iTAB <- with(toenail, table(visit, infect, ftrt))
> print(iTAB)
, , ftrt = Control

     infect
visit   0   1
    1  92  54
    2  92  49
    3  94  44
    4 103  29
    5 116  14
    6 107  10
    7 119  14

, , ftrt = Testing
```

```
     infect
visit   0   1
    1  93  55
    2  99  48
    3 105  40
    4 111  29
    5 125   8
    6 119   8
    7 125   6
```

```
>   ### Empirical probabilities of infection
> prop.table(iTAB[,,1], margin = 1)     ### proportions in Control group
     infect
visit          0          1
    1 0.63013699 0.36986301
    2 0.65248227 0.34751773
    3 0.68115942 0.31884058
    4 0.78030303 0.21969697
    5 0.89230769 0.10769231
    6 0.91452991 0.08547009
    7 0.89473684 0.10526316
```

```
> prop.table(iTAB[,,2], margin = 1)     ### proportions in Testing group
     infect
visit          0          1
    1 0.62837838 0.37162162
    2 0.67346939 0.32653061
    3 0.72413793 0.27586207
    4 0.79285714 0.20714286
    5 0.93984962 0.06015038
    6 0.93700787 0.06299213
    7 0.95419847 0.04580153
```

```
>   ### Empirical probabilities of infection again
> (pCont <- prop.table(iTAB[,,1], margin = 1)[,2])
> (pTest <- prop.table(iTAB[,,2], margin = 1)[,2])
```

```
>   ### Mean time of each visit
> (tCont <- with(subset(toenail, trt == 0), tapply(time, visit, mean)))
        1         2         3         4         5         6         7 
 0.000000  1.045339  2.058230  3.147998  6.307143  9.302808 12.506445 
```

```
> (tTest <- with(subset(toenail, trt == 1), tapply(time, visit, mean)))
        1         2         3         4         5         6         7 
 0.000000  1.019436  2.060345  3.120153  6.259130  9.319460 12.446565 
```

```
> (TLIM <- range(c(tCont, tTest)))
[1]  0.00000 12.50644
```

Empirical probabilities of infection can now be added to the scatterplot to get a better idea of their evolution over time in the two groups, see Figure 2.

```
>   ### Scatterplot with empirical probabilities of infection
>   ### per visit
> COL2  <- c("red3", "darkgreen")
> names(COL2)  <- levels(toenail[, "ftrt"])
> #
> par(mar = c(4, 4, 1, 1) + 0.1)
> with(toenail, plot(time, jinfect, pch = PCH[ftrt], col = COL[ftrt], bg = BG[ftrt],
+                     xlab = "Time [months]", ylab = "Infection"))
> abline(h = c(0, 1), col = "grey40", lty = 2)
> lines(tCont, pCont, col = COL2["Control"], lwd = 2)
> points(tCont, pCont, pch = PCH["Control"], bg = COL2["Control"], col = COL["Control"],
+         cex = 2)
> lines(tTest, pTest, col = COL2["Testing"], lwd = 2)
> points(tTest, pTest, pch = PCH["Testing"], bg = COL2["Testing"], col = COL["Testing"],
+         cex = 2)
> legend(13, 0.75, legend = names(PCH), pch = PCH, col = COL, pt.bg = BG, bty = "n")
> legend(11.5, 0.75, legend = c("", ""), pch = PCH, col = COL2, pt.bg = COL2, lwd = 2,
+         bty = "n")
```

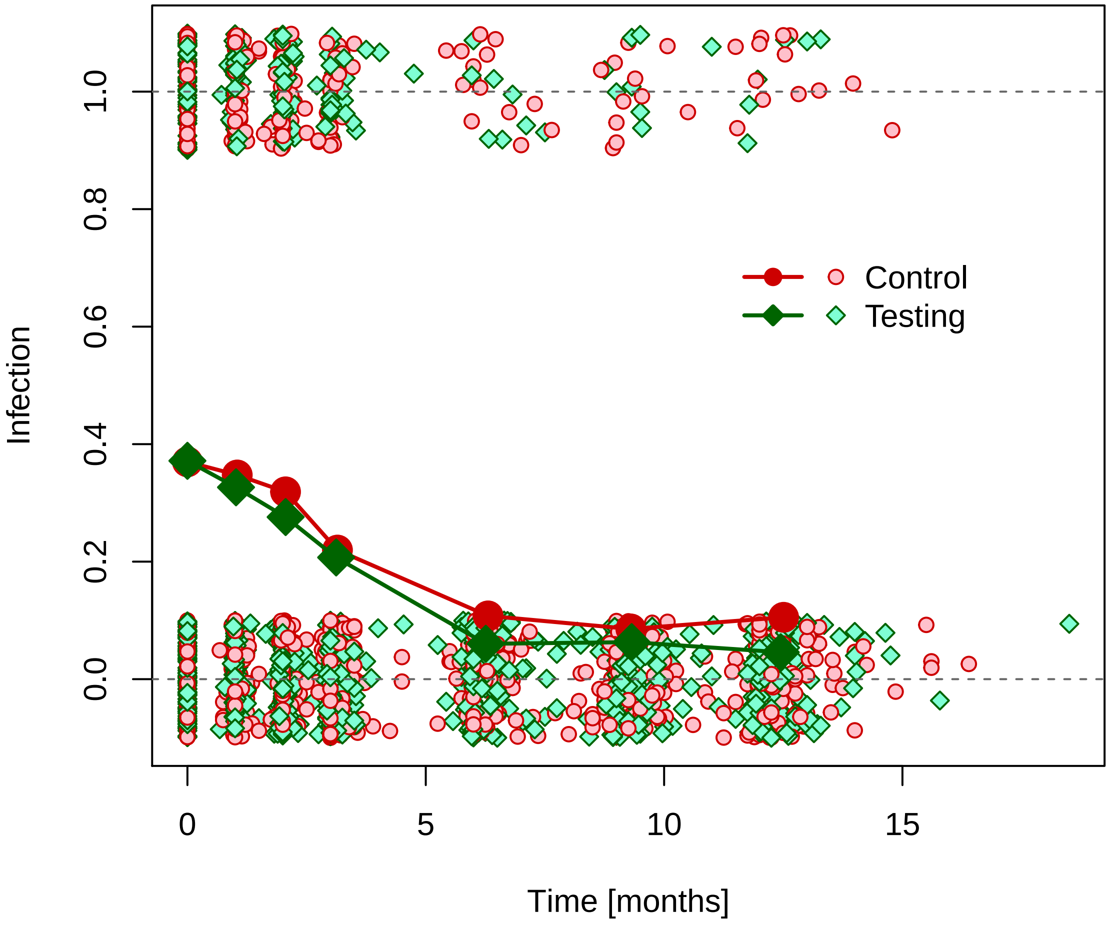

Figure 2: Observed data and empirical probabilities of infection in the two groups.

Next to empirical probabilities, their logits can be calculated and plotted, see Figure 3.

```
>   ### Empirical logits of infection
> (logitCont <- log(pCont / (1 - pCont)))
         1          2          3          4          5          6          7 
-0.5328045 -0.6299683 -0.7591051 -1.2674332 -2.1145329 -2.3702437 -2.1400662 
```

```
> (logitTest <- log(pTest / (1 - pTest)))
         1          2          3          4          5          6          7 
-0.5252663 -0.7239188 -0.9650809 -1.3422344 -2.7488722 -2.6996820 -3.0365543 
```

```
> (LLIM <- range(c(logitCont, logitTest)))
[1] -3.0365543 -0.5252663
```

```
> par(mar = c(4, 4, 1, 1) + 0.1)
> plot(TLIM, LLIM, xlab = "Time [months]", ylab = "Logit of prob. of infection",
+        type = "n")
> lines(tCont, logitCont, col = COL2["Control"], lwd = 2)
> points(tCont, logitCont, pch = PCH["Control"], col = COL["Control"],
+          bg = BG["Control"], cex = 1.5)
> lines(tTest, logitTest, col = COL2["Testing"], lwd = 2)
> points(tTest, logitTest, pch = PCH["Testing"], col = COL["Testing"],
+          bg = BG["Testing"], cex = 1.5)
> legend(7, -0.5, legend = names(PCH), col = COL2, lty = 1, lwd = 2)
```

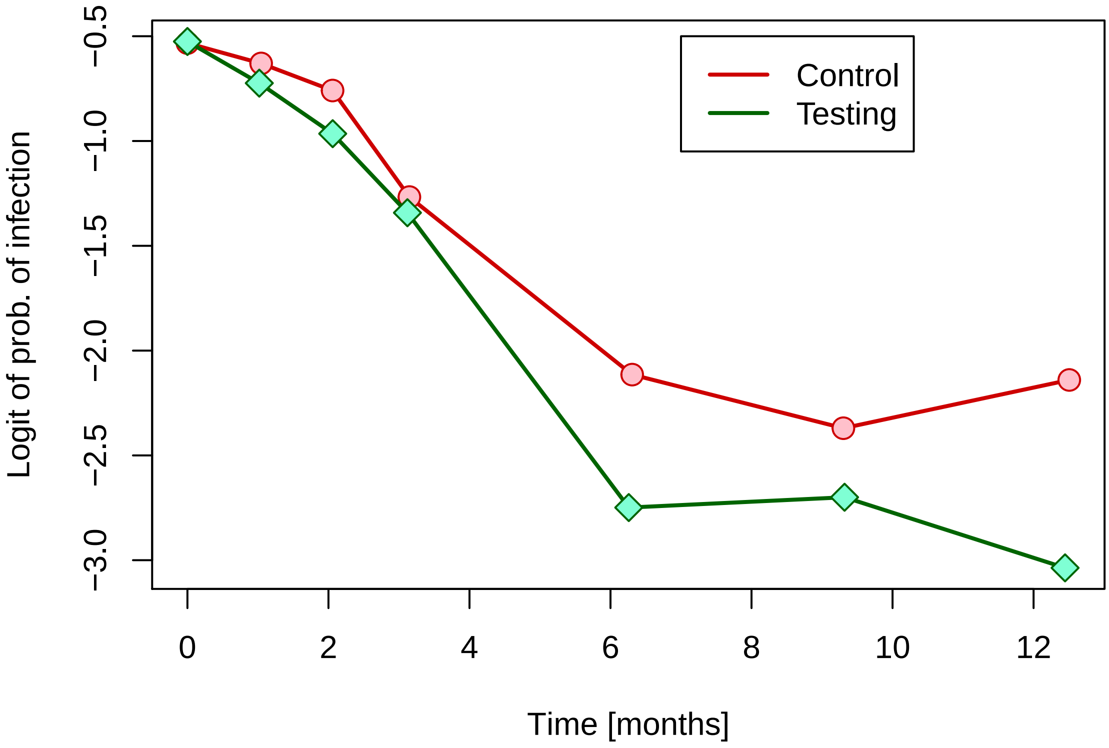

Figure 3: Logits of empirical probabilities of infection in the two groups.

TASK FOR YOU: Use standard logistic model (in which independence of observations is assumed) and develop reasonable model capturing evolution of probabilities (1) of infection over time. Explain, how those probabilities are modelled (in your final model) and provide estimates of the model parameters (including confidence intervals) obtained using a method of maximum-likelihood while assuming independence of observations. Plot estimated versions of the two functions (1) in one plot with empirical probabilities of infection per visit.

Hints:

(i) When checking your model, always (among the other things) calculate fitted probabilities (and/or their logits) as functions of time in each treatment group and plot those functions against above calculated empirical probabilities (and/or their logits) based on data grouped by visit.

(ii) Given the shown plots, it should be clear that a model that proposes to model the logit of probability of infection in each group as a linear function of time is not really appropriate...

TASK FOR YOU: Take your final model and specify a null hypothesis based on this model that will express a hypothesis of no difference in treatment efficacy between the two groups. Explain in words what it means from a clinician's perspective if this hypothesis is rejected. Perform the test (using standard methods for MLE estimated GLM with independence assumption), report the P-value and your conclusion.

Hint:

(i) The null hypothesis can be expressed either as a specific value for some linear combination of regression coefficients or as some submodel of your final model. If you express the null hypothesis by a linear combination of the regression coefficients, provide also related point estimate and confidence interval for exponential of this linear combination (= some odds ratio). Which odds is such an odds ratio comparing?

(ii) There exists more than one reasonable null hypothesis to compare the two treatments. Provide that one which in your opinion is the most relevant.