ECO 321 (Fall 2025): Term Project

Guidelines

Instructor: Aneesh Arora

Project Weight: 20% of Final Grade

Objective: The purpose of this project is to give you hands-on experience with **real-world data** and the tools of econometrics. You will:

- Identify an economic or social question of interest. You are free to choose topics from Biology, Chemistry etc provided they allow for econometric analysis. Make sure it is doable in due time and you are able to achieve its objective. Don't get carried away by over-ambition.
- Collect and clean real data.
- Apply **multivariate regression analysis** using Python.
- Interpret results in a meaningful economic context.
- Present your findings with supporting tables, plots, and discussion.

By the end, you should demonstrate that you can:

1. *Work with messy, real datasets.*
2. *Apply econometric methods correctly.*
3. *Draw economically sound conclusions from regression output.*

Project Deliverables

You must submit:

- ***Written Report (6–10 pages, double-spaced):***
  - *Clear statement of your research question.*
  - *Description of dataset and variables.*
  - *Data cleaning and preparation steps.*
  - *Regression models estimated and results.*
  - *Discussion of findings, limitations, and possible extensions.*
  - *Tables/figures summarizing descriptive stats, regressions, and plots.*
  - *Code with results (as a pdf in the end)*

- ***Python Code (Jupyter Notebook or .py file):***
  - *Should include data import, cleaning, analysis, and regression.*
  - *Code must be reproducible (i.e., runs without errors).*
    Each step should be well-documented with comments or markdown.

Format: Groups of 2-3 or Individual (your choice)

Due Date: *December 8, 2025*

Part 1: Project Proposal

Due Date: *October 13, 2025*

Before you begin your analysis, you must submit a one-page proposal that includes:

1. **Research Question:** A clear and concise question you aim to answer (e.g., "What is the effect of education level, work experience, and gender on personal income in the United States?").
2. **Dataset:** A description of the dataset you plan to use, including the source (with a link) and the key variables of interest.
3. **Initial Model:** A preliminary regression equation you plan to estimate, specifying your dependent variable (Y) and key independent variables (X's). For example:
   $wage = \beta_0 + \beta_1 educ + \beta_2 experience + \beta_3 female + \mu$

Part 2: Final Report and Submission

Due Date: *December 8, 2025*

Your final report submission must include all your code, outputs, plots, and written analysis in a clean, organized, and readable format. The notebook should be structured with the following sections:

I. Introduction

- **Motivation:** What is your question? Why is your research question interesting or important? Provide some background context.
- **Research Question:** State your final research question clearly.
- **Hypothesis:** What do you expect to find based on economic theory or intuition? For each key independent variable, state the expected sign of its coefficient (positive or negative) and why.

II. Data Description

- **Source:** Where did you get your data? Provide a direct link.
- Unit of observation (e.g., person, state, country, firm).
- Sample size and time period.
- Key variables (dependent and independent).

III. Data Cleaning & Preparation

- Describe the steps you took to clean the data. Did you handle missing values? Did you remove outliers? Did you create any new variables (e.g., log transformations, dummy variables)?
- **Summary Statistics:** Present a table with summary statistics (mean, standard deviation, min, max, count) for all the variables used in your final model.

IV. Exploratory Data Analysis (EDA)

This section is for visualizing your data *before* running the regression.

- Include at least **two** relevant visualizations. **For example**:
    - A **histogram** to show the distribution of your dependent variable.
    - A **scatter plot** showing the relationship between your dependent variable and one key independent variable.
- Provide a brief interpretation of what these plots tell you about the data and the relationships between variables.

V. Econometric Model & Results

- **Model Specification:** Write out your final multivariate regression equation. Discuss why you included certain variables.
- **Regression Output:** Present the summary output table from your regression analysis (e.g., from `statsmodels`).
- **Interpretation of Results:** This is the most important part of the project.
    - **Coefficient Interpretation:** For each independent variable, interpret its estimated coefficient. Be precise (e.g., "A one-year increase in education is associated with an X% increase in wages, holding all other factors constant.").
    - **Statistical Significance:** Discuss the statistical significance of your coefficients using p-values or t-statistics. Which variables have a statistically significant effect on your dependent variable?
    - **Model Fit:** Interpret the **R-squared** (or Adjusted R-squared) of your model. What does it tell you about how well your model explains the variation in the dependent variable?

VI. Conclusion

- **Summary of Findings:** Briefly summarize your key results. Did the data support your initial hypotheses?
- **Limitations:** What are some potential limitations of your analysis? (e.g., omitted variable bias, small sample size, measurement error).
- **Future Work:** Suggest one or two ways this research could be extended in the future.

VII. Technical Requirements

- **Software**: Python (pandas, numpy, statsmodels, matplotlib/seaborn etc).
- **Plots**: At least 2 relevant visualizations (scatter, regression line, residuals, trends, etc.).
- **Regression**: At least 1 multivariate regression (with ≥ 3 explanatory variables).
- **Reproducibility**: Code should run without errors. Use clear comments/markdown.
- **Academic Integrity**: Cite all data sources. Work must be original.

NOTE: Presentations will be held during the final exam slot if time permits; otherwise, the grade will be based solely on the written submission. You can find the final exam schedule here:
https://www.stonybrook.edu/commcms/registrar/registration/_exams/fall25-finals-2.php

Grading Rubric (100 Points Total)

Component	Points	Description
Introduction & Research Question	15	Clear motivation, well-defined questions, and plausible hypotheses.
Data & Exploratory Analysis	25	Appropriate dataset chosen, cleaning steps are explained, summary statistics and plots are present and correctly interpreted.
Econometric Model	30	Correct model specification, regression is run properly, and the results table is presented clearly.
Interpretation of Results	20	Accurate and detailed interpretation of coefficients, significance, and R-squared. Demonstrates a strong understanding of the findings.
Clarity & Organization	10	The Jupyter Notebook is well-structured, easy to follow, and free of major grammatical errors. Code is clean and commented where necessary.

Suggested Project Ideas & Data Sources

You are free to choose any topic that interests you! Here are a few ideas to get you started:

- **Labor Economics:** Determinants of wages (education, experience, union status).
- **Macroeconomics:** Factors affecting GDP growth (investment, government spending, inflation).
- **Finance:** Predicting stock returns based on company fundamentals (P/E ratio, market cap).
- **Real Estate:** Determinants of housing prices (square footage, number of bedrooms, location).
- **Sports Analytics:** What factors predict a team's winning percentage?

Great Sources for Data:

- **Kaggle:** A platform with thousands of user-uploaded datasets on a wide variety of topics.
- **FRED (Federal Reserve Economic Data):** Excellent source for U.S. macroeconomic and financial data.
- **World Bank Open Data:** Global development data covering hundreds of indicators.
- **U.S. Government Data:** Census Bureau, Bureau of Labor Statistics (BLS), etc.

Tips for Success

- Choose a dataset that excites you—it will make the project more enjoyable.
- Start early: data cleaning often takes longer than expected.
- Don’t chase statistical significance—focus on economic interpretation.
- Ask for help early if you’re stuck with data or coding.
- Remember: **a well-explained, simple regression beats a messy, complicated one.**

Final Submission*:* Upload your **report (PDF)** and **code file** to brightspace by the due date.

Good luck, and please don't hesitate to ask any questions.