Exam Introduction to Modelling (FIML) Semester 1 2020/2021

Student Number: 613234

Deadline: 22-10-2021 before midnight.

## General Instructions

There is no written exam for this course. Instead, you have to write a quantitative report where you answer six questions. These questions will be slightly different for each student but based on the same learning goals. The datasets used for the questions will also be unique for each student. In the last class (week 7) you will receive the questions and you will have exactly 1 week to make the report that has the answers to the questions. The percentages in the instructions indicate how much each questions weighs towards the final grade.

**Upload the report on handin.han.nl as a word or pdf file.**

All of the answers need to have a clear answer, the regression results should be presented in APA tables, and be accompanied by a working R-script.

The answers to the questions account for 75% of the grade. The remaining 25% is determined based on how well the answers are presented. This means:

* The regression analyses should be presented in a table according to APA standards (10%)
* The report should look professional (e.g. title page, table of contents, page numbers, APA referencing) (5%)
* The R-scripts should be working and should make clear use of notes (10%)

## Question 1 (15%)

The file Student\_Alcohol\_Consumption.csv contains data about students at a Portuguese school. The board of the school is increasingly worried about so-called binge drinking and decided to do research on the drinking behavior of students and how these relate to personal and school related characteristics. The board is also interested in whether the relationship between these characteristics and alcohol consumption is different between boys and girls. Thus, develop and present three models. The first model should contain the results for all students, the second the results for boys, and the third for girls. For each model, explain clearly what each variable means. Also discuss if there are any differences between the boys and the girls when it comes to the relationship between the variables and amount of alcoholc consumed. The variables you have to use are these:

* Alcohol: weekly alcohol consumption, 1 = Very low, 5 = Very high
* Sex: F = female, M = male
* Health: current health (1=very bad, 5=very good)
* Address: U = Urban, R = Rural
* Age: age in years
* Romantic: is the student currently in a romantic relationship?
* Activities: whether the student partakes in extra-curricular activities

## Question 2 (10%)

The file amazon\_books.csv contain data of 312 books crawled from amazon.com. Your job is to find out which variables are associated with the price that Amazon sets for the books (variable AmazonPrice). Check if there is multicollinearity and, if so, deal with it appropriately. Also standardize the appropriate variables and report which variable has the strongest influence on the price set by Amazon. Use the following variables:

* Price of the books (AmazonPrice)
* Height (Height)
* Hardback of Paperback (HardOrPaper)
* Number of pages (NumPages)
* Weight (Weight)
* Thickness (Thick)

## Question 3 (15%)

Open the file EconomicsJournalSubscriptionData.csv, it contains data about subscriptions to economic journals at US libraries for the year 2000. Your job is to create a model with the number of subscribers (subs) as the dependent variable. Use the following variables:

* Price of the journal (price)
* Characters per page (charpp
* Number of citations (citations)
* Year of first publication (foundingyear)

There might be non-linear relationships in the model, investigate if this is the case and if so, show this in a graph. Present a model without any transformation and one with the appropriate transformations. Which models fits better?

## Question 4 (10%)

The file EcommerceCustomers contains data about how much money customers spend on clothing online and how much time they spend on the website and on the app. The store offers in-store style and clothing advice sessions. Customers come in to the store, have sessions/meetings with a personal stylist, then they can go home and order either on a mobile app or website for the clothes they want. Your job is to create a predictive model that can predict how much a specific person will spend in a year (YearlyAmountSpent). Use the following variables:

* Average time spent with stylist (AvgSessionLength)
* Average length of time spent in the app (TimeonApp)
* Average length of time spent in the app (TimeonWebsite)
* How long they have been a member (LengthofMembership)

Present a model which can most effectively predict this, keeping in mind the principle of parsimony. How do you know that this model can be used to create accurate predictions? Compare the prediction of the model with the actual amount spent for the first person in the dataset. How big is the difference between the predicted amount of money spent and the actual amount of money spent for the first person in the dataset?

## Question 5 (10%)

The file Philadelphia\_Crime\_Rate contains data about the number of crimes per 1000 people (variable CrimeRate) for different cities in the state of Philadelphia. Create a model that predicts the number of crimes (CrimeRate) based on the following variables:

* Housing Price (HousePrice)
* Distance from the city center (MilesPhila)
* Change in population (PopChg)

Are there any outliers in this data? If so, show how you found these and what the effects are on the results. Also describe how you dealt with the outliers in your final model.

## Question 6 (15%)

The file graduate\_earnings.csv contains data about the earnings of alumni from different schools in the United States. Alumni were asked what their yearly earnings (Earn) were 5 years after graduating. Present a model predicting how much money alumni make after 5 year (Earn) using the following independent variables:

* Whether their university is public or private (Public)
* Their SAT (SAT)
* The tuition fee (Price)
* How much of their tuition fee was paid for through a scholarship (need\_fraction)

Some of the variables have missing observations, identify which these are and deal with them accordingly. Then present a model where you dealt with the missing values and compare this to the model that excluded the missing data. Is there a difference between the models? Which model would you prefer?