MR4571186 62G05 62D05 62G20

Che, Menglu (1-YALE-DBH)**; Han, Peisong** (1-MI-BPH)**;**
Lawless, Jerald F. (3-WTRL-S)

Improving estimation efficiency for two-phase, outcome-dependent sampling studies. (English. English summary)

Electron. J. Stat. **17** (2023), *no.* 1, 1043–1073.

When some covariates are expensive or difficult to measure, two-phase outcome dependent sampling (ODS) is often employed. To make the argument more concrete, denote the outcome, inexpensive-to-measure covariates, and expensive-to-measure covariates by Y, $\mathbf{X}$, and $\mathbf{Z}$, respectively. In two-phase ODS, $(Y, \mathbf{X})$ are taken from all individuals in Phase 1, and then $\mathbf{Z}$ is taken from selected individuals in Phase 2.

In this article, the authors study the problem of estimating the parameter $\boldsymbol{\beta}$ in a parametric model of the conditional density $f(Y|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$ under two-phase ODS. The conditional maximum likelihood (CML) method is popularly chosen, because it does not need to model $f(\mathbf{Z}|\mathbf{X})$ unlike the full maximum likelihood method and it can handle zero selection probabilities in Phase 2. However, CML leads to efficient loss due to discarding the information on those who do not enter Phase 2.

To pursue efficiency gain over CML, the authors summarize the information on Phase 1 through a parametric model of the conditional density $f(Y|\mathbf{X}; \boldsymbol{\theta})$. More specifically, they investigate the maximum empirical likelihood (MEL) estimation of $\boldsymbol{\beta}$, where the information is explicitly incorporated as some zero expectation restriction implied by this model, as well as the zero score restriction from CML.

Their analysis starts from the case with strictly positive selection probabilities in Phase 2 for all individuals. Under the oracle scenario where the selection probabilities are known, the MEL estimator for $\boldsymbol{\beta}$ is shown to attain efficiency gain over CML in the sense that the asymptotic variance of the former is smaller than the one for the latter. The analysis is further extended to the cases where positive selection probabilities are parametrically modeled (and estimated) and zero selection probabilities are allowed. Efficiency gain of MEL over CML in finite samples is confirmed via Monte Carlo simulations, and a real data analysis from the National Health and Nutrition Examination Survey is also conducted. *Masayuki Hirukawa*

[**References**]

1. Baksi, A. J., Treibel, T. A., Davies, J. E., Hadjiloizou, N., Foale, R. A., Parker, K. H., Francis, D. P., Mayet, J. and Hughes, A. D. (2009). A meta-analysis of the mechanism of blood pressure change with aging. *Journal of the American College of Cardiology* **54** 2087–2092.
2. Barnett, I. J., Lee, S. and Lin, X. (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic Epidemiology* **37** 142–151.
3. Bjørnland, T., Bye, A., Ryeng, E., Wisløff, U. and Langaas, M. (2018). Powerful extreme phenotype sampling designs and score tests for genetic association studies. *Statistics in Medicine* **37** 4234–4251. MR3879425 MR3879425
4. Breslow, N. and Cain, K. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75** 11–20. MR0932812 MR0932812
5. Breslow, N. E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59** 447–461. MR1440590 MR1440590
6. Breslow, N. E. and Holubkov, R. (1997). Weighted likelihood, pseudolikelihood

and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* **16** 103–116.

7. Chatterjee, N., Chen, Y.-H. and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association* **98** 158–168. MR1965682 MR1965682
8. Chatterjee, N., Chen, Y.-H., Maas, P. and Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association* **111** 107–117. MR3494641 MR3494641
9. Che, M., Lawless, J. F. and Han, P. (2020). Empirical and Conditional Likelihoods for Two-Phase Studies. *Canadian Journal of Statistics* **doi.org/10.1002/cjs.11566**. MR4267924 MR4267924
10. Derkach, A., Lawless, J. F. and Sun, L. (2015). Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika* **102** 988–994. MR3431569 MR3431569
11. Espin-Garcia, O., Craiu, R. V. and Bull, S. B. (2018). Two-phase designs for joint quantitative-trait-dependent and genotype-dependent sampling in post-GWAS regional sequencing. *Genetic epidemiology* **42** 104–116.
12. Han, P. and Lawless, J. F. (2016). Comment on "Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-level Information from External Big Data Source". *Journal of the American Statistical Association* **111** 118–121. MR3494642 MR3494642
13. Han, P. and Lawless, J. F. (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica* **29** 1321–1342. MR3932520 MR3932520
14. Han, P., Taylor, J. M. G. and Mukherjee, B. (2022). Integrating risk prediction models with no model details. *Canadian Journal of Statistics* **00** To appear. MR4595233
15. Hermansen, K. (2000). Diet, blood pressure and hypertension. *British Journal of Nutrition* **83** S113–S119.
16. Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47** 663–685. MR0053460 MR0053460
17. Huang, C.-Y., Qin, J. and Tsai, H.-T. (2016). Efficient estimation of the Cox model with auxiliary subgroup survival information. *Journal of the American Statistical Association* **111** 787–799. MR3538705 MR3538705
18. Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics* **20** 493–506. MR1973800 MR1973800
19. Keogh, R. H. and Cox, D. R. (2014). *Case-control studies* **4**. Cambridge University Press, Cambridge, UK. MR3443808 MR3443808
20. Kim, S., Cai, J. and Lu, W. (2013). More efficient estimators for case-cohort studies. *Biometrika* **100** 695–708. MR3094446 MR3094446
21. Kulich, M. and Lin, D. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association* **99** 832–844. MR2090916 MR2090916
22. Lawless, J., Kalbfleisch, J. and Wild, C. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61** 413–438. MR1680310 MR1680310
23. Lin, D.-Y., Zeng, D. and Tang, Z.-Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National*

Academy of Sciences of the United States of America **110** 12247–12252. MR3105371 MR3105371

24. Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data* **793**. John Wiley & Sons, Hoboken, New Jersey, USA. MR1925014 MR1925014
25. Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC, Boca Raton, FL, USA.
26. Pepe, M. S. and Fleming, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association* **86** 108–113. MR1137103 MR1137103
27. Piegorsch, W. W., Weinberg, C. R. and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* **13** 153–162.
28. Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika* **87** 484–490. MR1782493 MR1782493
29. Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* 300–325. MR1272085 MR1272085
30. Qin, J., Zhang, B. and Leung, D. H. (2009). Empirical likelihood in missing data problems. *Journal of the American Statistical Association* **104** 1492–1503. MR2750574 MR2750574
31. Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82** 299–314. MR1354230 MR1354230
32. Rivera-Rodriguez, C., Haneuse, S., Wang, M. and Spiegelman, D. (2020). Augmented pseudo-likelihood estimation for two-phase studies. *Statistical Methods in Medical Research* **29** 344–358. MR4064512 MR4064512
33. Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89** 846–866. MR1294730 MR1294730
34. Schaid, D. J., Jenkins, G. D., Ingle, J. N. and Weinshilboum, R. M. (2013). Two-phase designs to follow-up genome-wide association signals with DNA resequencing studies. *Genetic epidemiology* **37** 229–238.
35. Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84** 57–71. MR1450191 MR1450191
36. Scott, A. J. and Wild, C. J. (2011). Fitting regression models with response-biased samples. *Canadian Journal of Statistics* **39** 519–536. MR2842429 MR2842429
37. Tao, R., Zeng, D. and Lin, D.-Y. (2017). Efficient Semiparametric Inference Under Two-Phase Sampling, With Applications to Genetic Association Studies. *Journal of the American Statistical Association* **112** 1468–1476. MR3750869 MR3750869
38. Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media, New York, NY, USA. MR2233926 MR2233926
39. Weaver, M. A. and Zhou, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association* **100** 459–469. MR2160550 MR2160550
40. Zhang, Z. and Rockette, H. E. (2006). Semiparametric maximum likelihood for missing covariates in parametric regression. *Annals of the Institute of Statistical Mathematics* **58** 687–706. MR2345163 MR2345163
41. Zhao, Y., Lawless, J. F. and McLeish, D. L. (2009). Likelihood methods for regression models with expensive variables missing by design. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **51** 123–136. MR2667516 MR2667516
42. Zhou, H., Weaver, M. A., Qin, J., Longnecker, M. and Wang, M. C. (2002).

A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics* **58** 413–421. MR1908182 MR1908182

Note: This list reflects references listed in the original paper as accurately as possible with no attempt to correct errors.