Unbiased estimation of exposure odds ratios in complete records logistic regression

Jonathan Bartlett

London School of Hygiene and Tropical Medicine
www.missingdata.org.uk

Centre for Statistical Methodology
30th January 2015

Acknowledgements

This is joint work with:

- Ofer Harel (University of Connecticut)
- James Carpenter (LSHTM)

We are grateful to the UK Civil Aviation Authority for allowing the use of the data for the illustrative analysis, and to Bianca De Stavola for suggesting and facilitating the use of these data and providing useful comments on the manuscript.

I am grateful for support from the Medical Research Council (MR/K02180X/1).

Outline

When is the exposure odds ratio unbiased?

Application in practice

Illustrative example

Concluding comments

Outline

Setting and question

- Binary outcome Y, exposure X, confoudders C.
- In general X and/or C could be continuous, and we can use logistic regression for $Y|X, C$.
- We may have missing values in Y, X or one or more components of C.
- Let $R = 1$ denote complete records and $R = 0$ denote incomplete records.
- If we drop those with missing values, when will the complete records analysis (CRA) exposure (log) odds ratio $\hat{\beta}_X$ be (asymptotically) unbiased?

A simplification

- First we assume that there no confounders C, and that exposure X is binary.
- We can display the full data as

$$\begin{array}{cc|cc} & & \multicolumn{2}{c}{Y} \\ & & 0 & 1 \\ \hline X & 0 & a & b \\ & 1 & c & d \end{array}$$

- Odds ratio $= \frac{a \times d}{b \times c}$

Outcome dependent missingness

- Suppose missingness is outcome dependent, with $P(R = 1|X, Y = y) = k_y,\ y = 0, 1$.
- Then the expected complete records table is

		Y	
		0	1
X	0	$k_0 a$	$k_1 b$
	1	$k_0 c$	$k_1 d$

- Odds ratio $= \frac{k_0 a \times k_1 d}{k_1 b \times k_0 c} = \frac{a \times d}{b \times c}$
- i.e. unbiased
- This result of course justifies the validity of odds ratios in case-control studies.

Exposure dependent missingness

- Now suppose missingness is exposure dependent, with $P(R=1|X=x,Y)=k_x$, $x=0,1$.
- Then the expected complete records table is

		Y	
		0	1
X	0	$k_0 a$	$k_0 b$
	1	$k_1 c$	$k_1 d$

- Odds ratio $= \frac{k_0 a \times k_1 d}{k_0 b \times k_1 c} = \frac{a \times d}{b \times c}$
- i.e. unbiased
- This result (covariate dependent missingness) holds in fact more generally, e.g. for risk ratios and other regression models.

Confounders

- Now suppose we have a categorical confounder C with levels $c = 1, .., L$.
- Let the full data exposure/outcome table for level c of the confounder be

		Y	
		0	1
X	0	a_c	b_c
	1	c_c	d_c

- The full data odds ratio for the exposure effect at confounder level $C = c$ is $\frac{a_c \times d_c}{b_c \times c_c}$
- Assuming no effect modification, we then estimate the odds ratio by pooling across confounder levels, e.g. with Mantel-Haenszel or logistic regression.

Outcome/confounder dependent missingness

- Now suppose we have missingness dependent on Y and C, i.e. $P(R=1|Y=y,X,C=c)=k_{y,c}$, $y=0,1$, $c=1,..,L$.
- The expected complete record table is then

		Y 0	Y 1
X	0	$k_{0,c}a_c$	$k_{1,c}b_c$
	1	$k_{0,c}c_c$	$k_{1,c}d_c$

- Odds ratio $= \frac{k_{0,c}a_c \times k_{1,c}d_c}{k_{1,c}b_c \times k_{0,c}c_c} = \frac{a_c \times d_c}{b_c \times c_c}$.
- i.e. still unbiased

Exposure/confounder dependent missingness

- Now suppose we have missingness dependent on X and C, i.e. $P(R = 1|Y, X = x, C = c) = k_{x,c}$, $x = 0, 1,\ c = 1, .., L$.
- The expected complete record table is then

		Y 0	Y 1
X	0	$k_{0,c}a_c$	$k_{0,c}b_c$
	1	$k_{1,c}c_c$	$k_{1,c}d_c$

- Odds ratio $= \frac{k_{0,c}a_c \times k_{1,c}d_c}{k_{0,c}b_c \times k_{1,c}c_c} = \frac{a_c \times d_c}{b_c \times c_c}$.
- i.e. still unbiased

Exposure/outcome dependent missingness

- In general if missingness depends on X and Y, we will obtain biased estimates of the exposure effect.
- A special case exists however where we still obtain unbiased estimates.
- Specifically if $P(R = 1|Y, X, C) = s(X, C) \times t(Y, C)$ for some functions $s(X, C)$, $t(Y, C)$.
- This result can be viewed as applying the previous results in turn.

Exposure/outcome dependent missingness continued

- This could arise for example if two variables are partially observed with one mechanism depending on (X, C) and the second depending on (Y, C).
- If we let R_1 and R_2 denote the observation indicators for the two variables, we need

$$\begin{aligned} P(R = 1|Y, X, C) &= P(R_1 = 1, R_2 = 1|Y, X, C) \\ &= P(R_1 = 1|X, C) \times P(R_2 = 1|Y, C) \end{aligned}$$

Summary

The results apply more generally, to continuous exposure X, multiple confounders C, with logistic regression used for estimation.

Letting $\hat{\beta}_0$, $\hat{\beta}_X$, $\hat{\beta}_C$ denote the corresponding log odds ratios, we have

Missingness dependent on	$\hat{\beta}_0$	$\hat{\beta}_X$	$\hat{\beta}_C$
Neither Y, X nor C	Unbiased	Unbiased	Unbiased
Y	Biased	Unbiased	Unbiased
X, C	Unbiased	Unbiased	Unbiased
Y, C	Biased	Unbiased	Biased
Y, X, C	Biased	Biased*	Biased

* in general

MCAR/MAR/MNAR

- So far we have not said which variable(s) have missing values.
- Depending on what type of mechanism is assumed to be in action and which variable(s) have missing values, missingness could be MCAR, MAR or even MNAR.
- e.g. if exposure is partially observed, with missingness dependent on exposure level, and possibly confounders, this is MNAR.
- But so long as missingness is independent of Y, given X and C, $\hat{\beta}_X$ is unbiased.

Extensions

- If we include interactions between X and some components of C, the results continue to apply.
- With survival data, if the event rate is low and follow-up similar for subjects, the results also apply approximately due to link between logistic and Cox regression.
- In this case the event indicator takes the place of the binary outcome Y.

Caveat

- The arguments implicitly assume that the outcome model is correctly specified.
- i.e. no effect interactions (if not included).
- Provided any misspecifications are not severe, results should still approximately apply (see illustrative) example.

Outline

Application in practice

- The results depend on how the probability of being a complete record depend on (Y, X, C).
- In some cases our study specific knowledge may support one of the sufficient conditions needed for β_X to be estimated without bias.

Missingness in a confounder

Divide the confounders into $C = (C_1, C_2)$, with C_1 partially observed

Missingness in C_1 found related to	Plausible missingness mechanism	$\hat{\beta}_X$ unbiased
C_2	C_2	Yes
X and possibly C_2	X and C_2	Yes
Y and possibly C_2	Y and C_2	Yes
	X and Y	Generally no
X, Y and possibly C_2	C and X	Yes
	C and Y	Yes

Missingness in exposure

Missingness in X found related to	Plausible missingness mechanism	$\hat{\beta}_X$ unbiased
C	C	Yes
Y	Y	Yes
	C and Y	Yes
C and Y	X and Y	Generally no
	X and C	Yes

Missingness in outcome

Missingness in Y found related to	Plausible missingness mechanism	$\hat{\beta}_X$ unbiased
X	X	Yes
C	C	Yes
	X and C	Yes
X and C	Y and C	Yes
	X and Y	Generally no

Investigating missingness

- So in some situations we may from the data be able to make some tentative conclusions regarding missingness and thus the unbiasedness of $\hat{\beta}_X$.
- With missingness in multiple variables things inevitably become more complex.
- Here our substantive knowledge is crucial to judge the plausibility of assumptions.

Outline

Illustrative example

- We illustrate using data from a cohort study of professional pilots in the UK [1].
- This study's aim was to include all professional flight crew who held a license at some point between 1989 and 1999.
- For our analyses, follow-up starts at recruitment, where various variables of interest were recorded.
- Crew data was linked to UK health registers to obtain vital status up to 2006.

Outcome model specification

- For this illustrative analysis, we consider a binary outcome Y representing death in the first 15 years of follow-up.
- We consider a exposure 'number of accrued flying hours' at baseline, categorised into < 400 hours, $400 - 5499$, and > 5500 hours.
- Confounders were age at entry, smoking status, BMI and type of route.
- Our analyses here use data from 11,841 male crew, who had complete data at baseline for exposure and confounders.

Simulation setup

- We first fitted a logistic regression model to the full data.
- For each of 8 missingness mechanisms, we make data missing and perform the complete records analysis.
- For each mechanism we repeat 10,000 times, and present the mean estimates across 10,000 realizations of missingness process.

Results

Log odds ratios

Missingness dependent on	400-5499 hours vs $<$ 400 hours	$\geq$ 5500 hours vs $<$ 400 hours
Full data	0.55	0.59
MCAR	0.57	0.61
Event indicator (Y)	0.56	0.61
Age (C)	0.51	0.54
Age and flying hours (C, X)	0.52	0.56
Event indicator and age (Y, C)	0.66	0.74
Event indicator and flying hours (Y, X)	1.59	2.64
Event indicator and flying hours* (Y, X)	0.59	0.67

* $P(R = 1 | Y, X, C) = s(X)t(Y)$

Interpretation

- For the most part, results are as expected from theory.
- For missingness dependent on outcome and confounder (age), estimates are somewhat different to full data estimates.
- This is likely due to the fact the outcome model is not exactly correctly specified: the exposure effect varies somewhat by age.
- Missingness dependent on C means complete records have different C distribution, leading to some small bias.

Outline

When is the exposure odds ratio unbiased?

Application in practice

Illustrative example

Concluding comments

Summary

- Estimates of exposure effect from complete records logistic regression are unbiased under a wide range of missingness assumptions [2].
- Most of the results presented are not new [3, 4, 5], but perhaps are not widely appreciated.
- Directed acyclic graphs can be useful for considering possible missingness mechanisms and thus plausiblity of assumptions [6].

Summary

- Of course even when CRA is unbiased, it is not efficient.
- If missingness is in confounders, alternative approaches will gain efficiency.
- If the data are missing at random, multiple imputation can be used.
- If a confounder is missing not at random, but independently of outcome, a weighting/imputation approach can be used to improve upon CRA efficiency [7].

References I

[1] B DeStavola, C Pizzi, F Clemens, S A Evans, A D Evans, and I dos Santos Silva.

Cause-specific mortality in professional flight crew and air traffic control officers: findings from two uk population-based cohorts of over 20,000 subjects.

International Archives of Occupational and Environmental Health, 85:283–293, 2012.

[2] J W Bartlett, O Harel, and J R Carpenter.

Unbiased estimation of exposure odds ratios in complete records logistic regression.

under review, 2015.

References II

[3] W Vach and M Blettner.

Biased Estimation of the Odds Ratio in Case-Control Studies due to the Use of Ad Hoc methods of Correcting for Missing Values for Confounding Variables.

American Journal of Epidemiology, 134:895–907, 1991.

[4] M A Hernn, S H Hernndez-Diaz, and J M Robins.

A structural approach to selection bias.

Epidemiology, 15:615–625, 2004.

[5] D Westreich.

Berksons Bias, Selection Bias, and Missing Data.

Epidemiology, 23:159–164, 2012.

[6] R M Daniel, M G Kenward, S N Cousens, and B L de Stavola.

Using causal diagrams to guide analysis in missing data problems.

Statistical Methods in Medical Research, 21:243–256, 2012.

References III

[7] J W Bartlett, J R Carpenter, K Tilling, and S Vansteelandt.

Improving upon the efficiency of complete case analysis when covariates are MNAR.

Biostatistics, 15:719–730, 2014.