# Statistical Modelling 2020–2021: EXAM project - August

Upload your answer **as a single pdf file** by

**Friday 20 August, 2021 before 10:00 AM** .

Uploading it earlier is allowed, too late means a zero grade. Overwriting a previous upload before the deadline is allowed, only the last uploaded answer will be graded.

**Instructions:**

- □ Your answer to the following questions should be summarized in at most 8 pages (all included, no appendices are allowed), font size at least 11pt, no footnotes, regular spacing, regular margins. Everything beyond 8 pages will be completely ignored.

- □ All analyses should be performed in R, not with any other software package. The relevant R code that is asked for is needed to evaluate the correctness of the analysis. Providing an answer without code is not sufficient. Do not include output in the question about R code. Repetitive or similar lines of code can be shortened to show a relevant case only.

- □ Your answers + code should be uploaded as a single file in pdf format. Make sure that your name and student number appear on the first page. All answers should be typed, not handwritten.

- □ Clearly indicate each question number. Do not copy the question itself in your answer. Every question number should be present, even when you leave the answer blanc. No question number is no grade for that question.

- □ Use correct statistical notation.

- □ We will NOT verify any answers nor code before it is handed in. This is an exam.

- □ Grading: This exam counts for 16 points out of 20 on your final grade for this course. Question 1: 3 points, question 2: 4 points, question 3: 5 points, question 4: 4 points. The homework counts for 4/20. In case you did not turn in the homework it counts for 0/4. You are allowed to turn in the final exam without having turned in the homework, but in that case the maximum total is thus 16/20. There is no second chance for the homework.

- □ This exam should be individual work. The write-up must be your own. Committing plagiarism by copying parts of the work of another person and turning it in as your own, is strictly forbidden (and illegal). Do not post, distribute or pass on parts of solutions to someone else.
  **The university exam regulations apply to this exam.**

**Good luck!**

**Dataset**: You use a subset (see below) of the dataset in the file "OnlineCourse2021.txt" which consist of 14 columns and 3042 rows. Each row represents information about a student taking an online course. The columns in the file are explained below.

**pass** This is the response variable only for question 4. This variable is not used for questions 1–3. 0 = No, 1= Yes.

**score_exam** This is the response variable only for questions 1–3. This variable is not used for question 4. This is the actual exam score, maximum is 100.

**semester** `x1` 4-level factor indicating the semester in which the course has been taken.

**gender** `x2` 'M' for male, 'F' for female

**highest_education** `x3` 5-level factor, student's highest education level on entry.

**soc_eco_status** `x4` student's socio-economic rank, can be treated as continuous.

**age** `x5` student's age.

**num_of_prev_attempts** `x6` the number of times the student has attempted this course.

**studied_credits** `x7` the total number of credits for all courses taken by this student.

**disability** `x8` 1=Yes, 0=No, indicates whether the student has declared a disability.

**wscore_tutor** `x9` Weighted score in tutor marked assignments.

**wscore_comp** `x10` Weighted score in computer marked assignments.

**exercises** `x11` Total number of online exercises the student solved on the day before the exam.

Since this dataset is constructed for the purpose of the exam, it may not be used nor interpreted for any other purpose.

## Construction of your individual exam data set

Each one of you constructs a subset of this data of size 500 in the following way.

```
FullData = read.table("OnlineCourse2021.txt",header=T)
studentnumber = 123456  # fill in your student number here, this is an example!
set.seed(studentnumber)
rownumbers = sample(1:3042,size=500)
mydata = FullData[rownumbers,]
```

**You only work with the dataset mydata!**

# Questions to be answered

**1.** Use `score_exam` as the response variable and ignore the variable `pass` for this question. The models in this question should not treat covariates as random effects.

- **1.A** Use the AIC to select an appropriate distribution and link function for a parametric model for the exam score with all the covariates `x1`,...,`x11` included. Report using correct notation at least three such models with their AIC scores and indicate which model is the best.
- **1.B** Representative R code for Question 1.

**2.** Use `score_exam` as the response variable and ignore the variable `pass` for this question.

- **2.A** Construct and report in correct notation at least three different semiparametric models (the models are required to have at least one nonparametric component).
  - Describe which models were part of the search.
  - Use AIC as a method to select a final semiparametric model.

  Do not forget to mention the appropriate distribution that has been used. It is alright to use a general notation (e.g. $f(x_2)$) for a smooth function, but you have to state which (spline) functions you have used, and how the smoothing parameter was selected.
- **2.B** Only for the components of the selected model that are modeled in a nonlinear way, provide graphs of the estimated functions and briefly discuss.
- **2.C** Representative R code for Question 2.

**3.** For this question you use the response `score_exam` and (obligated) the covariates `x9` (`wscore_tutor`) and `x10` (`wscore_comp`), including other variables in the model as appropriate. Ignore the variable `pass`.

- **3.A** Construct a graph containing the bivariate smooth effect of `x9` and `x10` on the response variable, including other variables as appropriate.
- **3.B** Give the model that has been fitted to produce the graph using correct notation and briefly discuss the graph.
- **3.C** Consider now a parametric model where you test with an order selection test whether the model is additive in the variables `x9` and `x10`, other variables are included as appropriate. Report the full testing procedure from hypotheses to p-value and conclusion in a correct way. Briefly discuss whether the conclusion of the test corresponds to what is observed in the graph.
- **3.D** Representative R code for Question 3.

**4.** Use `pass` as the response variable and ignore the variable `score_exam` for this question. The variable `semester` is used as a random effect in this question.

**4.A** In a large model without preselection of variables use lasso estimation (penalized $L_1$ estimation) for generalized linear mixed models as provided in the R library `glmmLasso`. Here is some example code, to be adjusted to your data and variable names.

```
library(glmmLasso)
lassofit< −glmmLasso(pass∼x1+as.factor(x2), rnd = list(semester=∼1),
lambda = ..., family = ..., data = mydata)
```

Find a good lasso fit for these data using the BIC value as provided by that package to determine the value of the regularization parameter. Include a graph of the BIC values versus the regularization parameter and indicate the best value.
Note: depending on the version of R that is used the software might produce a warning of the following type that you may ignore. `In if (class(InvFisher2) == "try-error")` ... This is seemingly a build-in syntax problem.

**4.B** Construct a table containing the vector of estimated coefficients of the model using (i) the BIC-chosen regularization parameter and (ii) no regularization. Give a brief discussion of what is observed from this table.

**4.C** For lasso estimation write as a mathematical formula (no code) the corresponding formula for the expectation of a student passing this course.

**4.D** Representative R code for Question 4.