Case Study

January 16, 2018

The case study consists of the following:

- Each participant is given **two** data sets.
- Optionally each participant can find a **third** data set, related to their own interest.
- Each participant is expected to analyse the first two data sets following the instructions given below.
- For the third data the participantis required to show their own initiative in analysing the data.
- Each participant should write a small summary report describing how they have done the three analyses and describing their results.

1 The data

1.1 Instructions on how to analyse the first data set

The first data set is a subset of Body Mass Index (BMI) data obtained from the Fourth Dutch Growth Study, Fredriks *et al.* (2000a). The data contains BMI for different ages in years for Dutch boys. Each student will be given a different age, for example, 10 to 11 years old. The aim here is to find a suitable distribution of the BMI at this age.

(a) The original data, which contains all ages from zero to twenty, exists in the `gamlss.data` package under the name of `dbbmi`. Each participant should analyse a different age. Here we give an example how to analyse the 10 year olds. We first bring the data set in R and then create a subset `data.frame` containing only a specific age (here from 10-11). The following commands can be used:

```
library(gamlss)
data(dbbmi)
# this value will vary from student to student
old <- 10
da<- with(dbbmi, subset(dbbmi, age>old & age<old+1))
bmi10<-da$bmi
```

You can plot the data using:

```
hist(bmi10)
```

or even better:

```
library(MASS)
truehist(bmi10, nbins=30)
```

By increasing the argument `nbins` the histogram will become more dense. Find a suitable value for `nbins` for the histogram to look good.

(b) Fit different parametric distributions to the data and choose an appropriate distribution to the data. Justify the choice of the distribution.

(c) Output the parameter estimates for your chosen model using the function `summary` and interpret the fitted parameters. (You may have to refer to the GAMLSS distribution book to find the meaning of the fitted parameters).

1.2 Instructions on how to analyse the second data set

Cohen *et al.* (2010) analysed the of handgrip (HG) strength in relation to gender and age in English schoolchildren. There are l 3766 English boys. The data are stored in the packages `gamlss.data` under the name `grip` and contain the variables `grip` and `age`.

(a) Read the data file by typing `data(grip)` into `R`. Note that the **gamlss** packages have to be downloaded first i.e. `library(gamlss)`.

(c) Plot `grip` against `age`.

There is no need to power transform the `age` in this data. Explain why.

(d) Use the LMS method to fit the data[1]. That is, fit the BCCG distribution for `grip`.

`gbccg <- gamlss(grip ~pb(age), sigma.fo=~pb(age), nu.fo=~pb(age), data=da, family=BCCG)`,

where the smoothing for `age` uses the P-splines function `pb()`, i.e. `pb(age)`, for the predictors for parameter μ, σ and ν.

How many degrees of freedom were used for smoothing in the model? Use the function `edf()` or `edfAll()`.

(e) Use the fitted values from the LMS model in (d) as starting values for fitting the BCT and the BCPE distributions to the data, e.g.

```
gbct <- gamlss(grip~pb(age), sigma.fo = ~pb(age), nu.fo = ~pb(age), tau.fo
= ~pb(age), data=da, family=BCT, start.from=gbccg)
```

What are the effective degrees of freedom fitted for the parameters? Try to interpret the effective degrees of freedom.

(f) Use GAIC to compare the three models.

[1] You can simplify some of the steps below by using the `gamlss` package function `lms()` but you still have to justify the choice of the final model

(g) Plot the fitted parameters for the fitted models in (d) and (e) using for example

```
fitted.plot(gbccg, gbct, x=da$age)
```

where `gbccg` and `gbct` are the BCCG and BCT models respectively.

(h) Obtain a centile plot for the fitted models in (d) and (e) using `centiles()` or `centiles.split()` and compare them.

(i) Investigate the residuals from the fitted models in (d) and (e) using e.g. `plot()`, `wp()` (worm plot) and `Q.stats()` (Q-statistics).

(j) Choose between the models and give a reason for your choice.

1.3 Instructions on how to analyse the third data set (optional)

c) Perform a preliminary analysis on your data, This usually involves exploratory plots.

d) Find an appropriate statistical model for the response variable in your data using the explanatory variables. This usually involves selecting:

- an appropriate distribution for your response variable and
- a selection of relevant explanatory variables to explain the response.

e) Use diagnostics to check the assumptions of the model.

f) Use the model for prediction.

References

[1] **Cohen, D. D., Voss C., Taylor, M.J.D., Stasinopoulos, D.M., Delextrat, A. and Sandercock, G.** (2010). Handgrip strength in English schoolchildren. *Acta Paediatrica*, **99**: 1065–1072.

[2] **Fredriks, A.M., van Buuren, S., Burgmeijer, R.J.F., Meulmeester, J.F., Beuker, R.J., Brugman, E., Roede, M.J., Verloove-Vanhorick, S.P. and Wit, J. M.** (2000a). Continuing positive secular change in The Netherlands, 1955-1997. *Pediatric Research*, **47**: 316–323.