Group Assignment Brief
Introduction to Health Analytics

Spring 2025

1 Overview

The group assignment is designed to put the skills you're learning (both conceptual and coding) into practice. You should write a short report (2,000-2,500 words) on any topic you like using the publicly available IPUMS data.

I encourage you to frame your project around a causal question, such as 'What is the impact of X on Y?'. This will allow you to apply the empirical methods covered in the course and practice interpreting results within a causal framework. However, if a causal question is not feasible in your setting, a question about the relationship between two variables (e.g., 'What is the association between X and Y?') is also acceptable and you will not be penalized.

Whatever question you choose, your analysis must include at least one regression model and appropriate statistical tests to explore the relationship between your variables of interest. Be sure to justify your choice of model, interpret the results clearly, and test the robustness of your findings where possible

2 Deadlines

The deadlines are as follows:

- Project proposal (ungraded): January 24, 5pm
- Final project (graded): February 10, 5pm

3 Project proposal

The project proposal should be very short (1 page or less) and include the following:

- **Dataset:** List the IPUMS data series you will be using, including which countries and which years (e.g. IPUMS Global Health DHS Data, Kenya and Uganda, 1997-2010)
- **Question:** State clearly the question you want to answer in 1-2 sentences e.g. 'What is the effect of maternal education on child vaccination rates in East Africa?'
- **Key variables:** State the main variables you will be using e.g. 'A binary variable equal to 1 if a child has received all recommended vaccines'. If you will need to construct a variable, explain how you plan to do this.
- **Empirical strategy:** State clearly how you intend to answer this question in 2-3 sentences e.g. 'Estimate a linear regression of whether a child has received all recommended

vaccines on whether the child's mother completed primary education. Control for the following variables: X, Y, Z'.

4 Report sections

The project should contain the following sections. Marks will be allocated as follows.

- **Introduction** (10% of grade): Ask a clear research question and motivate why it's interesting. Describe related literature. You **do not** need to include a full literature review. Make sure you cite the data correctly.
- **Data description** (15%): Describe the data you use and produce 2 tables or figures containing descriptive statistics. If relevant, explain what survey weights you are using. Explain how you deal with missing data.
- **Empirical strategy** (15%): How are you going to provide convincing evidence relating to your research question?
- **Results** (25%): Present your results, including 1-3 regression tables.
- **Discussion and interpretation** (15%): Interpret the results and discuss how they relate to your research question. Discuss the major limitations and ideas for how these could be addressed with more time, data, funding etc.
- **Code Repository** (20%): At the end of your report, you should include a zipped file containing your code and a link to the public Github repository. It must be possible for the marker to completely replicate the analysis in your report using this repo. Code should be clear and concise. Please include comments. The markers will review the history of the repo to check every member of the group has contributed. **Do not** include the micro-data in the repo, but **DO** include a description of the IPUMS series you are using, including a list of the variables, time period and countries/states, as well as any other sample restrictions you make.

The number of words should be **less than 2,500**, excluding code, repository and references. There are no fixed word counts for individual sections.

5 Data

We strongly suggest that you use the IPUMS data. This is a fantastic resource that facilitates analysis of survey data. You can use any of the IPUMS series, which includes:

- IPUMS-Global Health: Offers harmonized international survey data focused on health, including Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS), which have been administered in many low- and middle-income countries.
- IPUMS-Health: Comprises harmonized data from the U.S. National Health Interview Survey (NHIS) and Medical Expenditure Panel Survey (MEPS), offering insights into health and healthcare trends in the U.S.
- IPUMS International: Offers harmonized international census microdata from hundreds of censuses, allowing for cross-national analysis of population trends.

If you want to use another data source, please contact either the professor or TA ASAP.

6 Prompts

You can look at any question you like! Here are some example questions to give you some inspiration. I encourage you to choose your own question, but if you are stuck you can use one of the below questions:

- **What is the impact of maternal education on child vaccination rates?**
  - Data Source: IPUMS-Global Health DHS surveys
  - Variables: Maternal education level, child vaccination status (e.g., measles vaccination), household socioeconomic indicators.
  - Ideas for empirical strategy: Use a regression model (linear probability model, logit or probit) to regress vaccination on maternal education, controlling for household and regional characteristics. What problems might there be with just using a regression model? Can you think of other ideas for providing convincing evidence e.g. using changes in compulsory schooling laws?
- **How does aging impact out-of-pocket healthcare costs in the United States?**
  - Data Source: IPUMS-Health (MEPS)
  - Variables: Age group, total out-of-pocket medical costs, type of insurance (e.g., Medicare, Medicaid, private).
  - Ideas for empirical strategy: Regression of out-of-pocket medical costs on age. Is this relationship linear? What might be driving it e.g. changes in health status, changes in type of insurance? What does the distribution of out-of-pocket healthcare costs look like?

7 Frequently asked questions

- **Can we combine IPUMS data with other data sources?** You are welcome to do this but make sure you do the following:
  - Check the data use agreement for the IPUMS series you are using to make sure this is not prohibited.
  - Check what geographic variables appear in the IPUMS data you are using. For example, in the NHIS survey we used in the first tutorial there is only a region variable, not state identifiers. So you would not be able to merge in state-level information.
- **Can I use programming languages other than R?** Yes, but we strongly recommend either R or Python. If you want to use something else please talk to me or the TA
- **How many observations should I use?** This is completely up to you. I would recommend using a dataset containing at least 500 people.
- **Am I allowed to write a report on a topic if another paper already exists on the same topic?** Yes, this is totally fine but please appropriately cite the other paper(s) you have found on the same topic. You do not have a lot of time, so we do not expect you to find every paper ever written on your topic!