Topic: the effect of health insurance coverage on health and financial outcomes

Note that for OHIE, a lot of paper choose to focus on both the effect of lottery selection and the effect of health insurance. I think due to time constraint, we should focus on one, but of course we can debate (just need to change the focus of literature review).

Literature Review

To be continued. I think the easiest thing to do is to split them up and scan through them while summarizing:

- Methods employed
- Specific data subset (subsample) used
- Potential problems with randomization
- Specific variables / variation they did not account for (intentional or not)
- List of outcomes (health, economics, others)
- Potential direction/concerns the authors have

Finkelstein et al. (2012):

- Used the 12-month mail survey data and some other administrative data (not publicly available) to estimate LATE associated with Medicaid using 2SLS.

Hattab et al. (2024):

- Used the in-person survey data to estimate CATE using instrumental forest. Found weak heterogeneity in effect across subgroups based on gender, age and race. Supporting materials Section 1.1 may be especially relevant.

Other non-ML papers
Other ML papers: Johnson et al (2022), Denteh and Liebert (2023)

Dataset

The Oregon Health Insurance Experiment (OHIE) is a randomized controlled trial (RCT) that started in 2008, with survey follow-ups until 2010. The public database and documentation, along with STATA replication package for some earlier papers, can be found here. It contains information on all the 74,922 individuals who enrolled in the lottery.

The 'userguide' document does a much better good job than me at introducing the data, so I will be brief. The data is split into 8 files with very few overlapping variables. The ones that concern us the most are 'descriptive_vars,' 'state_programs,' and 'survey12m' or 'inperson.' The first contains demographic information, the second enrollment status in government programs, and the last two health status and insurance needs for selected follow-up participants.

Note that we need to make a choice here. The datasets 'survey12m' (12-month-mail survey) and 'inperson' (in-person survey) both contain information on selected follow-ups approximately

a year from the start of the experiment, but they differ in terms of sample selection, number of observations, timeframe and specific variables. I propose focusing on one due to time constraint, and I recommend the 12-month-mail survey due to it having more observations. That said, the variables in 'inperson' are better-defined, so this is subject to change.

Variables

Covariates

Variables that 'userguide' mentioned the need to control for (part of randomization process):

- *numhh_list*: number of people in the household on the lottery list
- *wave_survey12m*: 12-month-mail survey only, specifies the lottery wave

Variables that Hattab et al. (2024) used:

- Gender: *female_list*
- Age (binned): from *birth_year*
- Race: from a series of indicator variables in 'inperson' (also in 'survey12m')
- High_risk: constructed from a series of pre-lottery indicators in 'inperson'

Other covariates: to be continued. Can start with the ones used in Finkelstein et al. (2012)

Outcomes

Some outcome variables shared across papers. (EM = extensive margin, TU = total utilization)

Self-perceived physical/mental health:

- Current overall happiness: *happiness_12m*
- Overall health: *health_gen_bin_12m, health_chg_bin_12m*
- Bad days: *baddays_phys_12m, 'baddays_ment_12m*
- Limited ability to work: *health_work_12m*
- Depression measures: *dep_interest_12m, dep_sad_12m*

Financial health/strain:

- Out-of-pocket cost: *cost_any_oop_12m, cost_tot_oop_12m*
- Owe money for medical expenses: *cost_any_owe_12m, cost_tot_owe_12m*
- Borrowed money for health expense: *cost_borrow_12m*
- Refused care due to money: *cost_refused_12m*

Health care utilization:

- Prescription medication: *rx_any_12m* (EM), *rx_num_mod_12m* (TU)
- Primary care visits: *doc_any_12m* (EM), *doc_num_mod_12m* (TU)
- ER visits: *er_any_12m* (EM), *er_num_mod_12m* (TU)
- Hospital visits: *hosp_any_12m* (EM), *hosp_num_mod_12m* (TU)

*In-person survey has similar variables.

Potentially other outcomes after lit review.

Justification for ML

- ML in insurance is usually used for prediction, as we can only find a few papers on the application of ML for causal investigation. It will be interesting to continue building this literature.
- More importantly, I think the fact that we have 5+ potential covariates means that classic non-parametric methods no longer satisfy our needs, which pushes for modern ML methods.

Methodology

1. Instrumental forest for conditional LATE (C-LATE):
    - Based on Hattab et al (2024) supporting information. See Supporting Information Section 1.1 and 3.1.
    - They only did 2-dimensional subgroups though (i.e. either age by gender and race by gender). I think it would be interesting 1) to see if we can replicate their results and 2) to take the subgroups one step further and see if that makes a difference
2. DML for LATE (from textbook)
3. DML for missing values (for LATE or for conditional mean function)
    - Inspired by Lecture 14-15 notes and HW 4. The idea behind this is that some people refused to answer the follow-up survey despite repeated requests and monetary incentives. Thus missing values are unlikely to be random, and we should DML to gain consistency.

Questions (should ask asap)

1. For the 3rd method, I am not sure how to go about estimating LATE with DML while accounting for missing values. Is it possible at all to combine the two?
2. For describing our motivation, how to graph propensity score with respect to new covariates while accounting for ones already included? And what is the best way to document heterogeneous treatment effect?
3. In the follow-up surveys, sample selection is down with weights - how to work with these in forests and DML?
4. How to work with censored data?
5. Should we take initial survey data into account (i.e. ‘survey0m’, done right after OHP approval), i.e. instead of measuring absolute magnitude, we should analyze the change in variables? This is giving me pause because the dataset is not exactly a panel.