

Project Information

Statistics for Data Science

2025/2026

Important remarks

- For cross-sectional data: the project should only be initiated after we finish chapter two - Multiple Linear Regression Model (\sim week 11 of the semester);
- For panel data: the project should only be initiated after we finish chapter 3 - Panel Data (\sim week 12 of the semester);
- For causal inference: the project should only be initiated after we finish chapter 4 - Causal Inference (\sim week 14 of the semester);

Project Description

The final project should be a scientific poster. The poster could be theoretical **OR** empirical (only choose one):

1. For **theoretical project**, the less common option, the students can develop a simulation study to evaluate the properties of an estimator (for time series or panel data) or the properties of a statistical test.
2. For the **empirical project**, the goal is to answer a specific research question related to a topic of interest. You have three settings you can opt (only choose one):
 - ***Cross sectional data***, where you will select relevant variables to answer your research question, specify a model, interpret critically the results and check the model diagnostics;
 - ***Panel data***, where you will select relevant variables to answer your research question, specify a model (FE or RE), interpret critically the results and check the model diagnostics.
 - ***Causal inference methods***, where you will select relevant variables to answer your research question, specify a model, interpret critically the results and check the model diagnostics using the Difference-in-Differences estimator or IV methods.

Reminder: the setting of the classes only covered continuous dependent variable, do not choose a discrete variable as dependent variable for your project.

Note: posters done in R Markdown will be valued.

Evaluation Criteria of Empirical Project

The evaluation criteria for the empirical project are:

1. (10%) **Originality** (use a unique dataset that is not from class examples).
2. (10%) **Presentation** (aesthetic quality of the poster).
3. (15%) **Organization and structure** (clear and concise introduction **outlining the research question and its significance**; methodology, detailing variable selection, model specification, and econometric techniques; description of the results, their implications, and the connection to the research question; discussion and conclusion that ties everything together).
4. (15%) **Content** (present evidence of reflection and analysis of the main literature; provide a rationale for variable selection and model choice, linking them to the research objective).
5. (20%) **Econometric component** (clearly specify the chosen econometric model; justify the choice of model with reference to the research question and data characteristics; perform robust diagnostics to validate the model; interpret the estimated coefficients and other key outputs, connecting them back to the research question; address potential issues such as omitted variable bias, endogeneity, or heteroskedasticity where relevant).
6. (15%) **R component** (Efficiency of the code, use of R Markdown to develop the poster is highly valuable)
7. (15%) **Conclusion and reflection** (address strengths and limitations of the analysis; discuss implication of the results; What are the implications of your work? Which recommendations/contributions do you find? What further research should be done?)

Evaluation Criteria of Theoretical Project

The evaluation criteria for the theoretical project are:

1. (10%) **Originality** (simulation study not addressed in class).
2. (10%) **Presentation** (aesthetic quality of the poster).
3. (15%) **Organization and structure** (clear and concise introduction **outlining the research question and its significance**; description of simulation framework: the design, assumptions, parameters, and steps; discussion of the implications of the findings for theory or practice).
4. (15%) **Content** (demonstrate a strong understanding of the theoretical foundations of the estimator or test; discuss relevant literature, including strengths and weaknesses of existing methods/tests.).
5. (20%) **Econometric component** (clearly define the estimator or test being evaluated and its theoretical properties (e.g., bias, consistency, efficiency); use simulation techniques to assess performance under various conditions (e.g., different sample sizes, distributions, or parameter settings); explain and justify the choice of simulation parameters and metrics used to evaluate the estimator/test; provide detailed diagnostics and interpretation of results (e.g., graphs showing bias across sample sizes)).
6. (15%) **R component** (Efficiency of the code, use of R Markdown to develop the poster is highly valuable)
7. (15%) **Conclusion and reflection** (address strengths and limitations of the analysis; discuss implication of the results; What are the implications of your work? Which recommendations/contributions do you find?)