UW Biostatistics Working Paper Series

12-30-2010

Modification and Improvement of Empirical Liklihood for Missing Response Problem

Gary Chan
University of Washington, kcgchan@u.washington.edu

Suggested Citation
Chan, Gary, "Modification and Improvement of Empirical Liklihood for Missing Response Problem" (December 2010). *UW Biostatistics Working Paper Series*. Working Paper 371.
http://biostats.bepress.com/uwbiostat/paper371

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

Modification and improvement of empirical likelihood for missing response problem

By KWUN CHUEN GARY CHAN

Department of Biostatistics and Department of Health Services, University of Washington, Seattle, Washington 98195, U.S.A.

kcgchan@u.washington.edu

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Summary

An empirical likelihood (EL) estimator was proposed by Qin and Zhang (2007) for a missing response problem under a missing at random assumption. They showed by simulation studies that the finite sample performance of EL estimator is better than some existing estimators. However, the empirical likelihood estimator does not have a uniformly smaller asymptotic variance than other estimators in general. We consider several modifications to the empirical likelihood estimator and show that the proposed estimator dominates the empirical likelihood estimator and several other existing estimators in terms of asymptotic efficiencies. The proposed estimator also attains the minimum asymptotic variance among estimators having influence functions in a certain class.

Keywords: Auxiliary information, Empirical likelihood, Missing data, Survey sampling

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper371

1 Introduction and existing estimators

Suppose we are interested in estimating the mean μ of a random variable Y but Y is partially observed subject to missingness. Let X be a vector of covariates that are fully observable and R be an indicator that Y is observed. The observed data are $(r_i, r_i y_i, x_i)$ for $i = 1, \ldots, n$ and are *i.i.d.* realizations from (R, RY, X). Under a missing at random assumption that $P(R = 1|Y, X) = P(R = 1|X) = \pi_0(X)$, μ can be consistently estimated by the inverse probability weighting (IPW) estimator

$$\hat{\mu}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{r_i}{\pi_0(x_i)} y_i$$

For missing data applications the nonmissing probability is usually not known but is being modelled. Suppose $P(R = 1|X) = \pi(X; \beta_0)$, where β_0 is a finite dimensional parameter. Based on $(r_1, x_1), \ldots, (r_n, x_n)$, the parameter β_0 can be estimated by solving a likelihood score equation $n^{-1} \sum_{i=1}^{n} s(x_i; \beta) = 0$ where $s(x; \beta) = [1 - \pi(x; \beta)]^{-1}[r_i - \pi(x; \beta)] \frac{\partial \pi}{\partial \beta}(x; \beta)$, and we denote $\hat{\beta}$ to be the solution. We usually replace $\pi_0(x_i)$ by the estimated probability $\pi(x_i; \hat{\beta})$ in IPW estimation.

The IPW estimator is intuitive and easy to implement but is inefficient in general, because information from X is not fully utilized when Y is not observed. To improve efficiency, an empirical likelihood estimator is proposed by Qin and Zhang (2007) where weights p_i are defined for complete case observations (*i.e.* when $r_i = 1$) and the following empirical log-likelihood function

$$l = \sum_{i=1}^{n} r_i \log p_i$$

is maximised subject to constraints

$$p_i \geq 0$$

$$\sum_{i=1}^{n} r_i p_i = 1$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$$\sum_{i=1}^{n} r_i p_i \pi(x_i; \hat{\beta}) = \hat{\theta}$$

$$\sum_{i=1}^{n} r_i p_i a(x_i; \hat{\beta}) = \hat{a}$$

where $a = (a_1, \ldots, a_p)$ is a vector function of p<n dimensions, $\hat{\theta} = n^{-1} \times \sum_{i=1}^{n} \pi(x_i; \hat{\beta})$ and $\hat{a} = n^{-1} \times \sum_{i=1}^{n} a(x_i)$. Let $s(x; \beta, \theta, a) = \{1 - \theta\pi^{-1}(x; \beta), \pi^{-1}(x; \beta)[a(x) - a]^T\}^T$ and $n_1 = \sum_{i=1}^{n} r_i$. Solving the constrained maximization problem, the empirical likelihood weights p_i^{EL} are expressed in terms of a vector of Lagrange multipliers $\hat{\lambda}_{EL}$

$$p_i^{EL} = \frac{1}{n_1} \frac{\hat{\theta}\pi^{-1}(x_i; \beta)}{1 + \hat{\lambda}_{EL}^T s(x_i; \hat{\beta}, \hat{\theta}, \hat{a})} \tag{1}$$

and the Lagrange multipliers satisfies a system of estimating equations

$$\sum_{i=1}^{n} \frac{r_i s(x_i; \hat{\beta}, \hat{\theta}, \hat{a})}{1 + \hat{\lambda}_{EL}^T s(x_i; \hat{\beta}, \hat{\theta}, \hat{a})} = 0. \tag{2}$$

The empirical likelihood estimator for μ is defined as

$$\hat{\mu}_{EL} = \sum_{i=1}^{n} r_i p_i^{EL} y_i.$$

Information from incomplete observations are utilized implicitly in the construction of weights p_i^{EL} from the constraints. When Y and $a(X)$ are correlated, the empirical likelihood estimator usually improves upon the IPW estimator in terms of estimation efficiency.

Although the empirical likelihood estimator has nice small sample properties shown in simulations, it does not uniformly dominate other existing estimators in terms of asymptotic efficiency. We consider two alternative estimators and compare them to the empirical likelihood estimator. A related survey calibration (CAL) estimator is defined by maximizing a pseudo empirical log-likelihood function (Chen, Sitter and Wu, 2002)

$$l_p = \sum_{i=1}^{n} \frac{r_i}{\pi(x_i; \hat{\beta})} \log p_i$$

subject to constraints

$$p_i \geq 0$$

http://biostats.bepress.com/uwbiostat/paper371

$$\sum_{i=1}^{n} r_i p_i = 1$$

$$\sum_{i=1}^{n} r_i p_i a(x_i; \hat{\beta}) = \hat{a}$$

Solving the constrained maximization problem, the calibration weights p_i^{CAL} are expressed in terms of a vector of Lagrange multipliers $\hat{\lambda}_{CAL}$

$$p_i^{CAL} = \frac{\pi^{-1}(x_i; \hat{\beta})[\sum_{i=1}^{n} r_i \pi^{-1}(x_i; \hat{\beta})]^{-1}}{1 + \hat{\lambda}_{CAL}^T [a(x_i) - \hat{a}]}$$

and the Lagrange multipliers satisfies

$$\sum_{i=1}^{n} \frac{r_i \pi^{-1}(x_i; \hat{\beta})[a(x_i) - \hat{a}]}{1 + \hat{\lambda}_{CAL}^T [a(x_i) - a]} = 0.$$

The calibration estimator is defined as

$$\hat{\mu}_{CAL} = \sum_{i=1}^{n} r_i p_i^{CAL} y_i$$

The calibration estimator and the empirical likelihood estimator is very similar but not identical.

The augmented inverse probability weighting (AIPW) estimator is another estimator proposed in the literature to improve efficiency of IPW estimation (Robins, Rotnitzky and Zhao 1994). A regression model is fitted using the complete case subsample, treating Y as outcome and $a(X)$ as covariates. Let $\hat{m}(X)$ be the prediction from the fitted model. An AIPW estimator is defined as

$$\hat{\mu}_{AIPW} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{r_i}{\pi(x_i; \hat{\beta})} y_i + \left[\frac{1 - \pi(x_i; \hat{\beta})}{\pi(x_i; \hat{\beta})} \right] \hat{m}(x_i) \right\}$$

Empirical likelihood, survey calibration and augmented inverse probability weighting do not dominate one another in general in terms of asymptotic efficiencies. It has been shown that the influence function of $\hat{\mu}_{EL}$, $\hat{\mu}_{CAL}$ and $\hat{\mu}_{AIPW}$ belongs to the following class

$$\mathcal{L} = \left\{ \frac{R}{\pi_0(X)} [Y - m(X)] + [m(X) - \mu] : m(X) \text{ is linear in } \tilde{a}(X) \right\}.$$

Hosted by The Berkeley Electronic Press

where $\tilde{a} = (1, a_1, \ldots, a_p, (1-\pi_0)^{-1}\partial\pi^T/\partial\beta)^T$. Our main results in section 2 are to show that certain modifications of the empirical likelihood estimator has an influence function corresponding to the minimal asymptotic variance among the class $\mathcal{L}$. That is, the modified empirical likelihood estimator is at least as efficient as the EL, CAL and AIPW estimators when the same amount of covariate information is used. Section 3 will present simulation studies comparing the finite performance of estimators.

2 Modified empirical likelihood

In this section we propose several modifications to the empirical likelihood estimator and show that it attains the minimum asymptotic variance among estimators having influence functions in the class $\mathcal{L}$.

Since $\hat{\theta}$ is a consistent estimator for $P(R=1)$, in the modification we replace $\hat{\theta}/n_1$ in (1) by $1/n$. Also, we replace $s(x_i; \hat{\beta}, \hat{\theta}, \hat{a})$ by $s^*(x_i; \hat{\beta}, \hat{a}, \hat{b})$, where

$$s^*(x; \beta, a, b) = \frac{1-\pi(x;\beta)}{\pi(x;\beta)} \times \left(1, [a(x)-a]^T, \left[\frac{1}{1-\pi(x;\beta)}\frac{\partial\pi}{\partial\beta}(x;\beta) - b\right]^T\right)^T$$

and $\hat{b} = \sum_{i=1}^n (1-\pi(x_i); \hat{\beta})^{-1}\partial\pi(x_i; \hat{\beta})/\partial\beta$. The modified empirical likelihood (MEL) weights are defined as

$$p_i^{MEL} = \frac{1}{n}\frac{\pi^{-1}(x_i; \hat{\beta})}{1 + \hat{\lambda}_{MEL}^T s^*(x_i; \hat{\beta}, \hat{a}, \hat{b})} \tag{3}$$

where the pseudo Lagrange multiplier $\hat{\lambda}_{MEL}$ are obtained by solving

$$\sum_{i=1}^n \frac{r_i(1-\pi(x_i; \hat{\beta}))^{-1} s^*(x; \hat{\beta}, \hat{a}, \hat{b})}{1 + \hat{\lambda}_{MEL}^T s^*(x_i; \hat{\beta}, \hat{a}, \hat{b})} = e \tag{4}$$

where $e = (1, 0, \ldots, 0)^T$. Plugging (3) into (4) gives $\sum_{i=1}^n r_i p_i^{MEL} = 1$ and $\sum_{i=1}^n r_i p_i^{MEL} a(x_i) = \hat{a}$, which corresponds to the constraints in empirical likelihood estimation. In addition, we have $\sum_{i=1}^n r_i p_i^{MEL}(1-\pi(x_i); \hat{\beta})^{-1}\partial\pi(x_i; \hat{\beta})/\partial\beta = \hat{b}$. Unlike (1) and (2), (3) and (4) are not implied by constrained maximization problems. The reason is similar to the fact that not every estimating functions are derivative of log-likelihood functions. The modified empirical

http://biostats.bepress.com/uwbiostat/paper371

likelihood estimator is defined as

$$\hat{\mu}_{MEL} = \sum_{i=1}^{n} r_i p_i^{MEL} y_i.$$

The modified empirical likelihood estimator can be shown to achieve the minimum asymptotic variance among class $\mathcal{L}$. For $m(X)$ being any linear functions of $\tilde{a}(X)$, the variance of $\frac{R}{\pi_0(X)}[Y - m(X)] + [m(X) - \mu]$ is $Var(Y) + E(\frac{1-\pi_0(X)}{\pi_0(X)}(Y - m(X))^2)$. Let $m_0(X) = c_0^T \tilde{a}(X)$ where

$$c_0 = \arg\min_{c \in \mathbb{R}^q} E\left(\frac{1-\pi_0(X)}{\pi_0(X)}(Y - c^T \tilde{a}(X))^2\right)$$

where q is the dimension of $\tilde{a}$. By the definition of c_0, the following set of normal equations are satisfied.

$$E\left(\frac{1-\pi_0(X)}{\pi_0(X)}\tilde{a}^T(X)(Y - c_0^T \tilde{a}(X))\right) = 0 \tag{5}$$

Also, the variance of $\frac{R}{\pi_0(X)}[Y - m_0(X)] + [m_0(X) - \mu]$ is the minimum among the class $\mathcal{L}$. We note that

$$\begin{aligned}
\hat{\mu}_{MEL} - \mu =& \sum_{i=1}^{n} r_i p_i^{MEL} y_i - \mu \\
=& \sum_{i=1}^{n} r_i p_i^{MEL}(y_i - m_0(x_i)) + \sum_{i=1}^{n} r_i p_i^{MEL} m_0(x_i) - \mu \\
=& \sum_{i=1}^{n} r_i p_i^{MEL}(y_i - m_0(x_i)) + \frac{1}{n}\sum_{i=1}^{n}(m_0(x_i) - \mu) \\
=& \frac{1}{n}\sum_{i=1}^{n} r_i \left[\frac{1}{\pi(x_i;\hat{\beta})(1 + \hat{\lambda}_{MEL}^T s^*(x_i;\hat{\beta},\hat{a},\hat{b}))} - \frac{1}{\pi_0(x_i)}\right](y_i - m_0(x_i)) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi_0(x_i)}[y_i - m_0(x_i)] + [m_0(x_i) - \mu]\right\}
\end{aligned} \tag{6}$$

where the second last equality follows from (3), (4) and the definition of s^*. By asymptotic properties for estimating equations (Newey and McFadden, 1994), $\sqrt{n}(\hat{\lambda}^{MEL} - 0)$ and $\sqrt{n}(\hat{\beta} -$

Hosted by The Berkeley Electronic Press

β) converges weakly to Gaussian distributions. By Taylor Series expansions,

$$\frac{1}{n}\sum_{i=1}^{n} r_i \left[\frac{1}{\pi(x_i;\hat{\beta})(1+\hat{\lambda}_{MEL}^T s^*(x_i;\hat{\beta},\hat{a},\hat{b}))} - \frac{1}{\pi_0(x_i)}\right](y_i - m_0(x_i))$$

$$= A(\hat{\lambda}^{MEL} - 0) + B(\hat{\beta} - \beta) + c\lambda^T|_{\lambda=0}(\hat{a}^T - \mu_a^T, \hat{b}^T - \mu_b^T)^T + o_p(n^{-1/2}) \tag{7}$$

where $\mu_a = E(a(X))$, $\mu_b = E((1-\pi_0(X))^{-1}\frac{\partial\pi(X;\beta)}{\partial\beta})$, $A = E(\frac{1-\pi_0(X)}{\pi_0(X)}(\tilde{a}(X) - \mu_a)^T(Y - m_0(X)))$, $B = E(\frac{1}{\pi_0(X)}\frac{\partial\pi^T}{\partial\beta}(Y - m_0(X)))$ and $c = E(\frac{1-\pi_0(X)}{\pi_0(X)})$. Matrices A and B are both 0 following the normal equations (5). Note that the form of A is dependent on the method of estimating the Lagrange multipliers. For empirical likelihood and calibration, A will be different matrices and is generally non-zero. For modified empirical likelihood estimator, it follows from (6) and (7) that the influence function of $\hat{\mu}_{MEL}$ is $\frac{R}{\pi_0(X)}[Y - m_0(X)] + [m_0(X) - \mu]$ which attains the minimum variance among the class $\mathcal{L}$.

In the special case where $E(Y|X) = b_0 + b_1^T a(X)$ for some b_0 and b_1, $E[(Y - E(Y|X))^2|X]$ is minimized at each X and therefore $m_0(X) = E(Y|X)$. In this case, the modified empirical likelihood estimator attains the semiparametric efficiency bound. Also, empirical likelihood, calibration and AIPW estimators attain the same asymptotic variance as the modified empirical likelihood estimator under correct specification of the outcome regression model. However, when the outcome regression regression model is misspecified, the modified empirical likelihood has a smaller asymptotic variance than other estimators in general.

The modified empirical likelihood estimator also possesses a double robustness property as for the empirical likelihood estimator. Suppose $E(Y|X) = b_0 + b_1^T a(X) = m_0(X)$ but the missing data model $\pi(x;\beta)$ is misspecified. The estimates $\hat{\beta}$, $\hat{\lambda}_{MEL}$ and $\hat{b}$ converges in probability to some constants β^*, λ^* and μ_b^* and λ^* is usually non-zero. From (6) we note that

$$\begin{aligned}\hat{\mu}_{MEL} &= \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi(x_i;\hat{\beta})(1+\hat{\lambda}_{MEL}^T s^*(x_i;\hat{\beta},\hat{a},\hat{\beta}))}(y_i - m_0(x_i)) + \frac{1}{N}\sum_{i=1}^{N} m_0(x_i) \\ &\xrightarrow{p} E\left(\frac{\pi_0(X)}{\pi(X;\beta^*)(1+\lambda^{*T}s^*(X;\beta^*,\mu_a,\mu_b^*))}(E(Y|X) - m_0(X))\right) + E(E(Y|X))\end{aligned}$$

http://biostats.bepress.com/uwbiostat/paper371

$$= \quad 0 + \mu = \mu$$

That is, the modified empirical likelihood estimator is consistent when the outcome regression model is correctly specified even when the missing data model is misspecified.

3 Simulations

In this section we present simulation studies to evaluate the finite sample performance of the modified empirical likelihood estimator. The simulation studies followed the scenario in Kang and Schafer (2007) for estimating a population mean. The scenario was carefully designed so that the assumed outcome regression and missing data models are nearly correct under misspecification, but the AIPW estimator can be severely biased. Sample sizes for each simulated data set was 200, 500 or 1000, and 1000 Monte Carlo datasets were generated. For each observation, a random vector $Z = (Z_1, Z_2, Z_3, Z_4)$ was generated from a standard multivariate normal distribution, and transformations $X_1 = \exp(Z_1/2), X_2 = Z_2/(1 + \exp(Z_1)), X_3 = (Z_1 Z_3/25 + 0.6)^3$ and $X_4 = (Z_2 + Z_4 + 20)^2$ were defined. The outcome of interest Y was generated from a normal distribution with mean $210 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4$ and unit variance, and Y was observed with probability $\exp(\eta_0(Z))/(1 + \exp(\eta_0(Z)))$ where $\eta_0(Z) = -Z_1 + 0.5Z_2 - 0.25Z_3 - 0.1Z_4$. The correctly specified outcome and missing data models were regression models with Z as covariates, whereas we treated X to be the covariates in misspecified models instead of Z. Kang and Schafer (2007) showed that the missspecified models are nearly correctly specified. In each case we considered four possible combinations of correct and misspecified missing data and outcome regression models: (a) both correct; (b) correct missing data model and incorrect outcome regression; (c) incorrect missing data model but correct outcome regression and (d) both incorrect. For correctly specified outcome model, $a(Z) = (Z_1, Z_2, Z_3, Z_4)$ and to $a(X) = (X_1, X_2, X_3, X_4)$ for misspecified outcome model. We compared the performances of the augmented inverse probability weighted estimator $\hat{\mu}_{AIPW}$, the empirical likelihood estimator $\hat{\mu}_{EL}$, the survey calibration estimator

Hosted by The Berkeley Electronic Press

$\hat{\mu}_{CAL}$ and the modified empirical likelihood estimator $\hat{\mu}_{MEL}$. The results are shown in Table 1.

Simulation results showed that EL, CAL, AIPW and MEL estimators all had relatively small bias when either the missing data model or the outcome regression model was correctly specified. When both models were correctly specified, all estimators had very similar performances because all of them were semiparametric locally efficient. When only one of the two models were correctly specified, the empirical likelihood, calibration and modified empirical likelihood estimators were more efficient than the AIPW estimator. When both models were misspecified, the AIPW estimator had a considerable bias and variability but the other empirical likelihood based estimators showed much better performance. When the outcome regression model was misspecified, the modified empirical likelihood estimators had smaller bias and variability compared to the empirical likelihood and calibration estimators, consistent with the theoretical results. In this simulation study, the modified empirical likelihood estimator performs consistently better than other estimators.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper371

References

Chen, J., Sitter, R. & Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* 89 230–237.

Kang, J. & Schafer, J. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* 22 523–539.

Newey, W. & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* 4 2111–2245.

Qin, J. & Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 101–122.

Robins, J., Rotnitzky, A. & Zhao, L. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association* 89 846–866.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 1: Comparisons among estimators under the Kang and Schafer scenario with four possible combinations of correct and misspecified missing data and outcome regression models, (a) both correct, (b) correct missing data model and incorrect outcome regression, (c) incorrect missing data model but correct outcome regression and (d) both incorrect. RMSE represents the square root of sampling mean squared error.

		(a)		(b)		(c)		(d)	
n		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
200	$\hat{\mu}_{AIPW}$	0.02	2.50	0.28	3.77	0.01	2.55	-8.00	41.07
	$\hat{\mu}_{EL}$	0.02	2.50	0.49	2.90	0.02	2.50	-1.71	3.52
	$\hat{\mu}_{CAL}$	0.02	2.50	0.28	3.14	0.01	2.49	-2.73	4.83
	$\hat{\mu}_{MEL}$	0.02	2.50	0.21	2.62	0.03	2.50	-1.07	3.51
500	$\hat{\mu}_{AIPW}$	0.03	1.62	0.12	2.74	0.11	2.36	-39.66	898.58
	$\hat{\mu}_{EL}$	0.03	1.62	0.30	1.78	0.03	1.62	-2.06	2.81
	$\hat{\mu}_{CAL}$	0.03	1.62	0.16	1.94	0.03	2.61	-3.53	4.62
	$\hat{\mu}_{MEL}$	0.03	1.61	0.16	1.65	0.03	1.62	-1.06	2.14
1000	$\hat{\mu}_{AIPW}$	0.01	1.13	0.06	1.65	-0.01	1.25	-13.38	73.39
	$\hat{\mu}_{EL}$	0.01	1.13	0.19	1.22	0.01	1.13	-2.15	2.52
	$\hat{\mu}_{CAL}$	0.01	1.13	0.10	1.35	0.01	1.13	-4.16	5.05
	$\hat{\mu}_{MEL}$	0.01	1.13	0.11	1.16	0.01	1.13	-1.18	1.72

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper371