**Exam**

There is no written exam for this course. Instead, you have to write a quantitative report where you answer five questions. The datasets used for the questions will be unique for each student so if you copy the results from another students I can easily tell. In the last class (week 7) you will receive the questions and you will have exactly 2 weeks (not including holiday) to make the report that has the answers to the questions. This means the deadline is Thursday 3rd of October at 23:59.

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

The answers to the questions account for 85% of the grade. The remaining 15% 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
* The report should look professional (e.g. title page, table of contents, page numbers)
* The R-scripts should be working and should make clear use of notes

The report and r-scripts need to be uploaded on handin.han.nl

**Question 1 (20%)**

The file **Article\_Page\_Views.csv** contains data about various online articles. Your job is to build a model that can estimate the page views (dependent variable) based on the other variables in the data. For each variable in the model, clearly explain what the correct interpretation is. Also make sure to check for multicollinearity and deal with it accordingly.

Once this is done, create two more models: one for news articles and one for entertainment articles. Present all three models in a single table and describe the most important differences between the models

Definition of the variables:

* Page\_views: number of people who have viewed the article page (dependent variable)
* word\_count: number of words in the article
* read\_time: average number of minutes it takes to read the article
* gender: gender of the author
* pictures: number of pictures in the article
* type: whether it’s an entertainment or news article
* clickbait\_title: a judgment of whether it has a clickbait title (yes/no/maybe)

**Question 2 (15%)**

The file **Marathon\_Times.csv** contains data about the marathon time (variable minutes) of 600 runners. Build a model that can estimate the average time for the runners. First standardize the appropriate variables and describe which variable has the largest influence on running time. Second, identify if there are outliers in the data and if so, deal with them appropriately. Describe what you did to deal with the outliers. Also present two models in a single table: one with all the observations and one with the outliers you removed. Are there any differences between the two models?

Definition of the variables:

* participant: participant number
* minutes: Minutes it took to complete the marathon (dependent variable)
* n\_previous\_marathons: how many marathons the participant has completed in the past
* age: Age of the participant
* gender: gender of the participant
* hills: How many hills there are on the track
* weather: Weather conditions during the marathon (sunny, cloudy, light rain or heavy rain)

**Question 3 (15%)**

The file **Second\_Hand\_Laptops.csv** contains data about the selling price (variable Price) of 500 secondhand laptops. Based on this data, you want to build a model that can accurately predict the average price that second-hand laptops are sold at. When creating the model, keep in mind the principle of parsimony. Can you make accurate predictions using this data? How do you know?

You own a laptop that you want to sell and you want to use the model to predict the average price that it will sell at. It is a three year old HP laptop that is medium size, with four USB ports. After looking up some data on the internet, you found that worldwide it sold about one million copies. Using your model, what is the average price that you can expect to fetch for the laptop?

Definition of the variables:

* Brand: Brand of the laptop (Apple, HP or MSI)
* USB\_ports: how many USB ports the laptop has
* Size: size of the laptop (small, medium or large)
* Age: how many years old the laptop is
* Sold\_worldwide: how many laptops of this type have been sold worldwide (in thousands)
* Price: average price that the laptop is sold at in Euros (dependent variable)

**Question 4 (15%)**

The file **Book\_store.csv** contains data about the sales of a sample of books as well as several characteristics of these books. Some independent variables might have a non-linear relationship with the dependent variable. Identify which these are and show what kind of relationship they have using a graph. Then, make the appropriate transformation to the variables and describe whether this made the model better or not.

Definition of the variables:

* books\_sold: number of copies of the book sold in a week (dependent variable)
* weeks\_best\_seller: how many weeks the books has been on a best seller list
* n\_books\_published: how many books the author published before
* genre: genre of the book (mystery, science fiction or romance)
* page\_numbers: how many pages are in the book
* publication\_year: the year in which the book was originally published

**Question 5 (20%)**

The file **Hotel\_rooms.csv** contains data about the bookings in hotels during off-season; it also includes data about characteristics of these hotels. Your job is to estimate the number of hotel bookings using the other variables in the dataset. Unfortunately, the dataset contains many missing values which could bias the results. First present a regression model where the missing values are simply removed. Second, use multiple imputation to estimate a second model where the missing values have been imputed. Present both models and describe the differences between them.

Definition of the variables:

* perc\_rooms\_booked: the percentage of rooms that have been booked in a hotel
* pool: whether the hotel has a pool or not
* gym: whether the hotel has a gym or not
* breakfast: whether breakfast is included in the price
* price: average price of a room per night
* distance: distance from the city center
* rating: the most frequent rating on google maps