Analysis of Binary Outcomes with Missing Data: Missing=Smoking, Last Observation Carried Forward, and a Little Multiple Imputation

Donald Hedeker, Robin J. Mermelstein, and Hakan Demirtas
University of Illinois at Chicago

1

Smoking Cessation Studies

- Outcome is binary (yes/no)

- Often longitudinal

- Two-group design (control/tx) considered here
  - approach is easily extended to multi-group designs

- Same issues for other binary outcomes (*e.g.*, alcohol, drug use)

2

Missing Smoking (y/n) Outcomes

- Common in smoking cessation studies
- Missing=Smoking
  - posits a perfect relationship
  - typically favors the treatment group
- Last Observation Carried Forward (LOCF)
  - posits a perfect relationship
  - group and time are often confounded

Probabilistic Versions of Deterministic Imputations

- Missing subjects are perhaps more likely to be smoking
- Past outcomes are related to future (missing) outcomes
- Provides a sensitivity analysis for missing data assumptions
- Very do-able with standard software

Relationship Between Smoking & Missingness

2 x 2 table of `Miss` by `Smoke` for the sample of n individuals

	Smoke		
Miss	no	yes	total
no	n_{11}	n_{12}	$n_{1.}$
yes	n_{21}	n_{22}	$n_{2.}$
total	$n_{.1}$	$n_{.2}$	n

observed: $n_{11}, n_{12}, n_{1.}, n_{2.}$, and n

missing: n_{21} and n_{22} (and thus $n_{.1}$ and $n_{.2}$)

Missing=smoking posits $n_{22} = n_{2.}$ and $OR = \infty$

	Smoke		
Miss	no	yes	total
no	n_{11}	n_{12}	$n_{1.}$
yes	n_{21}	n_{22}	$n_{2.}$
total	$n_{.1}$	$n_{.2}$	n

notice $\frac{n_{22}}{n_{21}} = OR \frac{n_{12}}{n_{11}}$, and so

$$n_{22} \; = \; n_{2.} \frac{OR \times odds_1}{1 + (OR \times odds_1)} \; = \; n_{2.}\, \pi$$

- $odds_1$ = odds of smoking for observed individuals (n_{12}/n_{11})

- π = probability of smoking for missing individuals under an assumed OR and given $odds_1$

Two Group Study

- Assume OR is same for control and treatment groups

- $n_{22c} = n_{2.c} \, \pi =$ calculated number of missing control individuals who are smoking

- $n_{22t} = n_{2.t} \, \pi =$ calculated number of missing treatment individuals who are smoking

- Add these calculated numbers to observed frequencies

- Perform χ^2 test for crosstab of group by smoking

- Repeat for various assumed levels of OR

Example

Gruder, Mermelstein *et al.*, (1993) JCCP

- 489 subjects measured across 4 timepoints (post-intervention, 6 months, 12 months, 24 months)

- Subjects were randomized to
  - control: self-help manuals
  - tx1: group meetings (discussion)
  - tx2: enhanced group meetings (social support)

- Half of the subjects randomized to tx1 or tx2 never showed up to any meetings following the phone call informing them of where the meetings would take place (no-shows)

Here, we focus on 24-month timepoint and combine no-shows with controls and tx1 with tx2 (as-treated analysis)

$$\begin{array}{c}
\texttt{Smoke} \\
\begin{array}{r|cc|c}
\texttt{Miss} & \text{no} & \text{yes} & \text{total} \\
\hline
\text{no} & 78 & 294 & 372 \\
\text{yes} & n_{21} & n_{22} & 117 \\
\hline
\text{total} & n_{.1} & n_{.2} & 489
\end{array}
\end{array}$$

- Observed individuals, odds of smoking $= 294/78 = 3.77$

- Amount of missing data is rather large ($\approx 24\%$)
    - $83/299 = 27.8\%$ in control group
    - $34/190 = 17.9\%$ in treatment group

- 2 x 2 crosstab of $\texttt{Group}$ by $\texttt{Smoke}$
    - Available data ($n = 372$): $X_1^2 = 1.86, p < .17$
    - Missing=smoking ($n = 489$): $X_1^2 = 3.80, p < .051$

Assuming $OR = 2$ (*i.e.*, odds of smoking are double in missing subjects than in non-missing subjects)

$$\pi = \frac{2 \times 294/78}{1 + (2 \times 294/78)} = .8829$$

$\Rightarrow$ assumed smoking rate $= 88.29\%$ for missing individuals

Number of missing smokers in the two groups can be calculated:

$n_{22c} = .8829 \times 83 = 73.2793$

$n_{22t} = .8829 \times 34 = 30.0180$

Adding these to the observed yields

<div style="text-align:center">Smoke</div>

Group	no	yes	total
tx	38 + 3.9820	118 + 30.0180	190
control	40 + 9.7207	176 + 73.2793	299
total	91.7027	397.2973	489

$X_1^2 = 2.28, p < .131$

Repeating this yields two-tailed p-values .10, .09, and .08 as OR is 3, 4, and 5, respectively

Group by smoking analyses under different missing data assumptions

| | Smoking frequencies (proportions) | | | |
	control	treatment	X^2	$p <$
Available data ($n = 372$)	176/216 (81.48)	118/156 (75.64)	1.87	.17
Missing=smoking ($n = 489$)	259/299 (86.62)	152/190 (80.00)	3.80	.051
OR = 1 ($n = 489$)	241.60/299 (80.80)	144.87/190 (76.25)	1.45	.23
OR = 2 ($n = 489$)	249.28/299 (83.37)	148.02/190 (77.90)	2.28	.13
OR = 5 ($n = 489$)	254.82/299 (85.22)	150.29/190 (79.10)	3.07	.08
OR = odds ratio for Missing and Smoking				

Previous Smoking Information

		Smoke0 = Non-Smoking					Smoke0 = Smoking	

		Smoke					Smoke	
Miss	no	yes	total		Miss	no	yes	total
no	42	71	$n_{11.} = 113$		no	36	223	$n_{21.} = 259$
yes	n_{121}	n_{122}	$n_{12.} = 37$		yes	n_{221}	n_{222}	$n_{22.} = 80$
total	$n_{1.1}$	$n_{1.2}$	$n_{1..} = 150$		total	$n_{2.1}$	$n_{2.2}$	$n_{2..} = 339$

LOCF posits $n_{121} = 37$ $(OR = -\infty)$ and $n_{222} = 80$ $(OR = \infty)$

Instead, π_i = probability of smoking for ith table $(i = 1, 2)$

$$\pi_i = \frac{OR_i \times odds_{i1}}{1 + (OR_i \times odds_{i1})}$$

- OR_i = assumed odds ratio for ith table

- $odds_{i1}$ = observed odds of smoking for ith table (1.69 & 6.19)

Continuing with $OR = 2$ assumption for missing and smoking, assumed both for $t0$ smokers and non-smokers, we get

$$\pi_1 = \frac{2 \times 1.69}{1 + (2 \times 1.69)} = .772 \quad \text{and} \quad \pi_2 = \frac{2 \times 6.19}{1 + (2 \times 6.19)} = .925$$

Thus, among those who are missing at the final timepoint:

- $t0$ smokers: very high assumed probability of smoking (.925)

- $t0$ non-smokers: lower smoking probability (.772)

Using these, and the number of missing control and treatment subjects in each table, yields smoking rates 83.42% and 77.45% for control and treatment groups $(X^2_1 = 2.70, p < .101)$

	Smoking frequencies (proportions)			
	control	treatment	X^2	$p <$
Available data ($n = 372$)	176/216 (81.48)	118/156 (75.64)	1.87	.17
Missing=smoking ($n = 489$)	259/299 (86.62)	152/190 (80.00)	3.80	.051
Stratified OR = 1 ($n = 489$)	242.34/299 (81.05)	143.78/190 (75.68)	2.02	.16
Stratified OR = 2 ($n = 489$)	249.42/299 (83.42)	147.16/190 (77.45)	2.70	.10
Stratified OR = 5 ($n = 489$)	254.76/299 (85.21)	149.82/190 (78.85)	3.28	.07

OR = odds ratio for `Missing` and `Smoking`
stratified = stratified by $t0$ smoking status

More Sources of Variation

- individual variation - subjects with same covariate values should have different probabilities of smoking $V(y \mid x)$

- sampling variation - sample proportions of smoking are estimates and are not known

- missing data variation - imputed values are more uncertain than observed values

$\Rightarrow$ Previous imputations ignored these, but better approach is to incorporate these real sources of variation into the imputation and data analysis

Imputation as Logistic Regression Model

$$\log\left[\frac{\#\texttt{Smoking}}{\#\texttt{Non Smoking}}\right] = [\beta_0 + \beta_1\texttt{Miss}]\,[1 - \texttt{Smoke0}] \ + \ [\beta_2 + \beta_3\texttt{Miss}]\,\texttt{Smoke0}$$

$\texttt{Miss} = 0$ or 1 for observed or missing individuals, respectively
$\texttt{Smoke0}$ ($t0$ smoking status) $= 0$ or 1 for non-smokers or smokers

Previous imputations have used observed data to yield β_0 and β_2, and assumed OR provides β_1 and β_3

Individual Variation

Consider the following latent variable representation of this logistic regression model for subject i ($i = 1, \ldots, n$):

$$y_i^* \ = \ [\beta_0 + \beta_1\texttt{Miss}_i][1 - \texttt{Smoke0}_i] \ + \ [\beta_2 + \beta_3\texttt{Miss}_i]\texttt{Smoke0}_i \ + \ \varepsilon_i$$

- y_i^* is a latent variable that is related to the observed binary smoking outcome y_i through the "threshold concept"

  $\Rightarrow$ if $y^* > \gamma$ then $y = 1$, otherwise if $y^* < \gamma$ then $y = 0$

- A logistic regression model for y implies that the distribution of ε_i is standard logistic with mean 0 and variance $\pi^2/3$

$\Rightarrow$ For missing subjects, generate random draw from standard logistic, add to $\boldsymbol{x}_i'\boldsymbol{\beta}$, obtain y_i^*, and determine if y is 0 or 1 (for this, $\gamma = 0$)

Sampling Variation

Based on results for logistic regression with a single binary predictor, for $t0$ non-smokers:

$$V(\hat{\beta}_0) = (n_{111} + n_{112})/n_{111}n_{112}$$
$$V(\hat{\beta}_1) = 1/n_{111} + 1/n_{112} + 1/n_{121} + 1\ n_{122}$$
$$C(\hat{\beta}_0, \hat{\beta}_1) = -(n_{111} + n_{112})/n_{111}n_{112}$$

- frequencies are as in table of `Smoke0` by `Miss` by `Smoke`
- n_{121} and n_{122} are obtained depending on the assumed level of the odds ratio for missing and smoking

$\Rightarrow$ Take random draw from bivariate normal with mean $\hat{\boldsymbol{\beta}}_{ns}$ (vector with $\hat{\beta}_0$ and $\hat{\beta}_1$) and variance-covariance $V(\hat{\boldsymbol{\beta}}_{ns})$

$\Rightarrow$ do same thing for $t0$ smokers $\hat{\boldsymbol{\beta}}_s$

Missing Data Uncertainty

- Data from $n = 372$ is known, however data from $n = 117$ is missing and imputed
- Analysis should reflect this extra uncertainty of $n = 117$
- Repeating the imputation many times, doing multiple imputation, allows us to assess and incorporate the variation attributable to imputation
- `SAS PROC MIANALYZE` can be used to combine results from multiply-imputed datasets

Group by smoking analyses under multiple imputation.
Averaged results based on 100 imputations. ($n = 489$)

| | Smoking frequencies (proportions) | | | |
	control	treatment	X^2	$p <$
Stratified OR $= 1$	242.09/299 (80.97)	143.82/190 (75.70)	1.60	.21
Stratified OR $= 2$	248.87/299 (83.23)	146.95/190 (77.34)	2.28	.13
Stratified OR $= 5$	254.20/299 (85.02)	149.55/190 (78.71)	2.91	.09

OR = odds ratio for `Missing` and `Smoking`
stratified = stratified by $t0$ smoking status

How extreme is $OR = 5$?
for $t0$ smokers and non-smokers, we get

$$\pi_1 = \frac{5 \times 1.69}{1 + (5 \times 1.69)} = .894 \quad \text{and} \quad \pi_2 = \frac{5 \times 6.19}{1 + (5 \times 6.19)} = .969$$

Thus, among those who are missing at the final timepoint:

- $t0$ smokers: of 80 missing, $80 \times .969 \approx 78$ smoking
  odds of smoking $= 78/2 = 39$

- $t0$ non-smokers: of 37 missing, $37 \times .894 \approx 33$ smoking
  odds of smoking $= 33/4 = 8.25$

$\Rightarrow$ very extreme

Multiple Imputation SAS syntax

```
DATA one; INFILE 'c:\smoke.dat';
INPUT id smk miss smk0 grp;
```

- **id** is the subject identifier

- **smk** is the smoking status at the final timepoint (0=abstinent, 1=smoking, .=missing)

- **miss** is the missing indicator (0=observed or 1=missing at the final timepoint)

- **smk0** is the $t0$ smoking status (0=abstinent, 1=smoking)

- **grp** is the grouping variable (0=control, 1=treatment)

Uppercase letters designate SAS syntax, lowercase letters designate user-defined entities

---

Observed cell frequencies in the crosstab of **smk0** by **miss** by **smk**

```
n111 = 42; /* number of abstainers - smk0=abstinent */
n112 = 71; /* number of smokers - smk0=abstinent */
n211 = 36; /* number of abstainers - smk0=smoking */
n212 = 223; /* number of smokers - smk0=smoking */
```

Missing cell frequencies are based on the observed frequencies, numbers of missing subjects, and the assumed odds ratio

```
orat = 2;
n12dot = 37; /* number of missing for smk0=abstinent */
p122 = (orat*(n112/n111))/(1 + (orat*(n112/n111)));
n122 = p122*n12dot;
n121 = (1 - p122)*n12dot;

n22dot = 80; /* number of missing for smk0=smoking */
p222 = (orat*(n212/n211))/(1 + (orat*(n212/n211)));
n222 = p222*n22dot;
n221 = (1 - p222)*n22dot;
```

Mean values of the logistic regression coefficients are obtained based on the above frequencies and the assumed level of the odds ratio for missing and smoking

```
beta0m = LOG(n112/n111);
beta1m = LOG(orat);
beta2m = LOG(n212/n211);
beta3m = LOG(orat);
seed = 974677743;
```

The variances associated with the regression coefficients are calculated (these formulas can be found in Agresti, 2002)

```
beta0v = (n111 + n112)/(n111*n112);
beta1v = 1/n111 + 1/n112 + 1/n121 + 1/n122;
beta2v = (n211 + n212)/(n211*n212);
beta3v = 1/n211 + 1/n212 + 1/n221 + 1/n222;
```

Imputation is now done using a logistic regression model

- It is important to perform this imputation multiple times

- The subsequent code does this 100 times

- To get a random draw from a standard logistic distribution, we use the fact that this distribution can be obtained as the natural logarithm of the ratio of two independent standard exponential distributions (see McCullagh and Nelder, 1989, page 20)

- Random draws from standard exponential distributions are obtained using the SAS function `RANEXP`

- Random draws from standard normal distributions are obtained using the SAS function `RANNOR`

- The Cholesky factorization, or matrix square root, of the variance-covariance matrix associated with the regression coefficients is used, and applied to the standard random normal deviates that are obtained using `RANNOR`

```
DATA sim; SET one;
ARRAY smks(100) smks1-smks100;
DO i = 1 TO 100;
  IF miss EQ 0 THEN smks(i) = smk;
  IF miss EQ 1 THEN DO;
   exp1 = RANEXP(seed); exp2 = RANEXP(seed); stdl = LOG(exp1/exp2);
   ran0 = RANNOR(seed); ran1 = RANNOR(seed);
   /* the next lines incorporate the covariance between beta0 and beta1
   (likewise for beta2 and beta3) using the Cholesky factorization */
   beta0 = beta0m + ran0*SQRT(beta0v);
   beta1 = beta1m - ran0*SQRT(beta0v) + ran1*SQRT(beta1v - beta0v);
   ran2 = RANNOR(seed); ran3 = RANNOR(seed);
   beta2 = beta2m + ran2*SQRT(beta2v);
   beta3 = beta3m - ran2*SQRT(beta2v) + ran3*SQRT(beta3v - beta2v);
   ystar = (1-smk0)*(beta0+beta1*miss) + smk0*(beta2+beta3*miss) + stdl;
   smks(i) = 0; IF ystar > 0 THEN smks(i) = 1;
  END;
END;
```

Here, a new dataset `sim` is created which will contain 100 smoking variables named `smks1` to `smks100`. A `DO` loop is used to create these 100 variables, and the `ARRAY` statement is used to specify a vector named `smks` containing the 100 smoking repetitions. These are set to the original variable `smk` for observed individuals, and imputed otherwise.

Analysis of Multiply-Imputed Data

To analyze the multiply-imputed data, we first have to adjust the data so that they are in the "long" format. Namely, in the file `sim`, which is in the "wide" format, each of the 100 smoking variables are associated with one case, whereas, for the analyses to be performed, each needs to be a separate case, with a variable indicating the imputation number. The code below does this translation, yielding a variable `smki` that is the smoking variable, and the variable `_imputation_` that is the imputation number. These variables, and `grp`, are saved in the dataset `unisim`.

```
DATA unisim (KEEP = id grp smki _imputation_); SET sim;
ARRAY smks(100) smks1-smks100;
DO _imputation_ = 1 TO 100;
  smki = smks(_imputation_);
  OUTPUT;
END;
```

The data are now sorted by `_imputation_`.

```
PROC SORT; BY _imputation_ ;
```

The logistic regression analysis is performed, stratified by `_imputation_` (*i.e.*, performed 100 times), and the results from each analysis are saved in the dataset `outlreg`.

```
PROC LOGISTIC DESCENDING NOPRINT OUTEST=outlreg COVOUT;
MODEL smki = grp / LINK = LOGIT;
BY _imputation_ ;
```

The results corresponding to the regression coefficients (*i.e.*, for `intercept` and `grp`) from the 100 logistic regression analyses are combined using `PROC MIANALYZE`.

```
PROC MIANALYZE DATA=outlreg;
VAR intercept grp;
```

`PROC MIANALYZE` prints out the results of the multiple imputation process for the two logistic regression parameters `intercept` and `grp`.

Combining Results of Multiply-Imputed Datasets

For a given dataset, we can test whether the probability of smoking is the same in the control and treatment groups using

$$z = \frac{\hat{p}_c - \hat{p}_t}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_c} + \frac{1}{n_t}\right]}}$$

$\hat{p}_c$ is the smoking proportion in the control group
$\hat{p}_t$ is the smoking proportion in the treatment group
$\hat{p}$ is the smoking proportion in the entire sample
n_c and n_t are the control and treatment group sample sizes

$\Rightarrow$ Squaring this z-statistic equals the Pearson X^2 statistic for assessing the independence of group and smoking

Denote the numerator and denominator in z for imputation j ($j = 1, \ldots, m$) as $\hat{Q}^{(j)}$ and $\sqrt{U^{(j)}}$, and calculate the averages:

$$\bar{Q} = \frac{1}{m} \sum_{j=1}^{m} \hat{Q}^{(j)}$$

$$\bar{U} = \frac{1}{m} \sum_{j=1}^{m} U^{(j)}$$

The latter average represents the within-imputation variance

The between-imputation variance is calculated as

$$B = \frac{1}{m-1} \sum_{j=1}^{m} \left[\hat{Q}^{(j)} - \bar{Q}\right]^2$$

and the total variance is a modified sum of these two components

$$T = \bar{U} + (1 + 1/m)B$$

To test the null hypothesis that group and smoking are independent use $\bar{Q}/\sqrt{T} \sim t$ on ν df, where

$$\nu = (m-1)\left[1 + \frac{\bar{U}}{(1 + 1/m)B}\right]^2$$

If ν is large (*e.g.*, 100 or so), then

- $\bar{Q}/\sqrt{T} \sim$ standard normal
- $(\bar{Q}/\sqrt{T})^2 \sim \chi_1^2$