A New Semiparametric Profile Likelihood Approach for Biased Sampling Studies with Applications

Syracuse University
Office of Undergraduate Research & Creative Engagement

AiJing Wu

(feel free to contact me at: awu117@syr.edu)

WiSE
WOMEN IN SCIENCE AND ENGINEERING
SYRACUSE UNIVERSITY

INTRODUCTION

Biased Sampling

Definition:

A sampling method is called biased if it systematically favors some outcomes over others.

Applications:

Fields –

Social Sciences, Medicine, Rare Diseases, Rare Events, Epidemiology and Public Health, etc.

Topics –

Case and Control Studies, Causal Inference, Missing Data Problems, Case Cohort Studies, Exponential Tilting Genetic Mixture Models, etc.

Logistic Regression

Advantage:

Straightforward when handling multiple genetic variants

Disadvantage:

Not efficient as it fails to exploit the gene-environment independence assumption

Semiparametric Profile Likelihood

Semiparametric Model:

Parametric – finite-dimensional – understandable

Non-parametric – infinite-dimensional – can be manipulated while still offering a fair representation of the messiness that is involved in real life

Profile Likelihood:

For models with high–dimensional parameter spaces

METHOD & THEORY

In rare disease studies ($Pr(D = 1) < 5\%$), we have case-control observations ($D_i = 0, X_i, Z_i$) for $i = 1, \dots, n_0$ and

($D_i = 1, X_i, Z_i$) for $i = n_0 + 1, \dots, n = n_0 + n_1$.

$D = 1$: presence, $D = 0$: absence,

X: genetic factors, Z: environmental risk factors.

The disease occurrence model

$$h(d_i, x_i, z_i) = \frac{e^{(d(\alpha + m(x_i z_i \beta)))}}{1 + e^{(\alpha + m(x_i z_i \beta))}},$$

where $m(.)$ is

$$m(x_i, z_i, \beta) = \beta_x x_i + \beta_z z_i + \beta_{xz} x_i z_i.$$

Goal:

Estimate α and β efficiently

Challenges:

1. Data sample is not random
2. Do not know the distribution of X or Z or both

Current Practices & Their Limitations:

1. Treat data as they were randomly collected
2. Make strong distribution assumptions
3. Require disease rate
4. Biased or misleading results
5. Estimators are not efficient

New Method:

Semiparametric Profile Likelihood

Advantages –

1. View the biased sample as they were from a hypothetical population
2. Data are a random sample from the hypothetical population
3. Do not require distribution assumption on X or Z
4. Do not require disease rate
5. Estimate α and β efficiently

Super Solver Algorithm:

$\theta = [\alpha, \beta_x, \beta_z, \beta_{xz}]$, γ: density of X, ξ: density of Z, π_1: disease rate

(a) Set initial values: $\tilde{\theta} = \theta_{LR}, \tilde{\gamma}_j = \frac{X_j}{\sum X}, \tilde{\xi}_k = \frac{Z_k}{\sum Z}$

(b) Calculate $\tilde{\pi}_1 = \sum_{k=1}^{n} \xi_k \{\sum_{j=1}^{n} \gamma_j h(1, x_j, z_k)\}, \tilde{\pi}_0 = 1 - \tilde{\pi}_1$

(c) Update γ and ξ with the following equations, denoted as $\hat{\gamma}$ and $\hat{\xi}$

$$\frac{n_1}{\pi_1} + \left(\frac{n_0}{\pi_0} - \frac{n_1}{\pi_1}\right) \sum_{k=1}^{n} \xi_k h(0, x_l, z_k) = \frac{1}{\hat{\gamma}_l}$$

$$\frac{n_1}{\pi_1} + \left(\frac{n_0}{\pi_0} - \frac{n_1}{\pi_1}\right) \sum_{j=1}^{n} \gamma_j h(0, x_j, z_s) = \frac{1}{\hat{\xi}_s}$$

(d) Check if $\hat{\gamma}$ and $\hat{\xi}$ both sum to 1, and Update $\tilde{\pi}_1$ with them

(e) Update θ by solving the following equation, denoted by $\hat{\theta}$

$$\sum_{i=1}^{n} \left\{ \frac{\partial h(d_i, x_i, z_i)/\partial\theta}{h(d_i, x_i, z_i)} - \frac{\sum_{k=1}^{n}\sum_{j=1}^{n} \gamma_j \xi_k \partial h(d_i, x_i, z_i)/\partial\theta}{\sum_{k=1}^{n}\sum_{j=1}^{n} \gamma_j \xi_k h(d_i, x_i, z_i)} \right\} = 0$$

(f) Repeat (b) to (e) until $\hat{\theta}$ converges

(g) Calculate final $\hat{\pi}_1$

RESULTS

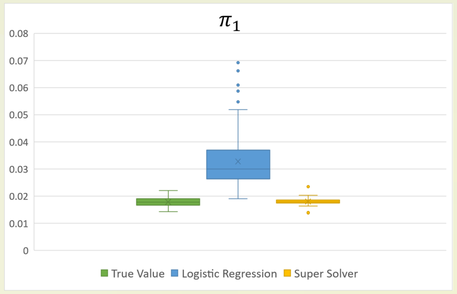
Simulation Settings:

$Simulation\ Rounds = 200,$

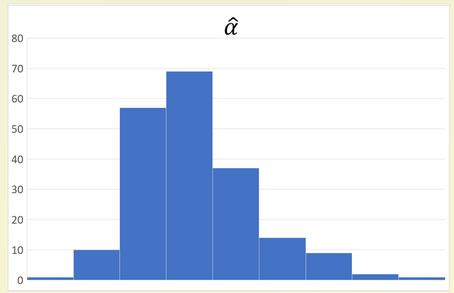
$n_0 = n_1 = 100,\ n = 200,$

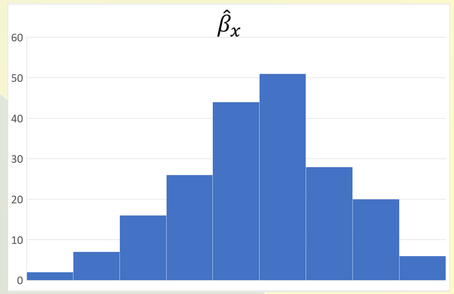
$X \sim Norm(0, 1),\ Z \sim Norm(0, 1),$

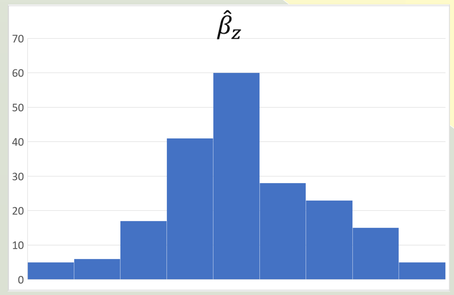
$True\ \alpha = -4.165,\ True\ \beta = [log(1.5), log(1.2), log(1.3)].$

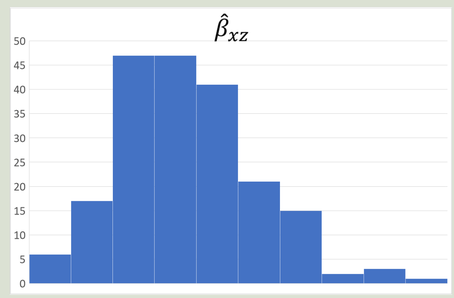


	Logistic Regression				Super Solver			
VAR	$\hat{\alpha}$*	$\hat{\beta}_x$	$\hat{\beta}_z$	$\hat{\beta}_{xz}$	$\hat{\alpha}$	$\hat{\beta}_x$	$\hat{\beta}_z$	$\hat{\beta}_{xz}$
TRUE	-4.1650	0.1761	0.0792	0.1139	-4.1650	0.1761	0.0792	0.1139
MEAN	-4.1635	0.4300	0.1978	0.2728	-4.1011	0.4339	0.1972	0.2716
BIAS	0.0015	0.2539	0.1186	0.1589	0.0639	0.2578	0.1180	0.1576
STD	0.0968	0.1458	0.1606	0.1601	0.0030	0.1198	0.1548	0.0842
MSE	0.0094	0.0862	0.0401	0.0511	0.0041	0.0812	0.0381	0.0321
EFF	2.2901	1.0607	1.0524	1.5932		*LR only estimates κ		









CONCLUSIONS & CONTRIBUTONS

Theoretical Contribution:

1. Work directly with biased-sampled data
2. Assumption free (distribution and modelling)
3. Do not require disease rate
4. Estimators are efficient

Accuracy:

1. Significant improvement in estimating $\hat{\alpha}$ and $\hat{\beta}_{xz}$
2. Improvement in estimating $\hat{\beta}_x$ and $\hat{\beta}_z$

Computing Time for the code package:

2300s VS 5500s per simulation

FUTURE WORK

1. Implement this method for various data structures
   eg. Five Correlated Single Nucleotide Polymorphisms, Discrete Environment Variables
2. Derive the asymptotical properties of β

REFERENCES

Chatterjee, N., Kalaylioglu, Z. & Carroll, R. J. (2005). Exploiting gene-environment independence in family-based case-control studies: Increased power for detecting associations, interactions and joint effects. *Genet. Epidemiol.* 28, 138–156.

Stalder, O., Asher, A., Liang, L., Carroll, R. J., Ma, Y. & Chatterjee, N. (2017). Semiparametric analysis of complex polygenic gene-environment interactions in case-control studies. *Biometrika*. 104(4), 801–812.

ACKNOWLEDGEMENT

This research is guided by Dr. Jianxuan Liu (jliu193@syr.edu) and is funded by SOURCE and WiSE.