Propensity Score Methods with Multilevel Data

March 19, 2014

Multilevel data

- Data in medical care, health policy research and many other fields are often multilevel.
- Subjects are grouped in natural clusters, e.g., geographical area, hospitals, health service provider, etc.
- Significant within- and between-cluster variations.
- Ignoring cluster structure often leads to invalid inference (1) Inaccurate standard errors, (2) Cluster-level effects could be confounded with individual-level effects.
- Hierarchical/Multilevel regression models provide a unified framework to study multilevel data.

Multilevel Data Example: Racial Disparity in Health Care

- Disparity: racial differences in care attributed to operations of health care system
- Breast cancer screening data in different health insurance plans are collected from the Centers for Medicare and Medicaid Services (CMS).
- Two-level structure: Level 1 – patients; Level 2 – insurance plans
- Data: 64 plans with a total sample size of 75,012

- Patients information: age, race, eligibility for medicaid, etc.

- Plan information: non/for profit status, practice model, geo code

- Goal: study the disparity in getting breast cancer screening between Whites and Blacks

Causal vs. Unconfounded Descriptive Comparison

- In the race disparity application, “race” is not manipulable, thus is not “cause” under the Rubin Causal Model.
- Indeed, the goal is not to study the “effect” of being a White/Black, but rather a descriptive comparison between two groups with similar background.
- Can we should use propensity score techniques?
- Yes! Remember the 1st property of propensity score is balancing: $\mathbf{W \perp X | e(X)}$, which has nothing to do with potential outcomes.
- It is important to differentiate between causal and “unfounded descriptive” comparisons.

Unconfounded (controlled) descriptive comparisons

- “Assignment": a nonmanipulable state defining membership in one of two groups. For example, different races, different years
- Objective: an unconfounded comparison of the observed outcomes between the groups
- Estimand: average controlled difference (ACD)

$$ACD = E\_\mathbf{x}[E(Y|\mathbf{X},Z = 1) - E(Y|\mathbf{X}, Z = 0)]$$

- The difference in the means of Y in two groups with balanced covariate distributions.

Causal Comparisons

- Assignment: a potentially manipulable intervention.
- Objective: causal effect - comparison of the potential outcomes under treatment versus control in *a common set of units*
- Estimand: average treatment effect (ATE)

$$ATE=E[Y(1)-Y(0)]$$

- Examples: evaluating the treatment effect of a drug, therapy or policy for a given population
- Under the assumption of “nonconfoundedness”, ACD=ATE.

Propensity Score: Recap

- Propensity score: $e(x)=Pr(W=1|X)$.
- Balancing property: balancing propensity score also balances the covariates of different groups.
- Using propensity score - two-step procedure:
- Step 1: estimate the propensity score, e.g., by logistic regression.
- Step 2: estimate the “treatment” effect by incorporating (matching, weighting, stratification, etc.) the estimated propensity score.

Propensity Score Methods for Multilevel Data

- Propensity score has been developed and applied in cross-sectional settings (single level data).
- How to extend the propensity score methods to multilevel data?
- Two central questions

1. Whether and (if true) how to incorporate multilevel structure into the modeling for propensity score?
2. Whether and (if true) how to incorporate multilevel structure into the estimation of ATE or ACD?

- Two relevant papers: (1) Matching (Arpino and Mealli, 2011), (2) Weighting (Li, Zaslavsky, Landrum, 2013)
- Crucial method: multilevel model.

Notations

Setting: (1) two-level structure, (2) treatment assigned at individual level.

- Cluster: $h = 1, ..., H$.
- Cluster sample size: $k = 1, ..., n_h$; total sample size: $n = \sum_h n_h$.
- "Treatment" (individual-level): Z_{hk}, 0 control, 1 treatment.
- Covariates: $\mathbf{X}_{hk} = (\mathbf{U}_{hk}, \mathbf{V}_h)$, $\mathbf{U}_{hk}$ individual-level; $\mathbf{V}_h$ cluster-level.
- Outcome: Y_{hk}.

Step 1: Propensity score models

- Marginal model-ignoring multilevel structure:

$$\text{logit}(e_{hk}) = \delta_0 + \mathbf{X}_{hk}\alpha.$$

- Fixed effects model - adding cluster-specific main effect δ_h:

$$\text{logit}(e_{hk}) = \delta_h + \mathbf{U}_{hk}\alpha.$$

  - Key: the cluster membership is a nominal covariate.
  - δ_h absorbs the effects of both observed and unobserved cluster-level covariates $\mathbf{V}_h$.
  - Estimates a balancing score without knowledge of $\mathbf{V}_h$, but might lead to larger variance than the propensity score estimated under a correct model with fully observed $\mathbf{V}_h$.
  - The Neyman-Scott problem: inconsistent estimates of δ_h, α given large number of small clusters.

Step 1: Propensity score models

- Random effects model:

$$\text{logit}(e_{hk}) = \delta_h + \mathbf{X}_{hk}\alpha, \quad \text{with } \delta_h \sim \text{N}(\delta_0, \sigma_\delta^2).$$

  - Due to the shrinkage of random effects, $\mathbf{V}_h$ need to be included.
  - Pros: "borrowing information" across clusters, works better with many small clusters.
  - Cons: produce a biased estimate if δ_h are correlated with the covariates.
  - Does not guarantee balance within each cluster.
  - Random slopes can be incorporated.

Multilevel regression models: Fixed effects and Random Effects Models

- Fixed effects: specify a different intercept for each cluster (dummy variable for cluster membership).
- Number of parameters increase with the number of clusters. When there is a large number of small clusters, estimates can be biased.
- Random effects: also specify a different intercept for each cluster, but assume these intercepts across clusters follow a distribution.
- More parsimonious, borrow strength across clusters. No balancing within each cluster.
- Random effects models can easily fitted with build-in packages "**lme4**" in R.

Step 2: Estimating ACD or ATE

- Weighting (Li, Zaslavsky, Landrum, 2013)
- Foundation: Horvitz-Thompson (inverse probability) weighting

$$w_{hk} = \begin{cases} \frac{1}{e(\mathbf{X}_{hk})}, & \text{for } Z_{hk} = 1 \\ \frac{1}{1-e(\mathbf{X}_{hk})}, & \text{for } Z_{hk} = 0. \end{cases}$$

- How to weight? Two choices:

1. Weighted average across clusters and individuals.

2. First calculate weighted average within a cluster and then calculate average of the cluster averages.

- Which one is better?

Weighting Estimators

- Marginal estimator - ignore clustering:

$$\hat{\pi}^{\text{ma}} = \sum_{Z_{hk}=1} \frac{Y_{hk} w_{hk}}{w_1} - \sum_{Z_{hk}=0} \frac{Y_{hk} w_{hk}}{w_0},$$

where w_{hk} is the HT weight using the estimated propensity score, and $w_z = \sum_{h,k:Z_{hk}=z} w_{hk}$ for $z = 0, 1$.

- Cluster-weighted estimator: (1) obtain cluster-specific ATE; and (2) average over clusters:

$$\hat{\pi}^{\text{cl}} = \frac{\sum_h w_h \hat{\pi}_h}{\sum_h w_h}.$$

where

$$\hat{\pi}_h = \frac{\sum_{k \in h}^{Z_{hk}=1} Y_{hk} w_{hk}}{w_{h1}} - \frac{\sum_{k \in h}^{Z_{hk}=0} Y_{hk} w_{hk}}{w_{h0}},$$

where $w_{hz} = \sum_{k \in h}^{Z_{hk}=z} w_{hk}$ for $z = 0, 1$.

Rule of Thumb

- Simulations and analytical results in Li et al. show that “ignoring cluster structure in both stages of propensity score methods is a very bad idea”.
- You should take into account of clustering in “at least one of the two stages” and “preferably in both stages”.

Step 2: Estimating ACD or ATE

- Matching (Arpino and Mealli, 2011)
- Two choices:

(1) Matching within clusters: maybe hard to find matches with smaller clusters.

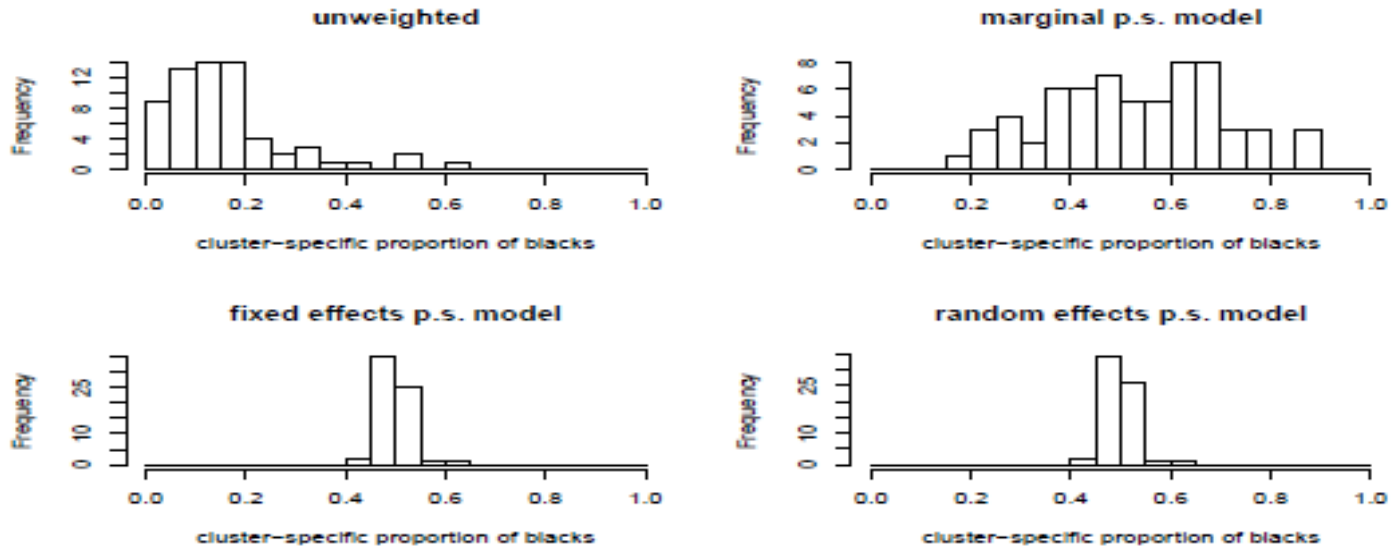
(2) Matching across clusters: more flexible.

- Simulations show similar suggestions as the weighting: it is important to take into account the clustering structure in at least one of the two stages. In particular, it is okay to do matching across clusters when clustering structure is taken into account in the propensity score model.

Real Application: Disparity in Breast Cancer Screening

Balance Check: different propensity score models

Figure : Histogram of cluster-specific weighted proportions of black enrollees.



All p.s. models suggest: living in a poor neighborhood, being eligible for Medicaid and enrollment in a for-profit insurance plan are significantly associated with black race.

Results

- Average controlled difference in percentage of the proportion of getting breast cancer screening between blacks and whites

	weighted	
	marginal	clustered
marginal	-4.96 (.79)	-1.73 (.83)
fixed	-2.49 (.92)	-1.78 (.81)
random	-2.56 (.91)	-1.78 (.82)

- All estimators show the rate of receipt breast cancer screening is significantly lower among blacks than among whites with similar characteristics.
- Ignoring clustering in both stages doubled the estimates from analyses that account for clustering in at least one stage.
- Between-cluster variation is large.