Analysis of binary outcomes with missing data: missing = smoking, last observation carried forward, and a little multiple imputation

Presenter: Boyoun Chung

Authors: Donald Hedeker, Robin J. Mermelstein, Hakan Demirtas

STAT 5095, Department of Statistics, UConn

Table of Contents

Table of Contents

Background

- Binary outcomes are common in many research areas, notably in studies of drug use, alcohol and smoking.
- How to handle missing values in an outcome variable has been a continuing source of controversy and debate [Schafer and Graham, 2002].
- In this paper, authors consider this problem in a simple two-group design (e.g. control versus treatment), where the groups are compared in terms of a single binary outcome [Hedeker et al., 2007].
- Also, authors focus on smoking behavior (currently smoking or not) as the outcome.

Background of deterministic imputations

- A common approach to dealing with missing smoking outcomes is to recode them as "smoking".
- This is often called a "conservative" assumption, in the sense that it is conservative to assume the worse of the two values of the outcome (i.e. smoking).
- Another relatively common deterministic imputation approach, is to recode the missing data on the final smoking outcome to the last available smoking.
- This procedure is termed "last observation carried forward", or LOCF.
- That is used in studies with more than one time-point where interest focuses on comparing two groups at a final time-point (longitudinal study).

Disadvantages of deterministic imputations (1)

- Deterministic imputations, like missing = smoking and LOCF, can be shown to yield severely biased results for several reasons [Schafer and Graham, 2002].
- First, if the missing data are related to group membership (i.e. control or treatment), then the comparison of groups on the outcome is confounded with the missing data.
- For example, if there are more missing data in the control group, then the "conservative" missing = smoking assumption yields a test that favors the treatment group.
- Similarly, under LOCF, if time of dropout is related to group (i.e. control subjects drop out early, whereas treatment subjects drop out late or not at all), then group comparisons of the outcome are confounded with the amount and/or time of treatment.

Disadvantages of deterministic imputations (2)

- A further problem is that these deterministic imputations produce variance estimates that are artificially small, because the missing data are treated as a constant.
- This leads to biased estimates of association involving the recoded outcome variable, and measures of uncertainty (i.e. variances, standard errors) that are too small.
- This yields standard errors of estimated associations are too small, resulting artificially low p-values and rates of Type I error that are higher than nominal levels.
- In the rest of this paper, authors illustrate why these deterministic imputations are logically problematic, and how they can be improved upon easily by positing them as non-perfect relationships.

The relationship between smoking behavior and group

- Suppose that there are two groups (treatment and control) and we are interested in whether smoking (yes or no) varies by group at a single time-point.
- A simple 2×2 table of group by smoking, with a Pearson's χ^2 test, can address this.
- Unfortunately, it may be complicated by the fact that not everyone has a measurement on the smoking variable.

Table: 1. Table of smoking outcome by treatment group

	Smoking	Not smoking	Total
Control	m_{11}	m_{12}	$m_{1.}$
Treatment	m_{21}	m_{22}	$m_{2.}$
Total	$m_{.1}$	$m_{.2}$	N

The relationship between smoking and missingness

- We will describe a way in which we can classify these missing individuals in the 2×2 table of group by smoking, based on an assumed relationship between missingness and smoking (e.x, odds ratio).
- While unknown, the numbers in the second row of Table 2 can be expressed as a simple function of the numbers in the first row and the odds ratio, denoted θ.
- We have $\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$, which reflects the association between missingness and smoking.
- Note that n_{21} and n_{22} are unobserved, though $n_{2.}$ is known.

Table: 2. Table of missingness by smoking outcome

	Smoking	Not smoking	Total
Observed	n_{11}	n_{12}	$n_{1.}$
Missing	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

Using assumed relationship between smoking and missingness

- Note that $\hat{\theta}$ can be written as follows.

$$\frac{n_{22}}{n_{21}} = \hat{\theta} \times \frac{n_{12}}{n_{11}} \tag{1}$$

- Since $n_{22} + n_{21} = n_{2\cdot}$ is known, we can specify n_{22} and n_{21} if $\hat{\theta}$ is known as θ^*.

- Specifically, we assume the probability of smoking for missing individuals under θ^* (i.e., $\pi(\theta^*)$) and the observed odds of smoking (i.e., n_{12}/n_{11}) for the non-missing subjects.

$$\pi(\theta^*) = Pr(smoking \mid missing, \theta^*) = \frac{\theta^* \times n_{12}/n_{11}}{1 + \theta^* \times n_{12}/n_{11}} \tag{2}$$

$$n_{21} = n_{2\cdot} \times \pi(\theta^*)$$
$$n_{22} = n_{2\cdot} \times (1 - \pi(\theta^*))$$

- After we obtain the complete table of Missing vs Smoking, we can specify the 2×2 table of smoking outcome by treatment group.

Table of Contents

Example: the smoking cessation study [Gruder et al., 1993]

- Gruder et al. [1993] evaluated the effectiveness of adding group-based treatment adjuncts to a smoking cessation intervention comprised of a television program and self-help materials.
- The primary hypothesis of the study was that participants in the social support training condition would have **higher** abstinence rates than those in the general group discussion condition.
- Participants were smokers who registered for the television-based program (to receive the self-help materials) and who indicated an interest in attending group-based meetings.

The smoking cessation study (continued)

- Although participants were randomized to condition, only 50% of the participants scheduled to attend a group session came to at least one meeting.
- Thus, for analysis, the original report considered four 'conditions': the two group treatments, a 'no show' group consisting of individuals who were randomized to a group treatment but never showed up for any group meetings, and the no-contact control condition.
- Here, for simplicity and exposition, we combine the controls and no-shows as one group and the two active treatments as a second group.
- Hereafter, we refer to this first group as control and the second as treatment.

Missing vs smoking status at the final time point

Table: 3. Table of missingness (Miss) by Smoke, [Gruder et al., 1993]

	Not smoking	Smoking	Total
Observed	78	294	372
Missing	n_{21}	n_{22}	117
Total	$n_{.1}$	$n_{.2}$	489

- Table 3 lists the frequencies for the cross-tabulation of missing by smoking status at the final study time-point (i.e. at 24 months).
- For the observed individuals, the odds of smoking equal to $294/78 = 3.77$.
- The amount of missing data is rather large, accounting for approximately 24% of the total number of individuals.
- Thus, it would seem that the statistical analysis and resulting conclusions could be influenced by whatever is assumed for the missing data.

Result of complete-cases analysis

- Analyzing only the available data (i.e. $n = 372$), the smoking rate in the control group is $176/216 = 81.48\%$ and $118/156 = 75.64\%$ in the treatment group.
- The simple Pearson's χ^2 test yields $X = 1.86$, d.f. $= 1$, $p - value = 0.086$ (one-sided), which is not significant.
- If one assumes "missing = smoking", then the smoking rates are $259/299 = 86.62\%$ and $152/190 = 80.00\%$ for the control and treatment groups, respectively.
- In this case, we obtain $X = 3.80$, d.f. $= 1$, $p - value = 0.025$ (one-sided p-value), which is significant at $\alpha = 0.05$.

The smoking cessation study (continued)

- Why are the results from these two analyses different?
- Of the 117 missing individuals, 83 are from the control group and 34 from the treatment group, while the total number of control and treatment subjects equals 299 and 190, respectively.
- Thus, missingness is more common among the control group ($83/299 = 27.8\%$) than the treatment group ($34/190 = 17.9\%$), and so the "conservative" missing = smoking assumption yields many more smokers in the control than the treatment group.

The smoking cessation study (continued)

- As an alternative to deterministically recoding missing to smoking, we can examine what happens under various assumed values of the θ^* for missing and smoking.

- For example, assuming that $\theta^* = 2$ (i.e. the odds of smoking are double in the missing subjects than the non-missing subjects).

$$\pi(2) = \frac{2 \times 294/78}{1 + 2 \times 294/78} = 0.8829 \tag{3}$$

- This implies an assumed smoking rate of 88.29% for missing individuals.

- As discussed in Eq. (2), the number of missing individuals who are smoking is calculated as $n_{22} = 0.8829 \times 117 = 103.2993$, and the number of non-smoking missing individuals is $n_{12} = 117 - 103.2993 = 13.7007$.

- These yield an odds of smoking of 7.54 for missing individuals (i.e. a doubling of the observed odds of smoking), implying that missing individuals have nearly an 8 to 1 odds of smoking under the $\theta^* = 2$ assumption.

The smoking cessation study (continued)

Table: 4. Table of Group by Smoke under $\theta^* = 2$ for Miss and Smoke.

	Not smoking	Smoking	Total
Treatment	38 + 3.9820	118 + 30.0180	190
Control	40 + 9.7207	176 + 73.2793	299
Total	91.7027	397.2973	489

- The number of missing smokers in the two groups can be calculated by multiplying this probability by the number of missing individuals in the group.

$$n_{22c} = 0.1171 \times 83 = 9.7207, \quad n_{22t} = 0.1171 \times 34 = 3.9814 \tag{4}$$
$$n_{22c} = 0.8829 \times 83 = 73.2793, \quad n_{22t} = 0.8829 \times 34 = 30.0180$$

- Using these calculated frequencies to augment the observed data yields the results presented in Table 4.
- The Pearson's χ^2 test yields $X = 2.28$, d.f. $= 1$, $p - value = 0.066$ (one-sided), which is not significant.
- This calculation can be repeated under other assumed values of the OR to obtain a sense of the robustness of the results.
- The one-sided p-values decrease to 0.05, 0.045 and 0.04 for $\theta^* = 3$, 4, and 5, respectively.

Table of Contents

Using previous smoking information

Table: 5. Table of missingness (Miss) by Smoke stratified by prior smoking outcome (Smoke()).

Smoke() = Non-smoking	Not smoking	Smoking	Total
Observed	$n_{111} = 42$	$n_{112} = 71$	$n_{11\cdot} = 113$
Missing	n_{121}	n_{122}	$n_{12\cdot} = 37$
Total	$n_{1\cdot1}$	$n_{1\cdot1}$	$n_{1\cdot\cdot} = 150$

Smoke() = Smoking	Not smoking	Smoking	Total
Observed	$n_{211} = 36$	$n_{212} = 223$	$n_{21\cdot} = 259$
Missing	n_{221}	n_{222}	$n_{22\cdot} = 80$
Total	$n_{2\cdot1}$	$n_{2\cdot1}$	$n_{2\cdot\cdot} = 339$

- Another simple imputation technique that is common in longitudinal studies is "last observation carried forward" (LOCF).
- LOCF posits a perfect relationship in the sense that missing observations are related perfectly to previously observed values.
- For this, consider the 2×2 cross-tabulation of missing by smoking status in Table 5, stratified by smoking status at a previous time-point.
- The smoking assessment at the previous time-point is denoted as Smoke().

Using information from previous smoking behavior

- The LOCF imputation would set all missing observations in the upper Table 5 to non-smoking (i.e. $n_{121} = 37$ and $n_{222} = 0$) and all missing observations in the lower Table 5 to smoking ($n_{221} = 0$ and $n_{222} = 80$).
- This would produce θ^* going to zero and positive infinity in the two tables, respectively.
- Such deterministic imputation can be improved easily upon using the ideas of the θ^*.
- To improve the result, we employ pre-assumed odds ratio between Missing and Smoking from the previous time point.

Using assumed odds ratio between missingness and smoking

- Define $\pi_i = Pr(\textit{smoking at time } i \mid \textit{missing at time } i, \theta^*)$ as follows

$$\pi_i = \frac{\theta_i^* \times n_{i12}/n_{i11}}{1 + \theta_i^* \times n_{i12}/n_{i11}}, \quad i = 1, 2 \tag{5}$$

- Here, θ_i^* is the assumed odds ratio for the ith table, and n_{i12}/n_{i11} are the observed odds of smoking for the ith table. π_i is the probability of smoking for the missing individuals under the assumed θ_i^* of the ith table.

- Note that, the observed odds of smoking equal to $71/42 = 1.6905$ and $223/36 = 6.1944$, depending on previous smoking status.

- LOCF implies $\theta_1^* = 0$, which seems strong assumption.

Probability of smoking for the missing individuals

- In this example, continuing with the assumption of $\theta_1^* = \theta_2^* = 2$ for missing and smoking, we can calculate π_1 and π_2 as follows.

$$\pi_1 = \frac{2 \times 71/42}{1 + 2 \times 71/42} = 0.7717 \tag{6}$$

$$\pi_2 = \frac{2 \times 223/36}{1 + 2 \times 223/36} = 0.9253$$

- Thus, among those who are missing at the current time-point, subjects who were smoking at the previous time point have a very high assumed probability of smoking (0.9253)

- On the other hand, subjects who were not smoking at the previouis time point have a lower smoking probability (0.7717).

- It is known that $n_{22\cdot} = 80$ people who smoked in the previous time are divided into 61 for control group and 19 for treatment group.

- Similarly, $n_{12\cdot} = 37$ people who did not smoke in the previous time are divided into 22 for control group and 15 for treatment group.

Estimating frequencies based on previous information

- Using π_1, π_2, and knowing the numbers of missing control and treatment subjects in each of these two tables, we can calculate the numbers of missing individuals who are smoking in the control and treatment groups as in Eq. (7)

$$
\begin{aligned}
n_{112c} &= (1 - 0.7717) \times 22 = 5.0226, \; n_{112t} = (1 - 0.7717) \times 15 = 3.4245 \\
n_{212c} &= (1 - 0.9253) \times 61 = 4.5567, \; n_{212t} = (1 - 0.9253) \times 19 = 1.4193 \\
n_{122c} &= 0.7717 \times 22 = 16.9774, \; n_{122t} = 0.7717 \times 15 = 11.5755 \\
n_{222c} &= 0.9253 \times 61 = 56.4433, \; n_{222t} = 0.9253 \times 19 = 17.5807
\end{aligned} \tag{7}
$$

- The smoking rates for the control and treatment groups now equal 83.42% and 77.45%, respectively, and the χ^2 test statistic is $X = 2.70$, d.f. $= 1$, $p - value = 0.055$ (one-sided).

Group	Not smoking	Smoking	Total
Treatment	38 + 3.4245 + 1.4193	118 + 11.5755 + 17.5807	190
Control	40 + 5.0226 + 4.5567	176 + 16.9774 + 56.4433	299
Total	92.4231	396.5769	489

Odds ratios from different assumption (θ^*)

Table: 6. Group by Smoke analyses under different missing data assumptions.

Scenario	Smoking frequencies (proportions)			
	Control	Treatment	χ^2	p-value
Available data (n = 372)	176/216 (81.48)	118/156 (75.64)	1.87	0.085
Missing = smoking (n = 489)	259/299 (86.62)	152/190 (80.00)	3.80	0.025
Marginal $\theta^* = 1$	241.60/299 (80.80)	144.87/190 (76.25)	1.45	0.114
Marginal $\theta^* = 2$	249.28/299 (83.37)	148.02/190 (77.90)	2.28	0.065
Marginal $\theta^* = 5$	254.82/299 (85.22)	150.29/190 (79.10)	3.07	0.039
Stratified $\theta_1^* = \theta_2^* = 1$	242.34/299 (81.05)	143.78/190 (75.68)	2.02	0.077
Stratified $\theta_1^* = \theta_2^* = 2$	249.42/299 (83.42)	147.16/190 (77.45)	2.70	0.050
Stratified $\theta_1^* = \theta_2^* = 5$	254.76/299 (85.21)	149.82/190 (78.85)	3.28	0.035

- Table 6 summarizes the results of this section for θ^* values of 1, 2 and 5.
- The 'marginal' results do not consider the previous smoking status, while the 'stratified' results consider the previous smoking status.
- $\theta^* = 1$ would imply that missing and smoking are independent, whereas a $\theta^* = 5$ implies strong relationship.

Table of Contents

Discussion

- In this paper, authors have described a relatively simple approach for dealing with missing data in two-group studies with a binary outcome.
- The focus has been to use the ideas behind missing = smoking and LOCF in a relational manner.
- We have argued that it is important to examine results under a range of plausible values for the association of missing and smoking, stratified by past smoking behavior.
- By performing a sensitivity analysis one can determine the robustness of results to the assumed association of missing and smoking.
- In this regard, Schafer and Graham [2002] note that "one hopes that similar conclusions will follow from a variety of realistic alternative assumptions; when that does not happen, the sensitivity should be reported".
- Thus, a major point of the argument is that blindly assuming "missing = smoking" provides a very unrealistic solution.

Reference

Charles L Gruder, Robin J Mermelstein, Susan Kirkendol, Donald Hedeker, Siu Chi Wong, Janice Schreckengost, Richard B Warnecke, Rebecca Burzette, and Todd Q Miller. Effects of social support and relapse prevention training as adjuncts to a televised smoking-cessation intervention. *Journal of consulting and clinical psychology*, 61(1):113, 1993.

Donald Hedeker, Robin J Mermelstein, and Hakan Demirtas. Analysis of binary outcomes with missing data: missing= smoking, last observation carried forward, and a little multiple imputation. *Addiction*, 102(10):1564–1573, 2007.

Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.