

Empirical likelihood inference in the presence of measurement error

Jiahua Chen

Department of Statistics & Actuarial Science

University of Waterloo

Presented at the Workshop on Developments
and Challenges in Mixture Models, Bump Hunt-
ing and Measurement Error Models.

June 2-4, 2002

Dively Center, Case Western Reserve Univer-
sity Cleveland, Ohio

This presentation is based on the joint work
with B. Zhong and J.N.K. Rao

Research partially supported by NSERC.

Outline of the talk

- Description of the Problem;
- Literature review;
- Empirical likelihood approach;
- Some results and conclusions;
- Simulation study.

Description of the Problem

We consider a general problem in sampling survey.

Assume that there is a finite population consists of N sampling units.

Each unit has some characteristic of interest.

The problem of survey is to make inference on the finite population parameters such as:

$$\text{Population mean : } \bar{Y} = N^{-1} \sum_{i=1}^N y_i;$$

$$\text{Population CDF : } F_N(y) = N^{-1} \sum_{i=1}^N I(y_i \leq y)$$

with $I(y_i \leq y)$ being an indicator function.

Typically, the inference is done through a sample survey.

First, we obtain a random sample from the finite population according to a sampling plan.

Second, we obtain measurements on sampled units.

Finally, we analysis the data and the inferences about the finite population parameters are then made.

We must choose or design a most appropriate statistical inference method given a sampling plan and some measurements.

We consider the following situation.

Several instruments might be used to take the measurements. They have different levels of precision.

Some (perfect) instruments are costly but have very high precision: measurements (almost) without error.

Some (imperfect) instruments cost little money, but the measurements obtained are less accurate.

When several instruments are used in a single survey, how should the statistical inferences be done?

In general, we are willing to assume the imperfect measurements are unbiased.

In this case, commonly used linear estimators for the population mean or total remain unbiased.

However, this does not generalize to the estimation of CDF.

Main problem of this talk

How to combine the information from both perfect and imperfect measurements to best estimate the population CDF?

Let us first introduce some notation and some existing methods.

Literature Review

Let the finite population under consideration be labeled as $\{1, \dots, N\}$;

The true value of the i th unit is y_{0i} .

Assume that we have two a simple random samples without replacement: s_0 and s_1 .

Perfect measurements are $\{y_{0i}, i \in s_0\}$.

Imperfect measurements are $\{y_{1i}, i \in s_1\}$.

Two natural estimators are:

$$\begin{aligned}\tilde{F}_1(t) &= n_1^{-1} \sum_{i \in s_1} I(y_{1i} \leq t); \\ \hat{F}_0(t) &= n_0^{-1} \sum_{i \in s_0} I(y_{0i} \leq t).\end{aligned}$$

If $y_{1i} = y_{0i} + e_i$ such that $E(e_i) = 0$, \tilde{F}_1 is not an unbiased estimator of $F_0(t)$.

The bias does not decrease when the sample size increases.

\hat{F}_0 is unbiased, but it does not utilize the entire sample information.

Two general estimators that utilize all the sample data are the regression-type estimator:

$$\hat{F}_{0,reg}(t) = \hat{F}_0(t) + B\{\tilde{F}_1(t) - \hat{F}_1(t)\} \quad (1)$$

and the ratio estimator

$$\hat{F}_{0,R}(t) = \hat{F}_0(t)\tilde{F}_1(t)/\hat{F}_1(t), \quad (2)$$

where

$$\hat{F}_1(t) = n_0^{-1} \sum_{i \in s_0} I(y_{1i} \leq t).$$

Remark: We need to assume $s_0 \subset s_1$ in this case.

In the regression-type estimator (1), we may choose an estimator of optimal B that minimizes the variance.

The estimators (1) and (2) are asymptotically unbiased but,

(i) the estimated CDF is not a monotone function in either case;

(ii) the best B depends on the value of t .

Consequently, there is a need of different method.

Empirical Likelihood Approach

When an iid sample $\{y_i, i = 1, \dots, n\}$ with common CDF, F , is available, the log-empirical likelihood function can be written as

$$\ell_n(F) = \sum_{i=1}^n \log p_i$$

such that $p_i = P(Y = y_i)$.

Inferences about F can be made by regarding $\ell_n(F)$ as an ordinary log-likelihood function.

For example, the MLE of F is given by the usual empirical distribution.

The empirical likelihood can further make use of auxiliary information to improve the efficiency of the inferences.

Suppose that $E\{g(Y, \theta)\} = 0$ is known for some $g(y, \theta)$.

Then, we maximize $\ell_n(F)$ subject to

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i g(y_i, \theta) = 0 \quad (0 \leq p_i \leq 1).$$

Given θ , we get

$$\hat{p}_i(\theta) = \frac{1}{n\{1 + \lambda'g(y_i, \theta)\}}, \quad i = 1, \dots, n,$$

where $\lambda = \lambda(\theta)$ satisfying $\sum \hat{p}_i(\theta)g(y_i, \theta) = 0$.

The empirical likelihood is maximized further with respect to θ .

When g 's dimension is larger than that of θ , the resulting

$$\hat{F}_m(t) = \sum_{i=1}^n \hat{p}_i(\hat{\theta})I(y_i \leq t),$$

is asymptotically more efficient.

Application to the measurement error problem

We assume that there are $H + 1$ instruments with $H + 1$ independent random samples s_0, \dots, s_H of sizes n_0, \dots, n_H .

We name instrument 0 as perfect, and the measurements as $\{y_{hi}, i \in s_h, h = 0, \dots, H\}$.

Remark: When the population size N is large and n is relatively small, the dependence between the sampled units are very weak.

In most sampling problems, unduly assuming independence results in more conservative inferences.

Further, we assume some auxiliary information exist and can be summarized as

$$E_h\{g_h(Y, \theta)\} = 0$$

for $h = 0, \dots, H$, where E_h is taking under the distribution F_h associated with instrument h .

All F_h 's share the parameter θ and g_h is a vector-valued estimating function with dimension d_h .

The log-empirical likelihood is given by

$$\ell_n(F_0, \dots, F_H) = \sum_{h=0}^H \sum_{i \in s_h} \log p_{hi} \quad (3)$$

where $p_{hi} = P(Y = y_{hi} | \text{instrument } h)$.

The inference is done by first maximizing $\ell_n(F_0, \dots, F_H)$ under constraints given by

$$\sum_{i \in s_h} p_{hi} = 1, \quad \sum_{i \in s_h} p_{hi} g_h(y_{hi}, \theta) = 0$$

for $h = 0, \dots, H$.

If θ is known, the resulting MEL estimator of F_0 is given by

$$\hat{F}_{0m}(t) = \sum_{i \in s_0} \hat{p}_{0i}(\theta) I(y_{0i} \leq t).$$

We have

$$\hat{p}_{hi}(\theta) = \frac{1}{n_h \{1 + \lambda'_h g_h(y_{hi}, \theta)\}}$$

for $i \in s_h$ and λ_h is a solution of

$$\sum_{i \in s_h} \hat{p}_{hi}(\theta) g_h(y_{hi}, \theta) = 0$$

for $h = 0, \dots, H$.

We then define the profile log-empirical likelihood ratio

$$r_n(\theta) = \sum_{h=0}^H \sum_{i \in s_h} \log\{1 + \lambda'_h g_h(y_{hi}, \theta)\}.$$

The second step is to maximize r_n with respect to θ to obtain its MEL estimator $\hat{\theta}_m$ and the corresponding

$$\hat{F}_{0m}(t) = \sum_{i \in s_0} \hat{p}_{0i}(\hat{\theta}_m) I(y_{0i} \leq t).$$

Does the empirical likelihood help?

The following result can be obtained by directly applying the techniques in Qin & Lawless (1994) or in Chen & Qin (1993).

Theorem. Suppose that as $n_0 \rightarrow \infty$, $n_h/n_0 \rightarrow k_h > 0$ for all h and that the y_{hi} 's are independent observations.

Assume that the g_h 's are all twice differentiable, that the derivatives are bounded by some integrable function, and that $\text{var}\{g_h(Y, \theta)\}$ is positive definite for each h in a neighborhood of the true parameter value θ_0 .

Then, there is a sequence of MEL estimators $\hat{\theta}_m$, and a positive definite matrix V_0 such that

$$\sqrt{n_0}(\hat{\theta}_m - \theta_0) \rightarrow_d N(0, V_0),$$

where \rightarrow_d denotes the convergence in distribution.

Corollary. Let the first component of θ be $F_0(t)$, t fixed. Then $\sqrt{n_0}\{\hat{F}_{0m}(t) - F_0(t)\}$ is asymptotically normal with variance W_0 , given by the (1,1)th element of the corresponding V_0 .

The asymptotic variance of $\hat{F}_{0m}(t)$ is always smaller than the variance of the empirical CDF based on the sample data obtained from the perfect instrument, 0, only.

Example 1: Common Mean Model

We assume $E(Y_1) = E(Y_0) = \theta_1$.

We then have two unbiased estimating functions

$$g_0(y, \theta_1) = g_1(y, \theta_1) = y - \theta_1.$$

We have two equations and one unknown parameter, hence the empirical likelihood approach is expected to be better.

The asymptotic variance of the MEL estimator

Let $\sigma_0^2 = \text{var}(Y_0)$ and $\sigma_1^2 = \text{var}(Y_1)$.

$$\frac{\partial g_0}{\partial \theta} = \frac{\partial g_1}{\partial \theta} = -1$$

and

$$E_0(g_0^2) = \sigma_0^2, \quad E_1(g_1^2) = \sigma_1^2.$$

In this case,

$$V = \left(\frac{1}{\sigma_0^2} + \frac{k_1}{\sigma_1^2} \right)^{-1}.$$

Is it reasonable?

We find that V/n_0 is the variance of the “optimal” linear estimator

$$\hat{\theta}_l = \frac{n_0\sigma_1^2}{n_1\sigma_0^2 + n_0\sigma_1^2} \bar{y}_0 + \frac{n_1\sigma_0^2}{n_1\sigma_0^2 + n_0\sigma_1^2} \bar{y}_1.$$

However, the optimal linear estimator cannot be used unless the variance ratio σ_1^2/σ_0^2 is known.

Asymptotic Variance of \hat{F}_{0m}

Let $\theta = [F_0(t), \theta_1]'$,

$g_0 = [I(y \leq t) - F_0(t), y - \theta_1]'$ and $g_1 = y - \theta_1$.

Using the Corollary,

$$V = \left[\begin{bmatrix} F_0(t)\{1 - F_0(t)\} & B \\ B & \sigma_0^2 \end{bmatrix}^{-1} + k_1 \begin{bmatrix} 0 & 0 \\ 0 & \sigma_1^{-2} \end{bmatrix} \right]^{-1}.$$

Its (1, 1)th element is the asymptotic variance of $\hat{F}_{0m}(t)$

$$W = F_0(t)\{1 - F_0(t)\} - \frac{n_1 B^2}{n_1 \sigma_0^2 + n_0 \sigma_1^2}.$$

Example 2: Additive Model

Assume that $E(Y_0) = E(Y_1) = \theta_1$ and $\text{var}(Y_1) - \text{var}(Y_0) = \sigma^2$ with σ^2 known.

The estimating functions are

$$g_0(y, \theta) = (y - \theta_1, y^2 - \theta_1^2 - \theta_2)'$$

and

$$g_1(y, \theta) = (y - \theta_1, y^2 - \theta_1^2 - \theta_2 - \sigma^2)',$$

where $\theta = (\theta_1, \theta_2)'$ and $\theta_2 = \sigma_0^2 = \text{var}(Y_0)$.

When $\mu_{h3} = 0$, the asymptotic variance of $\hat{\theta}_{1m}$ is $n_0^{-1} \{ \theta_2^{-1} + k_1(\theta_2 + \sigma^2)^{-1} \}^{-1}$, which equal the variance of the optimal linear estimator $\hat{\theta}_l$.

It can be shown that $\hat{F}_{0m}(t)$ has smaller asymptotic variance than $\hat{F}_0(t)$.

Example 3: Product Model

Assume that $E(Y_0) = E(Y_1) = \theta_1$ and $\text{var}(Y_1) = c\text{var}(Y_0) = c\theta_2$ with $c > 1$ known.

$$g_0(y, \theta) = (y - \theta_1, y^2 - \theta_1^2 - \theta_2)';$$

$$g_1(y, \theta) = (y - \theta_1, y^2 - \theta_1^2 - c\theta_2)'.$$

When the skewness $\mu_{h3} = 0$ for $h = 0, 1$. the asymptotic variance of $\hat{\theta}_{1m}$ equals $n_0^{-1}v_{11}$, where

$$\begin{aligned} n_0^{-1}v_{11} &= n_0^{-1} \left\{ \frac{1 + k_1/c}{\theta_2} + \frac{4k_1\theta_1^2(c-1)^2}{\mu_{14} + k_1c^2\mu_{04}} \right\}^{-1} \\ &\leq n_0^{-1} \left(\frac{\theta_2}{1 + k_1/c} \right), \end{aligned}$$

with equality if and only if $c = 1$.

That is, the MEL estimator $\hat{\theta}_{1m}$ again beats the optimal linear estimator.

The asymptotic variance of $\hat{F}_{0m}(t)$ is complex, but can be shown that it never exceeds the variance of $\hat{F}_0(t)$.

Simulation

We considered the case of two instruments and chose $(n_0, n_1) = (50, 600)$ and $(100, 400)$.

We studied two kinds of population distributions: (i) normal, (ii) chi-square.

We used the situations in three examples discussed.

Table 1: MSE of mean estimators ($\times 1000$).

Normal				
c	1	2	3	Opt.L
$n_0 = 100, n_1 = 400$				
1.50	2.73	2.75	2.74	2.73
2.00	3.36	3.38	3.37	3.36
$n_0 = 50, n_1 = 600$				
1.50	2.26	2.37	2.27	2.26
2.00	2.96	2.80	2.99	2.96
Chi-square				
c	1	2	3	Opt.L
$n_0 = 100, n_1 = 400$				
1.5	4.42	4.42	4.49	4.41
2.0	5.32	5.23	5.25	5.47
$n_0 = 50, n_1 = 600$				
1.5	3.65	3.73	3.76	4.66
2.0	4.49	4.61	4.77	4.60

*Estimators 1, 2, 3 refer to Examples 1, 2, 3 and Opt.L = Optimal linear estimator.

Table 2: MSE of CDF estimators ($\times 1000$):
50% extra variance($c = 1.5$).

	quantile				
	0.10	0.25	0.50	0.75	0.90
	normal: $n_0 = 100, n_1 = 400$				
perfect	0.89	1.90	2.53	1.89	0.89
imperfect	2.61	2.20	0.62	2.17	2.59
1	0.68	1.17	1.37	1.13	0.68
2	0.52	1.01	1.38	0.98	0.52
2*	0.54	1.04	1.36	1.02	0.55
3	0.48	0.97	1.38	0.94	0.48
3*	0.54	1.04	1.36	1.02	0.54
	chi-square: $n_0 = 100, n_1 = 400$				
perfect	0.89	1.86	2.51	1.91	0.92
imperfect	5.82	1.94	1.05	2.70	1.47
1	0.78	1.32	1.42	1.01	0.60
2	0.85	1.21	1.23	1.02	0.61
2*	0.82	1.24	1.23	1.02	0.61
3	0.87	1.25	1.26	1.01	0.59
3*	0.80	1.36	1.46	1.11	0.64

Table 2: Continued

	quantile				
	0.10	0.25	0.50	0.75	0.90
	normal: $n_0 = 50, n_1 = 600$				
perfect	1.85	3.74	5.01	3.70	1.78
imperfect	2.49	2.00	0.41	2.02	2.49
1	1.33	1.98	2.22	1.93	1.31
2	0.94	1.61	2.23	1.58	0.92
2*	1.04	1.70	2.16	1.75	1.03
3	0.92	1.60	2.22	1.57	0.90
3*	0.98	1.65	2.14	1.71	0.98
	chi-square: $n_0 = 50, n_1 = 600$				
perfect	1.79	3.82	4.94	3.80	1.81
imperfect	5.65	1.75	0.84	2.51	1.37
1	1.49	2.50	2.39	1.65	1.10
2	1.79	2.17	1.84	1.92	1.31
2*	1.74	2.12	1.86	1.93	1.30
3	1.82	2.23	1.88	1.88	1.27
3*	1.53	2.53	2.49	1.77	1.16

Estimators 2 and 3* refer to Examples 2 and 3 when σ^2 and c are estimated. “perfect” and “imperfect” refer to estimators based on perfect measurements only and on imperfect measurements only.

Table 3: MSE of CDF estimators ($\times 100$): 100% extra variance($c = 2.0$).

		quantile				
		0.10	0.25	0.50	0.75	0.90
		normal: $n_0 = 100, n_1 = 400$				
perfect		0.88	1.86	2.52	1.87	0.91
imperfect		7.14	4.96	0.63	4.96	7.11
	1	0.69	1.20	1.46	1.19	0.71
	2	0.57	1.09	1.47	1.08	0.59
	2*	0.66	1.15	1.44	1.12	0.63
	3	0.50	1.03	1.46	1.01	0.51
	3*	0.64	1.14	1.43	1.11	0.61
		chi-square: $n_0 = 100, n_1 = 400$				
perfect		0.91	1.94	2.48	1.88	0.89
imperfect		13.34	4.49	1.26	6.47	4.62
	1	0.79	1.43	1.48	1.09	0.61
	2	0.88	1.33	1.30	1.07	0.61
	2*	0.85	1.27	1.36	1.08	0.62
	3	0.92	1.42	1.36	1.04	0.57
	3*	0.80	1.36	1.64	1.34	0.76

Table 3: Continued

	quantile				
	0.10	0.25	0.50	0.75	0.90
	normal: $n_0 = 100, n_1 = 400$				
perfect	1.76	3.76	4.96	3.72	1.80
imperfect	7.05	4.84	0.42	4.82	7.03
1	1.28	2.06	2.29	2.00	1.33
2	0.96	1.69	2.30	1.68	0.98
2*	1.30	2.04	2.30	1.98	1.30
3	0.91	1.63	2.29	1.64	0.92
3*	1.14	1.88	2.27	1.82	1.13
	chi-square: $n_0 = 50, n_1 = 600$				
perfect	1.79	3.72	5.01	3.70	1.74
imperfect	13.23	4.31	1.03	6.30	4.52
1	1.49	2.50	2.47	1.74	1.10
2	1.79	2.19	1.97	1.95	1.30
2*	1.77	2.20	1.98	1.97	1.33
3	1.89	2.34	2.05	1.86	1.21
3*	1.51	2.42	2.58	2.15	1.32

Estimators 2 and 3* refer to Examples 2 and 3 when σ^2 and c are estimated. “perfect” and “imperfect” refer to estimators based on perfect measurements only and on imperfect measurements only.

What is good?

The empirical likelihood approach helps for the low and high quantiles.

What is not satisfactory?

It fails to beat the estimator based on imperfect measurements only for the median.

What can be done in the future?

This is totally open.