Final Project details for MA 5013
Applied Regression Analysis, Fall 2025

Overview

At the end of the semester, each student will complete a group project. Groups may have at most two members, and each group will analyze a different dataset using regression tools covered in class.

Every student should take on a distinct and meaningful role within the group, and these roles must be communicated to the instructor.

- If you choose to work alone, that is perfectly acceptable.
- If you choose to work in pairs, please discuss and agree on responsibilities early. This will help avoid misunderstandings or disagreements later on.
- In the rare case that serious conflicts arise within a group, please inform the instructor as soon as possible so that a fair resolution can be reached.

The project is worth **100 points**, divided into three components:

- **Initial Report (5 points):** due **September 12, 2025**.
- **Presentation (55 points):** each team will give a 5 to 10-minute presentation on their project outcomes. These will be scheduled during the final lecture hours, or outside class hours if necessary.
- **Final Report (40 points):** due **November 7, 2025**.

Initial Report (5 points)

The initial report should **not exceed one page** and must be submitted by the team leader. It should include:

- Team details (who is working with whom, or if you are working alone).
- A description of the dataset, including:
  - a layman's description of the problem,
  - the response variable (note: the response must **not** be categorical),
  - the number of observations and predictors,
  - a link to the dataset,
  - and any other relevant details.
- Research questions you aim to investigate.
- Possible methods you plan to apply (restricted to techniques covered in class).

Data Collection

An important part of your project is selecting a dataset to analyze. Each group must use a **different dataset**, and the dataset should be chosen carefully to ensure it is suitable for regression analysis.

- The **response variable must be numerical (not categorical)**. For example, predicting house prices, exam scores, crop yields, or health-related measurements would be appropriate.
- The dataset should have a sufficient number of observations and predictors to allow meaningful analysis, but it should not be so large that it becomes unmanageable. As a guideline, datasets with at least 500 observations and several predictors are generally suitable.
- You may collect data from publicly available repositories or official statistics websites. Some recommended sources are:
  - `https://www.kaggle.com/datasets` (variety of applied datasets).
  - `https://archive.ics.uci.edu/ml/index.php` (classic machine learning datasets).
  - `https://data.gov.in` (open data from the Government of India).
  - World Bank Data: `https://data.worldbank.org`.
  - WHO Global Health Observatory: `https://www.who.int/data/gho`.
- If you wish to use a dataset not from these repositories (for example, data you collected yourself, or from another online source), please check with the instructor first to confirm suitability.
- Be sure to cite the source of your data properly in both your initial and final reports.

Presentation (55 points)

Each team will give a **5 to 10-minute presentation** summarizing the outcomes of their project.

- Presentations will be scheduled during the last few lecture hours. Depending on syllabus coverage, some presentations may be held outside class hours.
- All group members are expected to participate. If you are in a team, decide in advance how speaking roles will be divided to ensure fairness.
- Your presentation should include:
  - A brief description of your dataset and research questions.
  - A summary of the methods used.
  - Key results, illustrated with clear tables or visuals.
  - A short conclusion explaining the main findings and their significance.

- Time management is crucial: presentations that exceed the allotted time will be cut short. Practice in advance to ensure clarity and conciseness.
- Please maintain professionalism and respect during presentations. Listen carefully to other teams and avoid unnecessary disruptions.

Final Report (40 points)

The final report is due on **November 7, 2025**. It should be between **4–6 pages**, including figures and tables. Reports longer than 6 pages will not be accepted.

Your report must include the following elements:

- **Introduction:**
  - Research questions or issues (scientific or statistical).
  - The significance of the problem being studied.
- **Data Description:** Provide details about the dataset and relevant background information.
- **Preliminary Studies:** Include steps such as visualization, dimension reduction, feature extraction, feature selection, model assumption checks (normality, transformations, etc.).
- **Statistical Analysis:**
  - **Methods:** What analyses were done and why. If challenges arose, describe how you addressed them.
  - **Results:** Use a small number of well-designed tables and graphics. Do not copy-paste raw software output.
  - **Conclusion:** Summarize findings in a way that is accessible to a broad audience. Discuss broader implications.
- **Code Submission:** All computer code must be submitted separately in an individual message to the instructor. Do not include code in the report itself.
- **Writing Quality:** Typos and grammatical errors will be heavily penalized. Proofread carefully. If needed, consult *The Elements of Style*.
- **Audience:** Write as if your report will be read by college-educated individuals with only an elementary background in statistics. This is not a report written for your professor.
- **Formatting:** Length: 4–6 pages (including figures and tables). Begin with a clear and informative title highlighting your topic and analysis.