Semiparametric maximum likelihood inference by using failed contact attempts to adjust for nonignorable nonresponse

Jing Qin & Dean Follmann & Other collaborators

National Institute of Allergy and Infectious Diseases
National Institutes of Health

November 6, 2015

Contents

A statement

I am a simple minded person. I don't know how to derive complicated mathematical formula. If you have such a problem, please do not ask me. Professor Jae Kim is the right person to ask.

Missing not at random data and biased sampling problem

Denote X as salary in a survey. $D_i = 1$ is the i-th individual responses, and 0 otherwise.

$$P(D_i = 1|x_i) = \pi(x_i)$$

where $\pi(x)$ is a monotone decreasing function. Denote the observed data as

$$(X_1, D_1 = 1), ..., (X_{n_1}, D_{n_1} = 1), (?, D_{n_1+1} = 0), ..., (?, D_n = 0)$$

Likelihood

Assume $X_i, i = 1, 2, ..., n \sim f(x)$. The likelihood is

$$L = \prod_{i=1}^{n_1} \{\pi(x_i) f(x_i)\} \prod_{i=n_1+1}^{n} \int \{1 - \pi(x)\} f(x) dx$$

It can be decomposed as $L = L_1 L_2$, where

$$\begin{aligned} L_1 &= \prod_{i=1}^{n_1} \frac{\pi(x_i) f(x_i)}{\int \pi(x) f(x) dx} \\ L_2 &= \{\int \pi(x) f(x) dx\}^{n_1} \{1 - \int \pi(x) f(x) dx\}^{n - n_1} \end{aligned}$$

The first term is the contribution from $X|D = 1$, which is called biased sampling likelihood.

Example

$$\pi(x) = \exp(-x\theta), X \sim \exp(\lambda)$$

$$L_1 = \frac{\pi(x)f(x)}{\int \pi(x)f(x)dx} \propto \exp(-(\theta+\lambda)x)$$

In other words it is an exponential distribution with rate $\theta + \lambda$.
Based on L_1 alone it is not possible to identify θ and λ!
However the second term L_2 contributes an estimating equation

$$E[n_1/n] = P(D=1) = \int \pi(x)f(x)dx = \lambda \int \exp(-(\theta+\lambda)x)dx$$

Therefore it is possible to identify θ and λ by using $L = L_1 L_2$. The missing data likelihood L is more informative than the biased sampling likelihood L_1!

Biased sampling problem

Suppose the targeted population has a density $f(x)$. Due to various reason we can not have direct observations from $f(x)$. Instead, the observed data have density

$$X \sim \frac{w(x)f(x)}{\int w(x)f(x)dx}$$

$w(x) = x, x^2, x^3$ are corresponding to length biased, area biased and volume biased sampling, respectively.
Patil, G. P. , and Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, 179-189.

Nonparametric MLE for biased sampling data

Suppose

$$X_1, ..., X_n \sim \frac{w(x)dF(x)}{\int w(x)dF(x)}$$

where $w(x)$ is known. We are interested in estimating F! The likelihood is

$$L = \max \prod_{i=1}^{n} \frac{w(x_i)dF(x_i)}{\int w(x)dF(x)}$$

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} w(x_i)^{-1} I(x_i \leq x)}{\sum_{i=1}^{n} w^{-1}(x_i)}$$

In the special case $w(x) = 1$, $X_1, ..., X_n \sim F$, then

$$\hat{F}(x) = \sum_{i=1}^{n} n^{-1} I(x_i \leq x)$$

How do make inference from biased and unbiased data

$$X \sim f(x), \quad Y \sim \frac{w(y)f(y)}{\int w(y)f(y)dy}$$

The likelihood is

$$\prod_{i=1}^{n_0} dF(x_i)] \left[\prod_{j=1}^{n_1} \frac{w(y_j)dF(y_j)}{\int w(y)dF(y)} \right]$$

Vardi (1982, 1985) gave us answers!

Case and control study

The logistic regression model is given by

$$P(D=1|x) = \frac{\exp(\alpha + x\beta)}{1+\exp(\alpha + x\beta)}, \quad X \sim h(x)$$

Case data:

$$X_1, ..., X_{n_1} \sim f(x|D=1)$$

Control data

$$Z_1, ..., Z_{n_0} \sim f(X|D=0)$$

where n_1 and n_0 are fixed! In general

$$n_1/(n_1+n_0) \neq P(D=1)$$

Equivalency

Using Bayes' formula

$$f(x|D=1) = \frac{P(D=1|x)h(x)}{P(D=1)}, \quad f(x|D=0) = \frac{P(D=0|x)h(x)}{P(D=0)}$$

$$P(D=1|x) = \frac{\exp(\alpha + x\beta)}{1+\exp(\alpha + x\beta)}$$

$$f(x|D=1)/f(x|D=0) = \frac{P(D=0)}{P(D=1)}\exp(\alpha + x\beta) = \exp(\alpha^* + x\beta)$$

$$\alpha^* = \alpha + \log\{P(D=0)/P(D=1)\}$$

Semiparametric model

The exponential tilting model for two densities is given by

$$f(x) = \exp(\alpha + x\beta)g(x) = \frac{\exp(x\beta)g(x)}{\int \exp(x\beta)g(x)dx}$$

where the baseline density $g(x)$ is not specified! Qin and Zhang (1997).
This is similar to the two sample proportional hazard model (Cox model) where the two hazards $\lambda_1(t)$ and $\lambda_2(t)$ satisfies

$$\lambda_2(t) = \theta\lambda_1(t)$$

Examples

In normal case

$$f(x)/g(x) = \exp\{-(x-\mu_1)^2/\sigma^2 + (x-\mu_2)^2/\sigma^2\} = \exp(\alpha + x\beta)$$

$$\alpha = 0.5(\mu_2^2 - \mu_2^2)/\sigma^2, \quad \beta = (\mu_2 - \mu_1)/\sigma^2$$

In Poisson case

$$f(x)/g(x) = \exp(\alpha + x\beta)$$

$$\alpha = \lambda_2 - \lambda_1, \quad \beta = \log\lambda_2 - \log\lambda_1$$

What happens for $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$?

Examples

$$f(x)/g(x) = \exp\{-(x-\mu_1)^2/\sigma_1^2+(x-\mu_2)^2/\sigma_2^2\} = \exp(\alpha+x\beta+\gamma x^2)$$

What the is connection between $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ and (α, β, γ)?

Heckman selection bias sampling model

Heckman (1979 Econometrica).

$$Y_1 = x_1\beta + \epsilon_1, \quad Y_2 = x_2\gamma + \epsilon_2$$

where (ϵ_1, ϵ_2) are bivariate normal. Y_1 is observable if and only if $Y_2 > 0$

$$\begin{aligned} & P(Y_1 = y_1 | Y_2 > 0, x_1, x_2) \\ = & \frac{P(Y_2 > 0 | Y_1 = y_1, x_1, x_2) P(Y_1 = y_1 | x_1, x_2)}{P(Y_2 > 0 | x_1, x_2)} \\ = & \frac{\bar{F}_2(0 | y_1, x_1, x_2)\phi(y_1 | x_1)}{\Phi(x_2\gamma/\sigma_2)} \end{aligned}$$

where $\bar{F}_2(0|y_1, x_2, x_2)$ is the selection bias!

Inference

$$E[Y_1|Y_2 > 0] = x_1\beta + \rho\sigma_1 \frac{\phi(x_2\gamma)}{\Phi(z\gamma)} \neq x_1\beta$$

The complete data only inference is biased!
Heckman's two-stage estimator!

Choice based sampling

In econometrics, the case and control sampling is called choice based sampling.
Econometricians Daniel McFadden and James Heckman won the 2000 Nobel Prize in economics for their works on discrete choice models and selection bias.
Heckman, James J. (1979): Sample Selection Bias as a Specification Error, Econometrica 47, 153-161. (Heckman got the Nobel prize for this article.)

Case and control study

A leading biostatistician in the world N. Breslow (2003 Biometrics) complained Nobel Prize in physiology and medicine never award for work on case-control studies in biostatistics or epidemiology.

Call back problem in survey

Let Y be a random vector of interested in a survey. $D_i = 1$ if individual i responses in the i-th call.

$$P(D_i = 1|y) = \frac{\exp(\alpha_i + y\beta)}{1 + \exp(\alpha_i + y\beta)} := \pi_i(y), \quad i = 1, 2$$

We are interested in estimating $\mu = E(Y)$. The density of Y, $f(y)$ is arbitrary! Note that we have to assume β is common in the first and second calls in order the model is identifiable!

Some existing approaches

Alho (1990 Biometrika), Wood, White and Hotopf (2006 JRSSA), Troxel, Lipsitz and Brennan. (1997 Biometrics), Wang (1999 Biometrics), Daniels, M. J., Jackson, D., Feng, W., and White, I. R. (2015 Biometrics), Kim and Im (2014 Biometrika) and others?

Likelihood

$$
\begin{aligned}
L &= [\pi_1(y,\beta)f(y)]^{D_1}[\{1-\pi_1(y)\}\pi_2(y)f(y)]^{(1-D_1)D_2} \\
&\quad [\int \{1-\pi_1(y)\}\{1-\pi_2(y)\}f(y)dy]^{(1-D_1)(1-D_2)}
\end{aligned}
$$

Alho (1990, Biometrika)

$$
E\left[\frac{\{D_1+(1-D_1)D_2\}Y}{\pi_1(Y,\beta)+\{1-\pi_1(y,\beta)\}\pi_2(y,\beta)}\right] = E(Y) = \mu
$$

The problem is how to estimate α_1, α_2 and β?

Maximum semiparametric likelihood estimation

Qin and Follmann (2014 Biometrika) are able to find the maximum semiparametric likelihood estimate by discreting $F(y)$ for each of observed y_i's, i.e, $D_{1i} + \{1 - D_{1i}\}D_{2i} = 1$. It is complicated.
To illustrate the basic concept, we consider no call back problem next first.

No call back problem

Denote Y as salary in a survey. $D_i = 1$ is the i-th individual responses, and 0 otherwise.

$$P(D_i = 1|Y_i) = \pi(Y_i) = \exp(-\theta y)$$

where $\pi(Y)$ is a monotone decreasing function. Denote the observed data as

$$(Y_1, D_1 = 1), ..., (Y_{n_1}, D_{n_1} = 1), (?, D_{n_1+1} = 0), ..., (?, D_n = 0)$$

Likelihood

Assume $Y_i, i = 1, 2, ..., n \sim f(y)$. The likelihood is

$$L = \prod_{i=1}^{n_1} \{\pi(y_i) f(y_i)\} \prod_{i=n_1+1}^{n} \int \{1 - \pi(y)\} f(y) dx$$

It can be decomposed as $L = L_1 L_2$, where

$$\begin{aligned} L_1 &= \prod_{i=1}^{n_1} \frac{\pi(y_i) f(y_i)}{\int \pi(y) f(y) dy} \\ L_2 &= \{\int \pi(y) f(y) dy\}^{n_1} \{1 - \int \pi(y) f(y) dx\}^{n-n_1} \end{aligned}$$

The first term is the contribution from $y|D = 1$, which is called biased sampling likelihood.

Details

Let $\pi = E(D) = \int \pi(y) dF(y)$. Denote $dF(y_i) = p_i, i = 1, 2, ..., n_1$. We need to maximize

$$\ell = \sum_{i=1}^{n_1} [\log \pi(y_i) + \log p_i] + (n - n_1) \log(1 - \pi)$$

subject to the constraints

$$\sum_{i=1}^{n_1} p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^{n_1} p_i \pi(y_i) = \pi$$

Details

The profile empirical likelihood is

$$\ell = \sum_{i=1}^{n_1} [\log \pi(y_i, \theta) - \log\{1 + \lambda(\pi(y_i) - \pi)\}] + (n - n_1) \log(1 - \pi)$$

where λ is the Lagrange multiplier determined by

$$\sum_{i=1}^{n_1} \frac{\pi(y_i) - \pi}{1 + \lambda(\pi(y_i) - \pi)} = 0$$

Next maximize ℓ with respect to θ!

Call back likelihood

$$
\begin{aligned}
L &= [\pi_1(y,\beta)f(y)]^{D_1}[\{1-\pi_1(y)\}\pi_2(y)f(y)]^{(1-D_1)D_2} \\
&\quad [\int \{1-\pi_1(y)\}\{1-\pi_2(y)\}f(y)dy]^{(1-D_1)(1-D_2)}
\end{aligned}
$$

Let $\Delta(y) = E[D_1 + (1-D_1)D_2|y] = \pi_1(y) + \{1-\pi_1(y)\}\pi_2(y)$. It can be decomposed as

$$
\left[\frac{\pi_1(y)}{\Delta(y)}\right]^{D_1} \left[\frac{\{1-\pi_1(y)\}\pi_2(y)}{\Delta(y)}\right]^{(1-D_1)D_2}
$$

$$
[\Delta(y)dF(y)]^{D_1+(1-D_1)D_2} \left[\int \{1-\Delta(y)\}dF(y)\right]^{(1-D_1)(1-D_2)}
$$

Profile likelihood

Denote $p_i = dF(y_i)$ if $D_{1i} + (1 - D_{1i})D_{2i} = 1, i = 1, 2...n_1$.

$$\begin{aligned} \ell &= \sum_{i=1}^{n} [D_{1i} \log \pi_1(y_i) + (1 - D_{1i})D_{2i} \log\{(1 - \pi_1(y_i))\pi_2(y_i)\} \\ &+ \sum_{i=1}^{n_1} \log p_i + (n - n_1) \log[\sum_{i=1}^{n_1} p_i\{1 - \Delta(y_i)\}] \end{aligned}$$

We can profile out p_i subject to the constraint

$$\sum_{i=1}^{n_1} p_i = 1, \quad p_i \geq 0$$

After profiling out p_i's we can maximize it with respect to α_1, α_2 and β!

Challenge

In the Alho's model we have to assume a common slope

$$P(D_i = 1|y) = \frac{\exp(\alpha_i + y\beta)}{1 + \exp(\alpha_i + y\beta)} := \pi_i(y), \quad i = 1, 2$$

How can we relax this assumption?
1). Use auxiliary information?
2). Use instrument variables?
3). Sensitivity analysis?
4). Other methods? Suggestions?
Take home message "Non-ignorable nonresponse=Missing not at random problem=hard problem!!!"
The best solution is to collect data without missing values!

Acknowledgements

1. I would like to thank Professor Jae Kim for the arrangement of this talk.
2. I would like to thank my collaborators over the years! If you think anything is good in my talk, that is my contribution. You should blame my collaborators if you find something is bad in my talk.
3. Most importantly I would like to thank everybody in this room for your attention!