Pursuant Methods

Thomas Hsiao

2024-08-02

Methods flow

1. Merge in external data (census data, covariates)
2. Clean variables and exclude unreasonable values
3. **Deduplicate sessions**
4. **Adjust outcome definition for unknown controlled hypertension**
5. **Fit unit-level regression adjusting for non-representativeness**
6. Poststratification of estimates (county, state, national)
7. Compare to BRFSS and NHANES

Deduplication

What we’ve tried

- Pursuant created a “pseudo-ID” for us to use.
- ID: DOB, age, sex, race, ethnicity, and kiosk address.
- Pursuant reports 94% concordance between Pseudo ID and account ID
- We aggregated by taking mean SBP and DBP over all measurements in a 2-yr range (2017-2018, 2019-2020, etc.)
- However, sessions decreased from ~80 million to ~40 million

Example BP for a pseudo-ID in 2023-2024

Solutions

- Make use of session time along with the pseudo ID (aggregate if same account/pseudo ID and multiple measurements made within 5 min of each other)
- Increase the number of time units we estimate (instead of deduplicating over 2 year ranges, only do so for every single year or 6 months)

Adjustment for controlled hypertension

Outcome definition

Hypertension definition: anyone who has either

1. SBP $\geq$ 140 mm Hg
2. DBP $\geq$ 90 mm Hg
3. Has been previously diagnosed or is taking medication for hypertension.

Pursuant dataset

Only ~1 million out of ~85 million sessions responded to “Have you ever been diagnosed with high blood pressure?” question.
We will underestimate hypertension if we ignore “controlled hypertension” (those who satisfy (3) only)

Solutions

Estimate $P(controlled|Z, SBP < 140, DBP < 90)$ where Z are individual-level covariates (age, race, sex, etc.). Simulate this probability at individual level, and decide whether each observation is controlled or not.
How to estimate this probability model?

1. The validation set of ~1 million diagnosis questions
2. NHANES microdata
3. Both?

Non-representativeness

Table 1

Our baseline demographics match the national very well (but may change depending on the de-duplication and unit of analysis).

Table 1: Baseline characteristics of Pursuant users (pseudo-ID deduplication)

Multilevel regression and poststratification (MRP)

Let p_i be the probability of hypertension for individual i.

$$\begin{aligned}\operatorname{logit}(p_i) = \gamma_0 + \alpha^{\text{county}}_{\text{c[i]}} + \alpha^{\text{state}}_{\text{s[i]}} + \alpha^{\text{age group}}_{\text{a[i]}} + \alpha^{\text{race}}_{\text{r[i]}} + \\ \beta^{\text{male}}\text{Male}_i + \alpha^{\text{race}\times\text{male}}_{\text{r[i], g[i]}} + \gamma_1^\top Z^{\text{county}} + \gamma_2^\top Z^{\text{state}}\end{aligned}$$

Poststratification

We have the following individual-level poststratification variables available in the Pursuant dataset.

- Age (single-year)
- Sex (2 levels)
- Race/ethnicity (5 levels)

Poststratification (cont.)

To post-stratify to any geographic level or demographic group we compute

$$\hat{\theta} \ := \frac{\sum N_j \hat{\theta}_j}{\sum N_j}$$

where j is a post-stratification cell, and N_j is the number of people in the population in that cell at the county, state, or national level. $\hat{\theta}_j$ is the model estimate of hypertension prevalence in that cell.

Covariates for county and state level

These should be predictive of hypertension prevalence.

We do not have data in Massachusetts (legal issues).

Good county/state level covariates will be important in data sparse areas.

Spatiotemporal modeling

Space-time models can help further if data sparsity is an issue at the small area level. However considering how much individual data and good geographic coverage we have, we can try MRP and observe spatial residuals.

Conclusion

Three main areas of attention

1. De-duplication of sessions
2. Adjustment for controlled hypertension in outcome definition
3. Adjustment for non-representativeness of the Pursuant users

Will write up the methods and send for review.

Questions?