This is a review submitted to Mathematical Reviews/MathSciNet.

Reviewer Name: Peng, Cheng

Mathematical Reviews/MathSciNet Reviewer Number: 82361

Address:
Department of Mathematics
West Chester University
West Chester, PA 19383
cpeng@wcupa.edu

Author: Cao, Yaqi; Chen, Lu; Yang, Ying; Chen, Jinbo

Title: Semiparametric maximum likelihood estimation with two-phase stratified case-control sampling.

MR Number: MR4607797

Primary classification:

Secondary classification(s):

Review text:

sectsty graphicx amsmath,amssymb

The article introduced a semiparametric logistic regression model under a special two-phase stratified case-control sampling plan.

Let Y be the binary response (1= case, 0 = control), and $\mathbf{X}$ be the vector of p covariates with values available for all subjects in phase sampling. Z is a covariate risk variable with values available only for randomly selected subjects in phase II Bernoulli sampling. Let A be the matching strata taking value $a = 1, 2, \cdots, S$. The logistic regression for the a-th stratum is defined to be

$$Pr(Y = 1|\mathbf{X}, Z : A = a) = \frac{\exp(\alpha_a + \beta^T)\mathbf{X} + \beta_2 Z}{1 + \exp(\alpha_a + \beta^T)\mathbf{X} + \beta_2}$$

where α_a is the intercept of a-th stratum for $a = 1, 2, \cdots, S$. A parametric model for $Pr(Z|\mathbf{X}; A)$ with density $p_\theta(Z|\mathbf{X}; A)$ is adopted where θ is the index parameter vector. Let δ_x^a denote the probability mass of $X = x$ in the a-th matching stratum ($a = 1, 2, \cdots, S$), which satisfies $\sum_x \delta_x^a = 1$. Assume further that R is the Bernoulli variable in phase II sampling, then the complete **retrospective** log-likelihood function is given by

$$\ell_0 = \log \prod_{i=1}^{n} P(R_i|Y_i, \mathbf{X}_i, A_i) + \log \left\{ \prod_{i \in P_I/P_{II}} P(\mathbf{X}_i|A_i, Y_i) \times \prod_{i \in P_{II}} Pr(\mathbf{X}_i, Z_i|A_i, Y_i) \right\}$$

where P_I and P_{II} denote subjects in Phase I and Phase II, respectively. Since the missingness of Z_i is at random, the authors proposed a semiparametric ML estimator of $(\beta^T, \theta)^T$ by maximizing the second part of the above log-likelihood function which is expressed in the following

$$
\begin{aligned}
\ell(\beta,\theta,\delta) = \sum_{i=1}^{n} \Big\{ & R_i[Y_i(\beta_i^T \mathbf{X}_i + \beta_2 Z_i)] + \log p_\theta(Z_i|\mathbf{X}_i, A_i) \\
& +(1-R_i)Y_i \log \sum_z [\exp(\beta_1^T \mathbf{X}_i + \beta_2 z) p_\theta(Z=z|\mathbf{X}_i, A_i)] + \log \delta_{\mathbf{X}_i}^{A_i} \Big\} \\
& - \sum_a n_{1a} \log \big\{ \exp(\beta_1^T \mathbf{x} + \beta_2 z) p_\theta(Z=z|\mathbf{x}, a) \cdot \delta_{\mathbf{x}}^{a} \big\}
\end{aligned}
\tag{1}
$$

Since δ is a high-dimensional nuisance parameter, the profile likelihood of (β, θ) is given by

$$
\ell_P(\beta, \theta, \hat{\delta}(\beta, \theta)) = \sup_{\delta} \ell(\beta, \theta, \delta)
$$

The odds ratio parameters will then be estimated from the above log-likelihood function. Some asymptotic results of the profile MLE were presented along with a numerical example using a real-world cancer data.