

# Final Term Project

Data Analysis 2 and Coding with R

MS in Business Analytics,  
2021/2022 Fall



## 1 The task

Your task is to analyse your research question, based on your chosen dataset. You need to choose one of the following broad topics

- Inequality
- Sport success
- Pricing and Sales

You will need to pick a topic, find a research question, get the data and answer the question. Find an interesting question that focuses on the potential relationship between  $y$  and  $x$ . Indeed, a research question is defined with a  $y$  (dependent/outcome), and  $x$  (explanatory/causal) variable, and some  $z$  (control/conditioning) variables.

There is no need for a literature review, but you need to argue why to expect a relationship between  $y$  and  $x$ .

Note that one potential result is that while we expected a link, there is no relationship between  $y$  and  $x$  - that is perfectly okay.

## 2 General rules

This assignment is evaluated for both *Data Analysis 2: Finding Patterns with Regressions* and *Coding 1: Data Management and Analysis with R*.

- **Sign-up:** You need to sign up and get an approved topic not later than **10th of December 11.55 PM** in this [LINK](#).
  - To have an approved topic, you need the followings:
    - \* Have more than 50 non-missing observations.
    - \* Have at least two control variables ( $Z$ ).
    - \* **Not** using the same or very similar datasets as we used in class
    - \* Use an already analysed dataset from the web. (We will check both data and codes later on.)
    - \* Use a topic, which can be done by cross-sectional analysis (thus no time-series)
- **Submission:** You need to upload the knitted pdf file to *Data Analysis 2: Finding Patterns with Regressions* **AND** the pdf file with the '.Rmd' file to *Coding 1: Data Management and Analysis with R* ceulearning's site.
  - You must put your data and codes into your github account and take care that we can run your code only by running the .Rmd, thus you call your data directly from your github repo.
- **Submission deadline:** **Wednesday, 22th of December 2021, 11:55 PM**
  - Late submission: 1 day delay -25%, after that no points.

### 3 General description

The aim of this assignment is to show your skills in data analysis and coding, learnt during this Fall semester. What we are looking in your final term project is the following:

- You show a compact report on an issue - which is easy to read (nicely formatted) and understand (not using complicated technical terms or itemized steps of data analysis).
- The report has a clear message - what we can learn from the data - and there is a proper argument based on learnt statistical tools.
- You can state your research question based on your data and you can provide an answer.
- You are confident in, what we can learn from the data and what you can not claim based on your dataset.
- You understand the possible challenges of your dataset and can access the uncertainties, which comes from data quality issues. These uncertainties are well articulated when conclusions are made.
- You can show pattern of association between variables, you can transform these variables in order to handle them in a (linear) regression model.
- You understand the nature of your outcome variable and you can use proper model(s) to handle this nature.
  - Within this model you can show you have mastered what we have learnt and can use them in analysis to make your argument more robust.
- You can show how to generalize your results and what are the constraints of this generalization.
- Formatting requirement:
  - PDF knitted by rmarkdown.
  - **Length:** max 4 pages of report and then appendix (this can be as long as you wish)
- You should study the provided materials in details (see the step-by-step for analysis and the example for term project) and use them as guidelines.

### 4 Structure and strong recommendations

We require a strongly structured output:

1. Introduction: introduce the problem, why interesting
2. Data: present the dataset, describe key features
3. Model:
  - (a) Present the model you estimate, argue for your model choice
  - (b) Discuss your variables, process of feature engineering. (In words, put all needed graphs, estimations into the appendix if needed.)
  - (c) Show core results. Interpret precisely what you got.
4. Generalization and external validity (robustness check)
  - (a) Show some robustness / alternative models.
5. Causal interpretation / main summary
  - (a) Summarize your findings. Discuss room for a causal interpretation.
6. Conclusion
  - (a) Conclude and make business / policy comments/recommendations

When writing your report, keep in mind the followings:

- Create rather a scientific paper or newspaper article style of report than an itemized description what you have done.
- Pay attention to format your graphs and tables:
  - Names, theme, labels, notes, ticks/values of y and x axis, ect.
  - Use a unified theme and be consistent during the report to use only one type.
  - Regression tables - name your variables (avoid e.g. `ln_gdp_per_capita_sq`, rather use  $\ln(\text{gppc})^2$ ).
  - When reporting numbers use 2 or 3 digits.
- Never include code chunks or outputs in your report. Always use some kind of formatting!
- Think carefully what you put into your first 4 pages. Put everything which is not essential for understanding the main results into the appendix!
- Note: If you use weighted linear regression - there is a different interpretation than simple linear regression.

## 5 Evaluation

### 5.1 DA2

For DA2 only pdf is going to be evaluated. Overall you can earn 110%.

- Have a research question, motivation and conclusion. Inference (incl external validity), causality discussed. 20%
- Data is well described, with sources, important features. May be (not needed) graph(s), table(s) to support. 15%
- Modelling choices are well presented, and supported by graph(s), table(s). Discuss filtering (if any), variable transformation (if any), estimated model. 20%
- Results are logically presented. Key coefficient(s) explained precisely. Results are appropriately interpreted. 20%
- Robustness, sensitivity presented 10%
- The paper is well structured, easy to follow, graphs and tables look good. 15%
- There is something extra 10%

### 5.2 Coding in R

Overall you can get 50 points for the take-home examination.

- Rmarkdown file runs and produces the attached pdf file (5p)
  - Data is read from your github repo (2/5p)
  - The attached file is the same as the produced file (3/5p)
    - \* Title and sections (1/5p)
    - \* Tex formatting (2/5p)
- Readability of the code (8p)
  - Proper commenting - easy to understand what the code (intended) to do (6/8p)

- Proper spacing of the code and general outlook (2/8p)
- Code does what it is intended/claimed to do in pdf/html (15p)
  - Data cleaning and munging (5p)
  - Data descriptives (2p)
  - Data visualization (3p)
  - Running regressions (3p)
  - Robustness checks and generalization (2p)
- Visualization - how graphs looks and annotated (10p)
  - Size and placing of the figure (3/10p)
  - Labels, readable theme, axis ticks, limits, legends if used (7/10p)
- Tables - how tables and regression results are presented and annotated (8p)
  - Size and placing of the figure (3/8p)
  - Variable names, number of digits, notes if needed, reported values and variables inc case of regressions (5/8p)
- Appendix formatting and readability. (4p)

Good luck!