Exercises

1. Describe in detail why in both datasets, one cannot make use of basic statistical models, like linear regression or ANOVA. What assumptions are violated? What special kind of data is the hip fracture dataset?

# The SII dataset

1. Read in the data in R (variables names are included in the textfile). Use the function str() or use the “environment” tab in RStudio to check if all variables have the correct type. If not use the co-ercion function as.factor() to define a factor variable as in the example:

# lizard$SEX<-as.factor(lizard$SEX)

1. Explore the mathgain values in boys and girls in the different schools using an index plot, similar to the plot in the slides for the lizard data. Within the for-loop, calculate the mean value (by school) for boys and girls, ans plot it versus the schoolID. You can use the code below to start from. We will only consider the clustering on school-level and ignore the clustering on the class-level (this will be included in a later stage).

allMOTHC<-unique(lizard$MOTHC)

nMOTHC<-length(allMOTHC)

plot(as.numeric(lizard$MOTHC),lizard$DORS,type="n",xlab="Mother ID",

ylab="DORS Mean",main="Within-mother comparisons")

for (i in 1:nMOTHC){

tmp<-lizard[lizard$MOTHC==allMOTHC[i],]

tmp$MOTHC<-as.numeric(tmp$MOTHC)

#mean for the males

tmpMale<-tmp[tmp$SEX=="1",]

points(allMOTHC[i],mean(tmpMale$DORS),pch=16)

#mean for the females

tmpFemale<-tmp[tmp$SEX=="2",]

points(allMOTHC[i],mean(tmpFemale$DORS),pch=1)

}

legend("topleft",c("Male","Female"),pch=c(16,1))

If necessary (yes, it is), adjust the range of the Y-axis using the ylim=c() argument.

* What does the plot tell about the difference in Mathgain between boys and girls?
* Why is the default range of the Y-axis not appropriate?

# The hip fracture dataset

1. Read in the data in R, check the variable types via str() or the “environment” tab in RStudio, and co-erce all variables into the correct type. Assign the appropriate factor levels.
2. Plot the MMSE evolutions over time. You can make one plot with all the patients, using a different color for neuro and non-Neuro patients, or generate two separate plots for the neuro and non-neuro persons, respectively. Discuss what you see here.