Adjusting for non-response in case-control studies

Yannan Jiang, Alastair Scott* & Chris Wild

Clinical Trials Research Unit and Department of Statistics
University of Auckland

a.scott@auckland.ac.nz

Introduction

In case-control sampling, we draw separate samples of cases (subjects with a condition of interest) and controls (subjects without the conditions) from some population of interest:

– very efficient when cases are relatively rare;

– widely used in practice.

An appreciable degree of non-reponse is common, particularly when some of the information is sensitive or intrusive.

A standard analysis can give misleading results if the chance of responding depends on unobserved variables and nothing is done to adjust for the non-response.

Example: The Women's Cardiovascular Health Study

The WCHS was a stratified population-based case-control study investigating the relationship between oral contraceptive use and the incidence of stroke, conducted in 3 counties around Seattle over a 5 year period. (Schwarz $et\ al$, 1997.)

Cases - women under 45 with a first stroke, whether fatal or non-fatal.
Tried to recruit all of them.
Controls were randomly sampled from 5 age groups using selection probabilities ranging from 4×10^{-5} for the $18 - 24$ age group to 58×10^{-5} for the $40 - 45$ age group.

Example: The Women's Cardiovascular Health Study

The WCHS was a stratified population-based case-control study investigating the relationship between oral contraceptive use and the incidence of stroke, conducted in 3 counties around Seattle over a 5 year period. (Schwarz $et\ al$, 1997.)

Cases - women under 45 with a first stroke, whether fatal or non-fatal. Tried to recruit all of them.
Controls were randomly sampled from 5 age groups using selection probabilities ranging from 4×10^{-5} for the $18-24$ age group to 58×10^{-5} for the $40-45$ age group.

There was a reasonable amount of non-response (29% among cases and 25% among the controls). All non-respondents were contacted by phone to obtain values of a few key variables (in particular contraceptive use) that were believed to influence response rates.

Introduction

The data from the WCHS study were analysed using weighted methods, originally developed for use in survey sampling, to adjust for non-response. (See Arbogast $et\ al$ (2002) for details.)

It is well-known that survey weighting leads to very inefficient estimates when the weights vary widely (e.g. Elliot & Little, 2000, recommend that w_{max}/w_{min} be no more than about 10).

In case-control studies the variation in weights is usually extremely large - $w_{max}/w_{min} = 2.5 \times 10^5$ in the example $\Rightarrow$ we should be able to get much better estimates.

Formal Set-Up

We have a binary response, Y, with $Y = 1$ denoting a case and $Y = 0$ a control, and a vector of potential explanatory variables, $\boldsymbol{x}$.

We want to fit a logistic regression model,

$$P\{Y = 1 \mid \boldsymbol{x}\} = \frac{e^{\boldsymbol{x}^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}^T\boldsymbol{\beta}}} = p_1(\boldsymbol{x}; \boldsymbol{\beta}) \quad \text{say},$$

for the conditional probability that someone with covariate values $\boldsymbol{x}$ is a case.

Formal Set-Up

We have a cohort of N units with values $\{(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, \ldots, N\}$, generated from the joint distribution Y and $\boldsymbol{x}$.

Suppose that we know the value of Y and some components of $\boldsymbol{x}$, say $\boldsymbol{x}^{(1)}$, for all members of the cohort.

We first partition the cohort into H strata, $S_1, \ldots, S_H$, based on the value of $\boldsymbol{x}^{(1)}$. Let $N_{\ell h}$ be the number of units with $Y = \ell$ in S_h.

Then we choose a sample of $n_{\ell h}$ of these units for further observation.

Let R_{1i} be an indicator variable taking the value 1 if the ith unit is in the chosen sample and 0 otherwise and set $\pi_{1i} = P(R_{1i} = 1)$.

Formal Set-Up

Some of the chosen units will respond and some will not.

We collect information on the remaining components of $\boldsymbol{x}$ from all respondents.

Let $R_{2i} = 1$ if the ith chosen unit responds and $R_{2i} = 0$ otherwise, and let $\pi_{2i} = P(R_{2i} = 1)$.

If we set $R_i = R_{1i}R_{2i}$, then we have complete information for those units with $R_i = 1$. This happens with probability $P(R_i = 1) = \pi_{1i}\pi_{2i}$

Formal Set-Up

If we had complete data on all members of the original cohort, we would estimate $\boldsymbol{\beta}$ from the logistic likelihood equations

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_1^N \mathbf{U}_i(\boldsymbol{\beta}) = \sum_1^N \boldsymbol{x}_i[y_i - p_1(\boldsymbol{x}_i;\boldsymbol{\beta})] = \mathbf{0}.$$

It is well-known that we can get very misleading results if we just leave out all the units with incomplete data and work with

$$\widehat{\mathbf{S}}_N(\boldsymbol{\beta}) = \sum_1^N R_i \mathbf{U}_i(\boldsymbol{\beta}).$$

However, there are ways of fixing this by adjusting the terms in this complete-case score function.

Estimating Equations

Consider the class of estimating equations $\mathbf{S}_0(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, with

$$\mathbf{S}_0(\boldsymbol{\beta}) = \sum_1^N R_i \mathbf{W}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\beta}, \pi) = \sum_1^N \mathbf{S}_{0i}$$

where $\mathbf{S}_{0i} = \mathbf{S}_{0i}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\beta}) = R_i \mathbf{W}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\beta})$ satisfies the condition $E\{\mathbf{S}_{0i} \mid \boldsymbol{x}_i\} = \mathbf{0}$ at the true value of $\boldsymbol{\beta}$.

Estimating Equations

Consider the class of estimating equations $\mathbf{S}_0(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, with

$$\mathbf{S}_0(\boldsymbol{\beta}) = \sum_1^N R_i \mathbf{W}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\beta}, \pi) = \sum_1^N \mathbf{S}_{0i}$$

where $\mathbf{S}_{0i} = \mathbf{S}_{0i}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\beta}) = R_i \mathbf{W}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\beta})$ satisfies the condition $E\{\mathbf{S}_{0i} \mid \boldsymbol{x}_i\} = \mathbf{0}$ at the true value of $\boldsymbol{\beta}$.

Note that this:

– only uses data from completely observed units;

– depends on π, the vector of selection probabilities.

Important Special Cases

- The Horvitz-Thompson estimator.

Here

$$\mathbf{S}_{0i}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\beta}, \pi) = \frac{R_i}{\pi_i} U_i(\boldsymbol{\beta})$$

with $\pi_i = P\{R_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{y}_i\} = \pi_{1i}\pi_{2i}$

This is the standard approach used to counteract selection bias in survey sampling [e.g.Binder(1983)].

It is not very efficient but has the big advantage that software is widely available (for linear and logistic models in all major packages, any GLM in Thomas Lumley's $Survey$ package in R).

Important Special Cases

- The Horvitz-Thompson estimator.
- The Sample Data (or Conditional) Likelihood estimator.

$$\mathbf{S}_{0i}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\beta}, \pi) = R_i \frac{\partial \log f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, R_i = 1)}{\partial \boldsymbol{\beta}}$$

This approach was developed, more or less independently, in many fields including econometrics [Hausman & Wise (1981)], biostatistics [Breslow & Cain (1988)], survey sampling [Pfefferman & Sverchkov (2003)].

Particularly simple for logistic regression where we simply add a constant offset $\log\{\pi(\boldsymbol{x}_i, \boldsymbol{y}_i = 1)/\pi(\boldsymbol{x}_i, \boldsymbol{y}_i = 0)\}$.

Other Examples

The class includes a number of other methods that have been suggested in the literature, including

– the "mean score" method of Reilly and Pepe (1995);

– the hybrid estimators of Jiang (2004).

Results for known π_is

Assume for the moment that π_i is known for every completely observed unit (true for the design phase of our example, not for the non-response).

Then, under mild conditions, $\widehat{\boldsymbol{\beta}}$, the solution to $\mathbf{S}_0(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, is consistent and asymptotically normal with asymptotic covariance matrix of the 'sandwich' form

$$\mathbf{I}_{00}^{-1}\mathbf{C}_{00}\mathbf{I}_{00}^{-1},$$

where $\mathbf{I}_{00} = E\{-\partial\mathbf{S}_0/\partial\boldsymbol{\beta}^T\}$ and $\mathbf{C}_{00} = Cov\{\mathbf{S}_0\}$.

Unknown π_is

What can we do if $\pi_i = \pi(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is not known?

Unknown π_is

What can we do if $\pi_i = \pi(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is not known?

A standard approach with missing data is to assume that π_i can be modelled by some parametric function, say

$$\pi_i = \pi(\widetilde{\boldsymbol{x}}_i, \boldsymbol{y}_i; \boldsymbol{\alpha}),$$

where $\widetilde{\boldsymbol{x}}_i$ contains a subset of the $\boldsymbol{x}$ variables, and that we can get enough information about the values of $\widetilde{\boldsymbol{x}}_i$ and $\boldsymbol{y}_i$ for incompletely observed units to make the parameter $\boldsymbol{\alpha}$ estimable.

Unknown π_is

Then, given a model $\pi(\widetilde{\boldsymbol{x}}_i, \boldsymbol{y}_i; \boldsymbol{\alpha})$, we can:

- substitute it for π_i in the expression for $\mathbf{S}_0$;
- augment the resulting equations, $\mathbf{S}_0(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}) = \mathbf{0}$ with another set of estimating equations for $\widehat{\boldsymbol{\alpha}}$, say $\mathbf{S}_1(\widehat{\boldsymbol{\alpha}}) = \mathbf{0}$;
- apply standard estimating equation methods to the enlarged system - i.e. set $\mathbf{S}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}) = \mathbf{0}$, where

$$\mathbf{S}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{pmatrix} \mathbf{S}_0(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ \mathbf{S}_1(\boldsymbol{\alpha}) \end{pmatrix}.$$

Unknown π_is

We have two sources of missing data here, so we have to build models for:

- π_{1i}, say $\pi_{1i} = \pi_1(\boldsymbol{x}_i^{(1)}, \boldsymbol{y}_i; \boldsymbol{\alpha}_1)$ based on information available for the whole cohort; and
- π_{2i}, say $\pi_{2i} = \pi_2(\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, \boldsymbol{y}_i; \boldsymbol{\alpha}_2)$ based on information collected from a sample of the non-respondents.

Estimated π_is

Then we set $\mathbf{S}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_2) = \mathbf{0}$, where

$$\mathbf{S}(\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \begin{pmatrix} \mathbf{S}_0(\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \\ \mathbf{S}_1(\boldsymbol{\alpha}_1) \\ \mathbf{S}_2(\boldsymbol{\alpha}_2) \end{pmatrix}.$$

The resulting estimators are consistent & asymptotically normal with asymptotic covariance matrix $\mathbf{I}^{-1}\mathbf{C}(\mathbf{I}^T)^{-1}$ where $\mathbf{C} = Cov\{\mathbf{S}\}$ and

$$\mathbf{I} = \begin{pmatrix} \mathbf{I}_{00} & \mathbf{I}_{01} & \mathbf{I}_{02} \\ 0 & \mathbf{I}_{11} & 0 \\ 0 & 0 & \mathbf{I}_{22} \end{pmatrix},$$

where $\mathbf{I}_{00} = E\left\{-\partial \mathbf{S}_1/\partial \boldsymbol{\beta}^T\right\}$ etc

Estimated π_is

Using

$$\mathbf{I}^{-1} = \begin{pmatrix} \mathbf{I}_{00}^{-1} & -\mathbf{I}_{00}^{-1}\mathbf{I}_{01}\mathbf{I}_{11}^{-1} & -\mathbf{I}_{00}^{-1}\mathbf{I}_{02}\mathbf{I}_{22}^{-1} \\ 0 & \mathbf{I}_{11}^{-1} & 0 \\ 0 & 0 & \mathbf{I}_{22}^{-1} \end{pmatrix}$$

then leads to

$$ACov\{\widehat{\boldsymbol{\beta}}\} = \mathbf{I}_{00}^{-1}\mathbf{C}_{00}\mathbf{I}_{00}^{-1} - \mathbf{I}_{00}^{-1}\mathbf{I}_{01}\mathbf{I}_{11}^{-1}\mathbf{C}_{01}^{T}\mathbf{I}_{00}^{-1} - \mathbf{I}_{00}^{-1}\mathbf{I}_{02}\mathbf{I}_{22}^{-1}\mathbf{C}_{02}^{T}\mathbf{I}_{00}^{-1}.$$

- The first term is the value of $ACov\{\widehat{\boldsymbol{\beta}}\}$ with known π_is;
- the second term measures the effect of estimating π_{1i};
- the third term measures the effect of estimating π_{2i}.

Estimated π_is

We know the values of $(R_{1i}, y_i, \boldsymbol{x}_{1i})$ for every unit in the whole cohort and we can use all this information to fit $\pi_1(\boldsymbol{x}_i^{(1)}, \boldsymbol{y}_i; \boldsymbol{\alpha}_1)$. If we do this efficiently, then $\mathbf{C}_{01} = \mathbf{I}_{01}$ and the effect of estimating π_1 becomes

$$\mathbf{I}_{00}^{-1}\mathbf{I}_{01}\mathbf{I}_{11}^{-1}\mathbf{I}_{01}^{T}\mathbf{I}_{00}^{-1}.$$

Since this is positive definite, $ACov\{\widehat{\boldsymbol{\beta}}\}$ is always smaller with estimated design probabilities than with known values.

Even when we know the design probabilities exactly, we are still better off using estimated values !!

Estimated π_is

This is not as strange as it might seem at first glance.

What we are really doing when we estimate the selection probabilities is bringing in extra information on the values of y_i and $\boldsymbol{x}_{1i}$ from the unsampled units in the cohort (just as with post-stratification and calibration in survey sampling).

(See Robins, Rotnitzky & Zhao, 1994), for the same sort of effect in a similar context.)

Estimated π_is

If we can obtain values of the relevant variables from **all** non-respondents, then a similar result holds for the effect of estimating the response probabilities:
If we observe $\boldsymbol{x}_{2i}$ for all non-respondents (as well as the respondents), and we use an efficient estimator for $\boldsymbol{\alpha}_2$, then $\mathbf{C}_{02} = \mathbf{I}_{02}$ and the effect of estimating π_2 becomes $\mathbf{I}_{00}^{-1}\mathbf{I}_{02}\mathbf{I}_{22}^{-1}\mathbf{I}_{02}^{T}\mathbf{I}_{00}^{-1}$, which is positive definite.

Again, even in the unlikely event that we knew the response probabilities, we would be better off estimating them.

The effect is less clear cut when we only have information on a sample of the non-respondents. Our simulations suggest that we need a reasonably large sample before estimation becomes beneficial.

Efficient Estimation

This is a very ad hoc way of using the information in the partially observed units, and it is reasonable to think that we should be able to use it more efficiently.

Can we?

The answer is essentially no, provided we fit a sufficiently rich model for the selection probabilities (although we can squeeze a little bit more information out of the complete data).

Efficient Estimation

Suppose that we get information on $\boldsymbol{x}_2$ from **all** non-respondents and that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ both have finite support.

In this case, we can find an explicit expression for the semi-parametric efficient estimator (Lee, Scott & Wild 2010). This turns out to be equal to the conditional likelihood estimator, but with $(\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ treated as the parameter rather than just $\boldsymbol{\beta}$, augmented by selection probabilities estimated from saturated models in $(y, \boldsymbol{x}_1)$ and $(y, \boldsymbol{x}_1, \boldsymbol{x}_2)$ respectively.

This suggests a way of generating good estimators in situations when fully efficient ones are not available (e.g. when we only have information on $\boldsymbol{x}_2$ from a sample of non-respondents, or when some components of $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$ are continuous).

Simulation results

We carried out a series of simulations to compare the performance of Horvitz-Thompson, conditional likelihood and fully efficient estimators based on a simplified version of the fitted model from the WCHS example using just age, x_3, and oral contraceptive use, x_2, and age, x_3. (x_1 was a categorised version of age.)

The probability of non-response was generated using a logistic model containing stroke status (y) and contraceptive use, but not age. We generated estimates using the known response probabilities, probabilities fitted using the correct model and using a saturated model containing extra terms for age groups. We also looked at estimates based on samples of 10% and 50% of the nonrespondents

Efficiencies are calculated relative to the semi-parametric efficient estimator based on data from all the non-respondents.

Simulation results

Table 1. Efficiency of Weighted and CML estimators

	β_2 (OC use)		β_3(age)	
	Wtd	CML	Wtd	CML
Known response probs	71.7	73.8	43.4	53.2
Fitted: True model	94.4	97.5	43.5	53.2
Augmented model	97.9	99.8	53.3	68.5
10% NR sample	7.2	51.2	45.5	64.5
50% NR sample	19.2	78.0	49.6	67.3