

# MA7007 - Statistical Modelling and Forecasting

## Case Study Report 2023-2024

<b>This assessment carries 100% of the module mark</b>
--

This document describes the coursework for MA7007. The coursework involves the statistical analysis of real data sets in R and the writing of a *report* to describe the results. [Note, the emphasis is on the *report*, which means that producing just computer output it is **not** enough. The output should be complimented with intelligent comments and explanations].

The coursework consists of the following:

- Each student is given **two** data sets.
- Each student is required to find a **third** data set, related to their own interest.
- Each student is expected to analyse the first two data sets following the instructions given below.
- For the third data set the student is required to show their own initiative in analysing the data.
- Each student should write a small report (less than 5000 words) describing how they have done the three analyses and describing their results. Section 2 gives instructions on writing the report.
- The students will have a chance to obtain **feedback** on their work at two occasions:
  - by submitting the answers to the first question of the report to the tutor in week 10 (i.e., week commencing 8<sup>th</sup> of April 2024) for an informal assessment.
  - by participating at the tutorials in weeks 11 and 12 where they can show their progress and ask questions.

## 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.* 2000 [1]. 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 two, exists in the `gamlss.data` package under the name of `dbbmi`. Each student should analyse a different age. Here we give an example how to analyse age 10. 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
```

The data.frame `bmi10` now contains only the subject aged from 10 to 11. You can plot the data using:

```
hist(bmi10)
```

or:

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

or:

```
library(gamlss,ggplots) gamlss.ggplots::y_hist(bmi10)
```

Note that by increasing the argument `nbins` of the `truehist()` function the histogram will become more dense. Find a suitable value for `nbins` for the histogram to look good. You only need to show one histogram on your report.

- (b) Fit different parametric distributions to the data and choose an appropriate distribution to the data. Justify the choice of the distribution by explaining what you have done and why you select this specific distribution.
- (c) Output the parameter estimates for your chosen model using the function `summary()` and interpret the fitted parameters. (You may refer to the GAMLSS distribution book Rigby *et al.* (2019) [2] (or to its earlier version which can be found in the GAMLSS web-site <https://www.gamlss.com>) to find what the distribution parameters represent (i.e. location, scale, kurtosis etc.).

## 1.2 Instructions on how to analyse the second data set

Cohen *et al.* (2010) [3] analysed the handgrip (HG) strength in relation to gender and age in English schoolchildren. Here each student is required to analyse a different sample of 1000 from the original 3766 English boys. The data are stored in the packages **gamlss.data** under the name **grip** and contain the variables **grip** and **age**. The aim here is to create centile curves for **grip** given **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)`.
- (b) In order to select your individual sample a unique seed number will be given to you. (In the example below we use the seed number 243 for demonstration.)  
Select your individual sample using your own seed number:  

```
set.seed(243)
index<-sample(3766,1000)
mydata<-grip[index, ]
dim(mydata)
```
- (c) Plot **grip** against **age**.  
Note that there is no need to power transform the **age** in this data set. Explain why.
- (d) Use the **LMS** method to fit the data<sup>1</sup>. 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  $\mu$ ,  $\sigma$  and  $\nu$ .  
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 the generalised Akaike information criterion, **GAIC**, to compare the three models.
- (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.

---

<sup>1</sup>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

- (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

- a) Make sure that the data set is checked for its suitability by your tutor. We are dealing, in this module, with regression type of models, so the data should contain one *response* variable (the target) and more than one explanatory variables.
- b) Give the website source (or other relevant information) for your data set and explain the purpose of the analysis i.e. why you would like to analyse the data.
- c) Perform a preliminary analysis on your data, This usually involves exploratory plots. Comment on how reliable the data are and possible pitfalls on the original data collection.
- 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.

## 2 How to write the report

The Report should have the following structure

1. Introduction
2. First data set (fitting distributions to the data)
  - (a) Comment on the different distributions you are using.
  - (b) Which distribution did you choose?
  - (c) Give reasons why you chose the distribution in part (b).
  - (d) Plot the fitted distribution and comment.
  - (e) State the fitted parameter values of the final chosen model
3. Second data set (centile estimation)
  - (a) Comment on the different models you are using.
  - (b) Answer the explicit questions in section 1.2.
  - (c) Use residual diagnostics for checking the model
  - (d) Comment on how you selected your final model.
  - (e) Comment on the final centile plots.
4. Third data set (students' data)

- (a) Explain why you collected the data and what is the question you are trying to answer.
- (b) Give a preliminary analysis on the collected data and comment on the reliability of the data.
- (c) Use an appropriate model(s) to fit the data.
- (d) Comment on how you selected your final model including diagnostics.
- (e) Show how you will use the model for prediction.

#### 5. Peer Review

- (a) Choose one from a selection of student work on the third data set.
- (b) Write a short critique on the adequacy of the work. You should look at:
  - the quality of the explorative analysis of the data set
  - the choice of the distribution for the response (target) variable
  - the method for selecting and checking the model
  - the interpretation of the results
- (c) Give a grade [A, B, C, D, E, F] representing your estimation of the value of the work.

#### 6. Conclusions

#### 7. References

#### 8. Appendix

##### Important points:

- **DO NOT PUT DATA** in the report (only the first 20 cases in the Appendix).
- Any figure you use should have a caption below like:  
Figure 1: Showing the linear regression of y against time. (You should refer to Figure 1 in the text).
- You should only put the important results or figures in the report and comment on them.
- Put PAGE Numbers into your report.
- The report does not have to be long. Extensive output without comments will get little credit. Your comments and explanation are most important.

Introduction	5
First data set analysis	15
Second data set analysis	25
Third data set analysis	35
Conclusion	5
References	5
Peer review	5
Presentation	5
Total	100

**The completed Case Study Report (no more than 5000 words) must be submitted on Weblearn by Friday 10<sup>th</sup> of May 2024 at 15:00.**

A single pdf of your report saved with the appropriate filename (see below). The filenames should be of the form: A B C.pdf

A = Your ID, B = Case Study Report , C = YY (last 2 digits of the year) For example : 12345678 Case Study Report 24.pdf

## References

- [1] A.M. Fredriks, S. van Buuren, R.J.F. Burgmeijer, J.F. Meulmeester, R.J. Beuker, E. Brugman, M.J. Roede, S.P. Verloove-Vanhorick, and J. M. Wit. Continuing positive secular change in The Netherlands, 1955-1997. *Pediatric Research*, 47:316–323, 2000.
- [2] Robert A Rigby, Mikis D Stasinopoulos, Gillian Z Heller, and Fernanda De Bastiani. *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. CRC press, 2019.
- [3] D. D. Cohen, C. Voss, M.J.D. Taylor, D.M. Stasinopoulos, A. Delestrat, and G.R.H. Sandercock. Handgrip strength in English schoolchildren. *Acta Paediatrica*, 99:1065–1072, 2010.