Effect of Health Insurance Coverage on Health & Financial Outcomes

ECMA 31350 Final Presentation

Timothee Maret
Yasmine Ouattara
Bhavya Pandey
Donna Zheng

March 9, 2024

Introduction

Motivation

Economic theory offers ambiguous predictions for the impact of expanding health insurance on outcomes such as

- health-care use,
- economic well-being, and
- long-term health outcomes.

It is difficult to separate the effects of insurance from confounding factors such as income and initial health in observational studies.

The **Oregon Health Insurance Experiment (2008-2010)** provides for a great opportunity to investigate this through a randomized controlled design.

Setting and Background

- In 2008, Oregon implemented a limited expansion of its Medicaid program for low-income adults through a lottery, selecting names from a waiting list to fill a limited number of available spots. Those selected had the opportunity to apply for Medicaid and to enroll if they met eligibility requirements.
- The lottery can be used to assess both the effects of lottery selection and the effects of Medicaid coverage itself on a range of outcomes.
- For this project, we decided to study the affect of **health insurance coverage** itself, upon some **outcomes related to health and financial well-being**.

Research Gap → Need for ML → Our Approach

1. ML in health insurance literature is usually used for prediction, and in an experimental context, we could only find a few (very recent) papers on the application of ML for causal investigation. It will be interesting to continue building this literature.
2. Additionally, we have 5+ potential covariates implying that classic non-parametric methods no longer satisfy our needs, which pushes for modern ML methods.

In this project, we:

- Investigate the conditional local ATE using instrumental/causal forest;
- Then calculate the local ATE, in general, using DML while taking into account missing values from follow-up surveys through imputation.

Literature Review

Literature Review

1. 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.
2. 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.
3. Goto et al. (2024)
   - Used the in-person survey data and causal forest approach to detect heterogeneous effects of Medicaid coverage on depression. Found reduced the risk of depression and greater impacts for populations at-risk at baseline.

Data

Data

- The Oregon Health Insurance Experiment (OHIE) is a randomized controlled trial (RCT) that started in 2008, with survey follow-ups until 2010.
- It contains information on all the 74,922 individuals who enrolled in the lottery.
- We use the the 12-month-mail survey post enrollment due to it having more observations as well as relevant follow up variables for our outcomes.
- We also consider self-reported variables for demographic information, enrollment status in government programs, and health status and insurance needs, for follow-up participants.

Method and Analysis

Method I: Motivating Causal Forest DML

Gradient Boosting

Given the high-dimmensional data, we used gradient boosting via XGBoost to identify the most important features.

- Reduce dimension of data $\rightarrow$ computationally efficient analysis.
- We selected features based on a F-score threshold of 300.

Implementation Approach

Balance, HTE, NAs and Imputation

- The covariates chosen are not perfectly balanced.
- There are heterogenous effects.
- Imputed data using K-NN : Had little effect on characteristics of the data

Next: We can further provide sample weights, using propensity scores, in the CausalForestDML approach.

Method I: Causal Forest DML Model

First stage:

- Random Forest Classifier to model treatment, and Random Forest Regressor for the outcomes (cross-fitting).

Second stage:

- The treatment effects are estimated using the residuals obtained from the first stage.

Solves

$$E[(Y - E|X] - \langle\theta(x), T - E[T|X]\rangle - \beta(x))(T; 1)|X = x] = 0$$

We used heterogeneity score as the criterion which finds splits that maximize the pure parameter heterogeneity score.

Method I: Results Imputed Data

Variable	ATE	SE	p-value	Significant
happiness	-0.05143	0.00424	5.88E-11	TRUE
health (overall)	-0.00624	0.00310	0.05721	FALSE
health (change)	-0.02023	0.00255	9.54E-08	TRUE
bad days (physical)	0.79423	0.07458	6.37E-10	TRUE
bad days (mental)	0.44775	0.08063	1.64E-05	TRUE
limits ability to work	0.09005	0.00302	0	TRUE
disinterest	0.05536	0.00666	4.43E-08	TRUE
felt sad/depressed	0.04394	0.00617	5.04E-07	TRUE
any OOP health costs	-0.09862	0.00321	0	TRUE
total OOP health costs	-2656.39	453.36	8.14E-06	TRUE
owe for medical expenses	-0.05911	0.00296	3.77E-15	TRUE
borrowed for medical expenses	-0.09182	0.00279	0	TRUE
refused care due to money	-0.02025	0.00194	9.32E-10	TRUE
on medication	0.08994	0.00246	0	TRUE
number of meds	0.56342	0.01726	0	TRUE
primary care visits	0.19769	0.00298	0	TRUE
number of visits (PHC)	1.16298	0.02110	0	TRUE
any ER visits	0.06134	0.00279	4.44E-16	TRUE
number of visits (ER)	0.16585	0.00800	1.78E-15	TRUE
any hospital visits	0.06183	0.00250	0	TRUE
number of visits (hosp)	0.10667	0.00527	3.11E-15	TRUE

Method II: Motivating DML Approach for LATE with NA

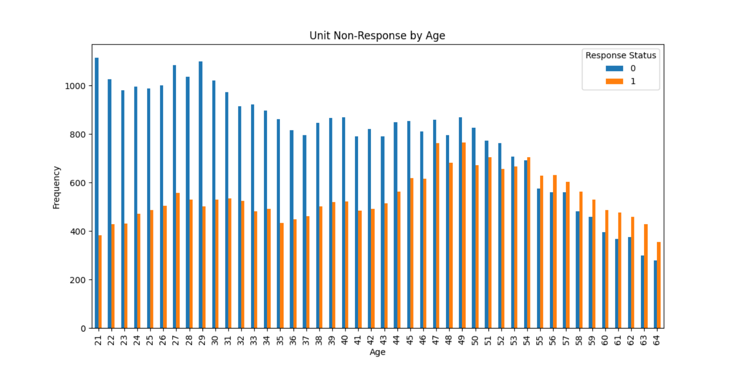
Quote from Finkelstein et al. (2012):

"The survey data [...] with a 50% effective response rate, are subject to potential nonresponse bias [...] These outcomes are only available for individuals who responded to the mail survey and may therefore not be representative of the full sample."

Problem: This is concerning because we do not have non-response at random!

- item non-response (controlled by K-NN imputation)
- unit non-response

Method II: Motivating DML Approach for LATE with NA



- Observed similar imbalances for gender, protocol intensity, access to phone, existing governmental programs, etc., all of which affect Y!

Question: Can we address this non-response bias in estimating LATE using machine learning?

Method II: Theorem

Set-up: Let

- Y - outcome (health & financial outcomes),
- D - treatment (insurance coverage),
- Z - instrument (lottery selection),
- W - covariates such that Z is a valid instrument,
- S - responder status,
- X - covariates such that missing at random, i.e. $(Y, D, Z, W) \perp S \mid X$, is plausible.

Then this score function is Neyman orthogonal:

$$g(A; \tau_{LATE}, \eta) = (g^{1,1;o}(A;\eta) - g^{1,0;o}(A;\eta)) - (g^{2,1;o}(A;\eta) - g^{1,1;o}(A;\eta))\tau_{LATE},$$

Method II: Theorem

where:

$$
\begin{aligned}
g^{1,1;o}(A; r, q, a_1, b_1) &= \frac{YZS}{rq} - (Z - r) \cdot \frac{a_1}{r^2} - (S - q) \cdot \frac{b_1}{rq}, \\
g^{1,0;o}(A; r, q, a_0, b_0) &= \frac{Y(1 - Z)S}{(1 - r)q} - (Z - r) \cdot \frac{a_0}{(1 - r)^2} \\
&\quad - (S - q) \cdot \frac{b_0}{(1 - r)q}, \\
g^{2,1;o}(A; r, q, f_1, g_1) &= \frac{Y(1 - Z)S}{rq} - (Z - r) \cdot \frac{f_1}{r^2} - (S - q) \cdot \frac{g_1}{rq}, \\
g^{2,0;o}(A; r, q, f_0, g_0) &= \frac{Y(1 - Z)S}{(1 - r)q} - (Z - r) \cdot \frac{f_0}{(1 - r)^2} \\
&\quad - (S - q) \cdot \frac{b_0}{(1 - r)q},
\end{aligned}
$$

Method II: Theorem

and the nuisance parameters $\eta = (r, q, a_1, a_0, b_1, b_0, f_1, f_0, g_1, g_0)$ are:

$$
\begin{aligned}
r(W) &= \mathbb{E}[Z \mid W] = \mathbb{P}[Z = 1 \mid W], \\
q(X) &= \mathbb{E}[S \mid X] = \mathbb{P}[S = 1 \mid X] = \mathbb{P}[S = 1 \mid W, X] = q(X, W), \\
a_1(W) &= \mathbb{E}[YZ \mid S = 1, W], \quad a_0(W) = \mathbb{E}[Y(1 - Z) \mid S = 1, W], \\
b_1(X, W) &= \mathbb{E}[YZ \mid S = 1, X, W], \quad b_0(X, W) = \mathbb{E}[Y(1 - Z) \mid S = 1, X, W], \\
f_1(W) &= \mathbb{E}[DZ \mid S = 1, W], \quad f_0(W) = \mathbb{E}[D(1 - Z) \mid S = 1, W], \\
g_1(X, W) &= \mathbb{E}[DZ \mid S = 1, X, W], \quad g_0(X, W) = \mathbb{E}[D(1 - Z) \mid S = 1, X, W].
\end{aligned}
$$

Method II: Results

	(1)	(2)	(3)	(4)	(5)	(6)
	Healthcare Utilization					
prescription drugs	0.088	-0.119	-0.121	0.347	-0.425	-0.436
outpatient visits	0.212	0.079	0.077	1.083	0.510	0.491
ER visits	0.022	-0.084	-0.084	0.026	-0.163	-0.165
hospital visits	0.008	-0.014	-0.015	0.021	-0.004	-0.004
	Financial Strain					
OOP medical expense	-0.200	-0.328	-0.332			
owe medical expense	-0.180	-0.390	-0.391			
borrowed money for medical expense	-0.154	-0.260	-0.263			
refused treatment due to money	-0.036	-0.070	-0.070			
	Health					
unhappiness	-0.191	-0.643	-0.649			
health (overall)	-0.133	-0.273	-0.278			
health (change)	-0.113	-0.208	-0.210			
bad days (physical)	-1.317	-4.563	-4.624			
bad days (mental)	-2.082	-5.464	-5.554			
limits ability to work	-1.585	-0.210	-0.213			

Columns (1)-(3) for extensive margin (any), (4)-(6) for total utilization. Columns (1) & (4) correspond to Finkelstein et al. (2012) LATE estimates, (2) & (5) correspond to the propensity score $p(X)$, (3) & (6) correspond to the propensity score $p(X, W)$.

Key Takeaways

For **Method I: CausalForestDML** with imputed values:

- Results are significant, effects align with previous analyses though diminished, however – hyperparameter tuning required to check for robustness and interpretability of effects

For **Method II: LATE with missing values**:

- columns (2) & (3) and (5) & (6) are similar $\implies$ some evidence that the independence assumption holds
- compared to Finkelstein et al. (2012):
  - our LATE estimates are diminished or even opposite in sign compared to estimates for healthcare utilization
  - but report greater reduction on financial strain and negative self-perceived health outcomes
  - formal interpretation will have to wait after obtaining SE

Robustness Checks/Next Steps

1. Checking for treatment and control balance at the baseline
2. Using simple propensity score matching + sample weights from survey to check for covariate balance in Method I
3. For DML for missing values:
   - Generate bootstrap standard errors
   - Ensuring item non-response imputation works properly
   - Use the alternative treatment variable used in Finkelstein et al. (2012) to see how well it aligns with current analysis as well as past results
4. Checking for/comparing results for other outcome variables corresponding to each treatment (lottery vs coverage)
5. Checking for/comparing results for 25-month survey responses

Appendix

Method I: Results Drop NA

Variable	ATE	SE	p-value	Significant
happiness_12m	-0.05143	0.00424	5.88E-11	TRUE
health_gen_bin_12m	-0.00624	0.00310	0.05721	FALSE
health_chg_bin_12m	-0.02023	0.00255	9.54E-08	TRUE
baddays_phys_12m	0.79423	0.07458	6.37E-10	TRUE
baddays_ment_12m	0.44775	0.08063	1.64E-05	TRUE
health_work_12m	0.09005	0.00302	0	TRUE
dep_interest_12m	0.05536	0.00666	4.43E-08	TRUE
dep_sad_12m	0.04394	0.00617	5.04E-07	TRUE
cost_any_oop_12m	-0.09862	0.00321	0	TRUE
cost_tot_oop_12m	-2656.39	453.36	8.14E-06	TRUE
cost_any_owe_12m	-0.05911	0.00296	3.77E-15	TRUE
cost_tot_owe_12m	2630427262	385992093.3	9.73E-07	TRUE
cost_borrow_12m	-0.09182	0.00279	0	TRUE
cost_refused_12m	-0.02025	0.00194	9.32E-10	TRUE
rx_any_12m	0.08994	0.00246	0	TRUE
rx_num_mod_12m	0.56342	0.01726	0	TRUE
doc_any_12m	0.19769	0.00298	0	TRUE
doc_num_mod_12m	1.16298	0.02110	0	TRUE
er_any_12m	0.06134	0.00279	4.44E-16	TRUE
er_num_mod_12m	0.16585	0.00800	1.78E-15	TRUE
hosp_any_12m	0.06183	0.00250	0	TRUE
hosp_num_mod_12m	0.10667	0.00527	3.11E-15	TRUE