The Canadian Journal of Statistics
Vol. xx, No. yy, 20??, Pages 1–25
La revue canadienne de statistique

Supplementary Material for "Empirical and conditional likelihoods for two-phase studies"

Menglu Che[1], Jerald F. Lawless[1]* and Peisong Han[2]

[1]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada*
[2]*Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan , USA*

1. ADDITIONAL SIMULATION STUDIES

1.1. Simulation Study 3

This study involves a binary covariate X and continuous covariate Z, which are correlated. We consider a phase 1 sample of 10,000 subjects with data generated as follows. A continuous standard normal covariate Z_i is first generated and then a Bernoulli covariate X_i is generated with probability $P(X_i = 1) = 0.2\mathrm{I}(Z_i > 0) + 0.5\mathrm{I}(Z_i \le 0)$. We then generate the response Y_i using a logistic regression model; with expit(u) denoting $e^u/(1 + e^u)$, it is

$$P(Y = 1|x, z) = \mathrm{expit}(\beta_c + \beta_x x + \beta_z z), \tag{1}$$

with $\boldsymbol{\beta}_0 = (-2.8, 0.5, 1)$; this results in N_1 subjects with $Y = 1$ and N_0 subjects with $Y = 0$. In phase 2, we randomly sample $n_1 = 150$ subjects with $Y_i = 1$, and $n_0 = 150$ subjects with $Y_i = 0$; the X_i are discarded for all other subjects and marked as unobserved.

This is a case of basic stratified sampling (BSS) with the phase 2 sampling depending only on the observed values of Y. The marginal sampling probability for $Y = 1$ cases is $p_1 = 150/N_1$ and for $Y = 0$ cases is $p_0 = 150/N_0$ but the R_i are not independent as for variable probability sampling (VPS). We can nevertheless use the VPS estimating equations and likelihoods, which are asymptotically valid under BSS; we do this, although finite sample adjustments for BSS could be made (e.g. Lawless et al. 1999). Under VPS we would use a logistic regression model for the sampling probabilities:

$$P(R = 1|y) = \pi_{est}(y; \boldsymbol{\alpha}) = \mathrm{expit}(\alpha_c + \alpha_y y), \tag{2}$$

but in the present case the design probabilities p_0, p_1 are random and not fixed, since they depend on N_0 and N_1. We denote estimates obtained using these design probabilities with the suffix est in Table 1. It is possible, however, to increase efficiency of estimation by using a stratified pseudo VPS sampling model that conditions on observed z values, similar to calibration or post-stratification in sampling contexts. We consider two such models, referred to with the suffixes sat1 and sat2 in Table 1. For sat1 we use a binary covariate $v = \mathrm{I}(z > 0.5)$and the model

$$P(R = 1|y, v) = \pi_{sat1}(y, v; \boldsymbol{\alpha}) = \mathrm{expit}(\alpha_c + \alpha_y y + \alpha_v v + \alpha_{yv} yv). \tag{3}$$

Even if the phase 2 VPS sampling probabilities depend only on the value of Y, using model (3) in estimating functions will give more efficient estimators than using model (2). The sat2 model uses the continuous covariate z in a more highly stratified logistic regression model for phase 2 selection, namely

$$P(R = 1|y, z) = \pi_{sat2}(y, z; \boldsymbol{\alpha}) = \mathrm{expit}(\alpha_c + \alpha_y y + \alpha_z z + \alpha_{yz} yz). \tag{4}$$

TABLE 1: Simulation results for Study 3.

Method	Mean (Empirical SE)[Estimated SE]		
	β_c ($\beta_{c0} = -2.8$)	$\beta_z(\beta_{z0} = 0.5)$	$\beta_x(\beta_{x0} = 1)$
CML-est	-2.813 (0.117)[0.123]	0.522 (0.247)[0.257]	1.018 (0.239)[0.250]
CML-sat1	-2.815 (0.115)[0.122]	0.524 (0.198)[0.200]	1.020 (0.239)[0.250]
CML-sat2	-2.814 (0.113)[0.120]	0.524 (0.124)[0.124]	1.021 (0.239)[0.250]
EL-est	-2.813 (0.117)[0.123]	0.522 (0.247)[0.257]	1.018 (0.239)[0.250]
EL-sat1	-2.814 (0.116)[0.122]	0.514 (0.130)[0.134]	1.020 (0.239)[0.249]
EL-sat2	-2.814 (0.114)[0.120]	0.520 (0.122)[0.123]	1.019 (0.240)[0.250]
SW-est	-2.813 (0.117)[0.123]	0.522 (0.247)[0.257]	1.018 (0.239)[0.250]
SW-sat1	-2.814 (0.116)[0.122]	0.515 (0.131)[0.130]	1.020 (0.239)[0.249]
SW-sat2	-2.814 (0.113)[0.120]	0.518 (0.121)[0.123]	1.018 (0.239)[0.250]

Note that working models (3) and (4) both include the true phase 2 sampling model (2) as special cases.

We also considered two pseudo empirical likelihood (PEL) estimators, where the $\boldsymbol{\alpha}$ parameters in models (2), (3), and (4) are first estimated by maximum likelihood from $\boldsymbol{S}_\pi(\boldsymbol{\alpha}) = 0$ and then fixed in the estimating function $\mathbf{U}(\boldsymbol{\phi}) = \mathbf{U}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}_{\rm ML})$. This EL procedure is slightly easier to implement since the estimating function $\boldsymbol{S}_\pi(\widehat{\boldsymbol{\alpha}}_{\rm ML})$ equals zero. Such estimators have been considered by others such as Qin et al. (2009) and Xie and Zhang (2017).

We mention that in this example the estimating equations $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are not linearly independent. Take the π_{sat1} model, for example; then $\dim(\boldsymbol{\beta}) = 3$ and $\dim(\boldsymbol{\alpha}) = 4$ so the dimension of $(\boldsymbol{S}_1^T, \boldsymbol{S}_2^T)^T$ is 7. However in Appendix Section A.3 we show that the actual rank of these 7 estimating equations is 4. Therefore we use here only the first element of $\boldsymbol{S}_2$ for the EL estimator. This phenomenon is an example of the well known fact that $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are not identifiable from the conditional likelihood $l_c(\boldsymbol{\beta}, \boldsymbol{\alpha})$ alone in this setting.

In Table 1, we compare the performance of CML, SW, and EL estimators based on 500 simulations, using each of the three π models (2) - (4). The EL0 and PEL estimators with each π model are asymptotically equivalent to the corresponding EL estimator so are omitted; their finite sample performances are close to those of the EL estimators. We show empirical standard deviations and average standard errors for each estimator; standard errors are obtained by estimating asymptotic covariance matrices with sample covariance matrices evaluated at estimates of $\boldsymbol{\phi}$. These are labelled empirical and estimated standard error (SE) in the table and they are seen to be close in value. In this case, CML performs about as well as the EL and SW methods. A substantial efficiency gain for estimation of β_Z, the coefficient for the covariate that is known for all individuals, occurs when the stratified selection model (3) is used instead of (2) for the EL and SW estimators. A big increase in efficiency for CML and small further increases in efficiency for EL and SW result from using the more highly stratified model (4).

1.2. Simulation Study 4

In Study 4, we simulate a normal linear regression model as in Study 2, but now with X and Z both continuous. We let X, Z follow a bivariate normal distribution with means and standard deviations $\mu = 0, \sigma = 1$, and correlation $\rho = 0.5$. The response model is $Y \sim \mathcal{N}(0.5X + Z, 1)$, and so $\boldsymbol{\beta}_0 = (0, 0.5, 1)$. The phase 1 sample size is $N = 500$ and the phase 2 sampling probability model is $P(R = 1|y, z) = \text{expit}(-1 + 0.5y + 0.5z)$, resulting in about 30% of subjects being selected in phase 2. In this case, we have the conditional likelihood

$$f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{\exp\{-(y - \beta_c - \beta_x x - \beta_z z)^2/(2\sigma^2)\}\text{expit}(\alpha_c + \alpha_y y + \alpha_z z)}{\int \exp\{-(y - \beta_c - \beta_x x - \beta_z z)^2/(2\sigma^2)\}\text{expit}(\alpha_c + \alpha_y y + \alpha_z z) dy}. \quad (5)$$

TABLE 2: Simulation results for Study 4.

Method	Mean (Empirical SE)[Estimated SE]			
	β_c ($\beta_{c0} = -2.8$)	$\beta_z(\beta_{z0} = 0.5)$	$\beta_x(\beta_{x0} = 1)$	$\sigma(\sigma_0 = 1)$
CML0	0.006 (0.102)[0.106]	0.494 (0.091)[0.092]	1.000 (0.093)[0.091]	0.985 (0.060)[0.062]
CML-est	0.008 (0.081)[0.093]	0.493 (0.075)[0.091]	1.000 (0.091)[0.089]	0.985 (0.061)[0.061]
CML-sat	0.005 (0.080)[0.092]	0.498 (0.076)[0.085]	1.000 (0.091)[0.089]	0.985 (0.061)[0.061]
EL-est	0.011 (0.084)[0.087]	0.489 (0.089)[0.090]	0.995 (0.093)[0.088]	0.980 (0.062)[0.060]
EL-sat	0.008 (0.082)[0.085]	0.499 (0.075)[0.081]	0.993 (0.092)[0.088]	0.979 (0.062)[0.060]
SW-est	0.005 (0.074)[0.086]	0.498 (0.076)[0.082]	1.000 (0.091)[0.089]	0.985 (0.061)[0.061]
SW-sat	0.005 (0.074)[0.086]	0.498 (0.076)[0.082]	1.000 (0.091)[0.089]	0.985 (0.061)[0.061]

We consider the two phase 2 selection models

$$\pi_{est}(y, z; \boldsymbol{\alpha}) = P(R = 1|y, z) = \text{expit}(\alpha_c + \alpha_y y + \alpha_z z) \tag{6}$$

$$\pi_{sat}(y, z; \boldsymbol{\alpha}) = P(R = 1|y, z) = \text{expit}(\alpha_c + \alpha_y y + \alpha_z z + \alpha_{yz} yz) \tag{7}$$

for CML, SW, and EL estimation. The performances of the estimators in 100 simulations are compared in Table 2. Once again we find that with the most highly stratified model (7), the three estimators have almost identical empirical standard errors for $\boldsymbol{\beta}_z$, and that EL and SW estimators are slightly more efficient for estimation of β_c.

2. A3. THE RANK OF CL ESTIMATING EQUATIONS FOR SIMULATION STUDY 1 AND 3

With the models in Simulation Studies 1 and 3, both the regression model and π model are in logistic form, so as discussed in Scott and Wild (2011), the conditional probability $p(Y = 1|X, Z, R = 1)$ is also a logistic form, with an offset term $\omega_i = \log\{\pi(y = 1, z_i)/\pi(y = 0, z_i)\}$, and so the conditional log-likelihood is

$$\begin{aligned} l_c(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} r_i [& y_i \log\{\text{expit}(\omega_i + \beta_c + \beta_x x_i + \beta_z z_i)\} \\ & + (1 - y_i) \log\{1 - \text{expit}(\omega_i + \beta_c + \beta_x x_i + \beta_z z_i)\}] \end{aligned}$$

and

$$\frac{\partial l_c}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} r_i \{y_i - \text{expit}(\omega_i + \beta_c + \beta_x x_i + \beta_z z_i)\}(1, x_i, z_i)^T,$$

$$\frac{\partial l_c}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{N} r_i \{y_i - \text{expit}(\omega_i + \beta_c + \beta_x x_i + \beta_z z_i)\} \frac{\partial \omega_i}{\partial \boldsymbol{\alpha}}.$$

When we use the "sat2" selection model, we have

$$
\begin{aligned}
&\frac{\partial \omega_i}{\partial \boldsymbol{\alpha}} = \frac{\partial}{\partial \boldsymbol{\alpha}}[\log\{\text{expit}(\alpha_c + \alpha_y + \alpha_z z_i + \alpha_{yz} z_i)\}] - \frac{\partial}{\partial \boldsymbol{\alpha}}[\log\{\text{expit}(\alpha_c + \alpha_z z_i)\}] \\
&= \{1 - \text{expit}(\alpha_c + \alpha_y + \alpha_z z_i + \alpha_{yz} z_i)\}(1, 1, z_i, z_i)^T \\
&- \{1 - \text{expit}(\alpha_c + \alpha_z z_i)\}(1, 0, z_i, 0)^T \\
&= \begin{pmatrix} \{1 - \text{expit}(\alpha_c + \alpha_y + \alpha_z z_i + \alpha_{yz} z_i)\} - \{1 - \text{expit}(\alpha_c + \alpha_z z_i)\} \\ 1 - \text{expit}(\alpha_c + \alpha_y + \alpha_z z_i + \alpha_{yz} z_i) \\ z_i[\{1 - \text{expit}(\alpha_c + \alpha_y + \alpha_z z_i + \alpha_{yz} z_i)\} - \{1 - \text{expit}(\alpha_c + \alpha_z z_i)\}] \\ z_i\{1 - \text{expit}(\alpha_c + \alpha_y + \alpha_z z_i + \alpha_{yz} z_i)\} \end{pmatrix}. \qquad (8)
\end{aligned}
$$

As Z is a continuous variable, it is easy to see that $\partial \omega_i / \partial \boldsymbol{\alpha}$ as in (8) is a full rank vector in this case (no row of it is a linear combination of other rows).

However, when we use the "sat1" selection model where $\pi(y, z; \alpha) = \pi(y, v(z); \alpha)$, and $v(z)$ is some coarsening of z so that we have two strata defined by the value of z, then at a given value of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we can write

$$
\begin{aligned}
\frac{\partial \omega_i}{\partial \boldsymbol{\alpha}} &= v_i \begin{pmatrix} -\text{expit}(\alpha_c + \alpha_y + \alpha_v + \alpha_{yv}) + \text{expit}(\alpha_c + \alpha_v) \\ 1 - \text{expit}(\alpha_c + \alpha_y + \alpha_v + \alpha_{yv}) \\ -\text{expit}(\alpha_c + \alpha_y + \alpha_v + \alpha_{yv}) + \text{expit}(\alpha_c + \alpha_v) \\ 1 - \text{expit}(\alpha_c + \alpha_y + \alpha_v + \alpha_{yv}) \end{pmatrix} \\
&+ \begin{pmatrix} -\text{expit}(\alpha_c + \alpha_y) + \text{expit}(\alpha_c) \\ 1 - \text{expit}(\alpha_c + \alpha_y) \\ 0 \\ 0 \end{pmatrix} \\
&=: v_i(a_1, a_2, a_1, a_2)^T + (1 - v_i)(b_1, b_2, 0, 0)^T \\
&=: v_i \boldsymbol{a} + (1 - v_i)\boldsymbol{b}
\end{aligned}
$$

where $\boldsymbol{a}, \boldsymbol{b}$ are constant vectors and thus the "Hessian" matrix can be written as

$$
E\left(\frac{\partial \log f_c}{\partial \boldsymbol{\phi}}\right)\left(\frac{\partial \log f_c}{\partial \boldsymbol{\phi}^T}\right) = E\left[r_i\{y_i - \text{expit}(\omega_i + \beta_c + \beta_x x_i + \beta_z z_i)\}^2 \mathbf{u}_i \mathbf{u}_i^T\right]
$$

where

$$
\begin{aligned}
\mathbf{u}_i =& (1, x_i, z_i, a_1 v_i + b_1(1 - v_i), a_2 v_i + b_2(1 - v_i), a_1 v_i, a_2 v_i)^T \\
=& \begin{bmatrix} 1 & 0 & 0 & b_1 & b_2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_1 - b_1 & a_2 - b_2 & a_1 & a_2 \end{bmatrix}^T \begin{bmatrix} 1 \\ x_i \\ z_i \\ v_i \end{bmatrix} := U \times (1, x_i, z_i, v_i)^T
\end{aligned}
$$

and where U is a 7×4 constant matrix. Thus $E\left(\partial \log f_c / \partial \boldsymbol{\phi}\right)\left(\partial \log f_c / \partial \boldsymbol{\phi}\right)^T$ has dimension 7×7 but rank 4.

Received 9 January 2020
Accepted 16 April 2020