Project Guidelines

Spring 2023

Professor: Borja Mesa Sánchez

This project is an integral part of the course. It consists of the identification of a real-world problem, formulation of appropriate hypotheses, collection and statistical analyses of data, presentation and interpretation of obtained results, and limitations of the model.

The project must be done in groups, using the statistical software we are using in class. The overall project grade will have a significant impact on your overall course grade, so try to do your best.

The theme is free, but here you can find a list of topics students chose in the past:

1. Factors Affecting the GPA of Students.
2. Life Expectancy & Gross Domestic Product per Capita.
3. Factors affecting consumption of tobacco.
4. Does a high/low income tax stimulate the country's GDP per capita?
5. Weather impact on stock market fluctuation.
6. The relationship between murder, unemployment, and crime rate
7. Education, unemployment and GDP
8. How did the economic crisis in the world affect imports and exports?
9. Determinants of Beer Consumption
10. Does income affect obesity rates?
11. Aid versus FDI
12. The problem of alcoholism
13. Happiness and money
14. Why suicide prevails in some countries
15. Electrical vehicles
16. GDP growth and R&D spending

Look for a topic that you find interesting. This topic should be related to the identification of a real-world problem taken from science, social sciences or any other field of interest. You are free to choose your topic since we are really Interested in the Statistical approach.

Read as much as you can about the topic and create a good theoretical framework so that everything makes sense. A good way to start is typing the keywords of your topic in Google Scholar.

Example: You are interested in explaining gender wage differentials in Spain. You type these keywords in Google Scholar and you note that there are hundreds of articles on the subject. You filter all these articles by using the following criteria: number of citations, year of publication (it is recommended that you start by reading more recent papers), and the similarity of the article to your line of research. You read some of these articles and you find that many of them use an empirical model to explain gender wage differentials. You also find that researchers very often use education and experience to explain the behavior of wages. Now you only need the data. If you are interested in Spain, try to read articles dealing with gender wage differentials in Spain. Most of these articles use data from the Spanish Wage Structure Surveys, available in www.ine.es and https://ec.europa.eu/eurostat.

You are on the right track. But you forgot something important! What is your hypothesis? You need a hypothesis to write an article. Are wage differentials in Spain higher than those in Europe? Has the gender wage gap in Spain increased with trade liberalization? What are the factors behind such wage differentials? Look for similar hypotheses in the literature, extract the most important information, and you will be ready to write your own introduction.

Once you define the hypothesis of your article, you will have to work hard on it. Let me tell you about the structure of your project. Your article should encompass the following sub-divisions:

o *Cover page:* lists the project title, the group members, and the date

o *Title:* should reflect the basic idea of your research.

o *Table of Contents*.

o *Abstract:* The abstract is a 5 to 10 lines summary of the entire project. It should include an introductory sentence followed by the objectives, methodology, major findings, and a final sentence on results interpretations and recommendation(s).

o *Keywords:* list the words you think are key in your project. These words should be limited to five.

o *Introduction:* motivate the article. You should specify the reason behind choosing a particular topic. Here is important to include a detailed literature review. Describe briefly what you expect to accomplish in your research. These objectives can be divided into general and specific and might change as you progress in your work: additional ones might be incorporated to the work and original ones modified or discarded. Sometimes, the descriptive analyses might reveal interesting facts about your data that you might want to pursue. You do not need to write the Introduction at the first stages of the project. In fact, I recommend you write it after getting the empirical results. Minimum length: 2 pages.

o *Data:* Population of interest. It is very important to specify and describe your population of interest as well as your sample (size, attributes, etc.). There exist many websites on the Internet where you can find reliable datasets. Here you can find some examples:

- Organization for Economic Co-operation and Development (OECD) (http://www.oecd.org/) (https://data.oecd.org)
- Quandl Financial and Economic Data (https://www.quandl.com/)
- Gapminder (www.gapminder.org)
- Data and Story Library (http://lib.stat.cmu.edu/DASL/): is an online library of datafiles and stories that illustrate the use of basic statistical methods.
- Gallup Organization (www.gallup.com)
- US Census Bureau (www.census.gov)
- US Federal Government (www.usa.gov)
- Bureau of Labor Statistics (www.bls.gov)
- World Bank (www.worldbank.org)
- http://datacatalog.worldbank.org
- http://www.imf.org/external/data.htm
- https://ec.europa.eu/eurostat

Data Cleaning and Preparation: Once you have downloaded the data, proceed to the following:

- Familiarize yourself with the data.
- Identify and classify your variables.
- Prepare the data for analysis.
- Clean the data (missing values, irrelevant variables, etc.)

o *Descriptive Statistics:* Now that your datasets are clean, you should use some of the methods discussed in class to summarize them numerically and graphically (graphs, frequency tables, histograms, boxplots, scatterplots, measures of central tendency, variability, correlation, covariance, etc.). Make sure to check for outliers (Boxplot method and standard deviations) and Normality, decide the best methods to deal with the data. All results should be discussed and supported by the literature.

Example: You have data about education, experience, wages, gender, etc. What are the average years of education in Spain? Average years of experience? How is the spread of the data? Are there outliers? Do the main variables of your model follow a Normal distribution? Are these variables correlated? Try to use all the tools you have learned in the first topic of the course.

o *Methodology*: Carefully explain the tools you are using to test your hypotheses. Choose a method of analysis to meet the objectives of the project (confidence intervals and the performance of hypothesis testing are both required). Check the assumptions to be met to carry out your analysis. Do not forget to correctly define your hypothesis and to be statistically and theoretically consistent.

o *Inferential Statistics*: In this part of the project, you should build some confidence intervals and perform hypothesis testing.

Example: Your project tackles with gender wage discrimination. You collect data about individuals' wages, years of education, experience, and gender. On the one hand, you can obtain some confidence intervals for these variables to compare them with other European countries. On the other hand, it might be interesting to test whether individuals with the same years of education and/or experience, but different gender have the same wage. If you collect data about individuals from more than two countries, you can check whether the average wage gap is the same across countries. All these questions should be closely related to the main topic addressed by the article.

o *The Model*: The core part of the project is the inferential analysis. You are expected to develop a model to generalize results from samples. For this purpose, you will use confidence intervals and hypothesis testing. This section should encompass the following aspects:

- Assumptions of the model (independence, Normality, etc.)
- Formulation of the hypotheses.
- Results
- Analysis of the results
- Interpretation of the most relevant test

All the results should be presented either in a tabular or graphic form. All tables and graphs should be properly numbered, labeled, and discussed. In the discussion section, you should try, as much as possible, to support your findings with similar results from the literature. You are expected to interpret the results.

o *Conclusions*: Write down your conclusion, limitations, and recommendation for future work.

o *Bibliography*: List all the references you have used in your work. No matter the style, but please be consistent.

o *Appendix*: Include all the material not important enough to be included in the main sections of the work.

The final report should be written professionally in the form of a journal article. Try to follow the instructions as closely as possible. Limit the number of graphs and tables you will be presenting. Use succinct and concise phrases. Avoid paraphrasing and plagiarism.

You are expected to submit the final version of your project in PDF format through Campus by using a Turnitin link provided by your teacher. The file with your code, and a video with a 10-minutes presentation will be shared with the teacher using Drive. The submission deadline will be announced by the professor.

Project Rubric Scoring Sheet (the project will be graded on a 0-100 scale)

- Title: reflects the basic idea of the research **(2 points)**
- Abstract: The abstract is clear and concise. The abstract summarizes the entire project properly. The objectives, methodology, major findings, results and interpretations are clearly reported. The word-limit (no more than 200 words) is not exceeded. Keywords describe the work **(3 points)**
- Introduction: The article is motivated adequately. 10 related articles are used in the literature review. Students clearly state what is expected to accomplish in their research. Objectives can be easily identified. Minimum length: 2 pages. Maximum length: 3 pages **(15 points)**
- Data: the source of the dataset is an official webpage. The number of observations is large enough to satisfy the assumptions of the model. Data have been cleaned adequately. **(10 points)**
- Descriptive Statistics: Students use numerical and graphical procedures seen in the first topic of the course to summarize and describe the data. Students use the tools that better fit the data. The explanation of the summary statistics and the graphs is clear and concise **(20 points)**
- Methodology: The students clearly explain the models they are going to use to undertake the analysis. **(5 points)**
- Inferential Statistics:
    - The students build confidence intervals for at least three variables **(15 points)**
    - Perform hypotheses testing for at least three variables (one sample tests) **(20 points).**
    - In order to get the maximum score, students should accomplish the following:
        - The assumptions of the model are clearly stated.
        - Students check whether the main assumptions of the model are met
        - All the tests should be related to the main topic of the project.
        - Hypotheses are clearly stated
        - Results are satisfactorily compared with the literature.
        - All the results are presented either in a tabular or graphic form. All tables and graphs are properly numbered, labeled, and discussed. Students interpret results satisfactorily. Students support their findings with similar results from the literature. The contribution of the article must be clearly stated.
- Conclusions: Students summarize in one paragraph the main results obtained in the work. They point out the limitations of their research, discussing future research. **(5 points)**
- Bibliography: at least 10 references related to the main topic of the article are used. References are sorted in alphabetical order. A generally accepted citation system is used. **(5 points)**

The article extension should be either 6,000 words or 10 pages (maximum extension and minimum extension) (Cover page, Table of contents, and Appendix not included). There will be a penalty of 10 points if this requirement is not met.