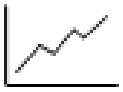
2024 ■ Vol. 20 ■ no. 3

Teaching Quantitative Methods Vignettes

Dealing with missing data in covariates: The missing indicator method

Shahab Jolani [a,b] and Pawel Weinstein [a]

[a]Department of Methodology and Statistics, CAPHRI, Maastricht University, The Netherlands
[b]School of Health Professions Education, FHML, Maastricht University, The Netherlands



Abstract ■ This vignette presents the missing indicator method for handling missing data in covariates. The method unfolds through two activities, guiding students in the practical implementation of the method and comparable alternatives using statistical software. We further evaluate these activities and discuss conditions under which the missing indicator method yields valid results. We conclude that the missing indicator method can be safely used in experimental studies characterized by randomization protocols.

Keywords ■ Imputation, Missing data, Observational studies, Randomization. **Tools** ■ R, SPSS.

✉ s.jolani@maastrichtuniversity.nl

10.20982/tqmp.20.3.p032

Acting Editor ■ Sébastien Béland (Université de Montréal)

Reviewers ■ One anonymous reviewer.

Concept to be presented

The term "*missing observation*" usually refers to a planned observation or measurement, that for a particular reason has not been measured. For instance, the baseline intensity of depression might not have been measured for some patients in a study that investigates the effects of a treatment on depression, while controlling for symptom severity of depression.

Missing data are a common and known problem in any research domain. For many researchers, however, it remains unclear how to correctly compensate for missing data and analyze the data with as little bias and as much efficiency as possible. Since missing data can occur for a multitude of reasons and at different observations (e.g., missing data in the independent vs. dependent variable), different methods of analyzing missing data have been developed and discussed over time (see, e.g., Little & Rubin, 2019; Schafer & Graham, 2002). Among the most common methods of analyzing missing data in covariates, we focus on the *missing indicator method* in the present vignette.

The missing indicator method is an approach that aims to account for the patterns of missing data in the independent variables (i.e., covariates) while the association between the independent variable of interest (e.g., treatment) and the dependent variable is sought (Groenwold et al., 2012). The missing indicator method utilizes dummy variables in the analysis model to visualize whether a particular case includes missing data or not. To conceptualize the methodology, let us consider the following hypothetical dataset which consists of 5 subjects. Suppose we want to estimate the effect of a treatment on the intensity of depression, controlling for baseline measurements. Here, the baseline intensity of depression is not measured for some individuals while the other variables are fully observed (see Table 1). The analysis model of interest can thus be defined by

$$\text{Post Score} = \beta_0 + \beta_1 \text{Treatment} + \beta_2 \text{Baseline score} + \varepsilon, \quad (1)$$

where β_0, β_1, and β_2 are the regression weights, and ε represents the random error. We throughout assume that model (1) is correct and β_1 is the coefficient of interest, which implies the ordinary least squares estimate of β_1 is unbiased.

The missing indicator method replaces all missing observations of the baseline intensity of depression with a fixed value (e.g., zero) and adds a dummy variable M to the dataset (see Table 2). The analysis model is then extended to

$$\text{Post Score} = \alpha_0 + \alpha_1 \text{Treatment} + \alpha_2 W + \alpha_3 M + \varepsilon \quad (2)$$

Table 1 ■ Example dataset of 5 participants with missing baseline scores

Participant ID	Treatment	Baseline Score	Post Score
1	1	85	56
2	1		74
3	2	78	27
4	2		36
5	2	92	80

Table 2 ■ Example dataset of 5 participants with missing baseline scores after creating the missing indicator variable

Student ID	Treatment	Baseline Score	Post Score	Missing Score
1	1	85	56	0
2	1	0	74	1
3	2	78	27	0
4	2	0	36	1
5	2	92	80	0

where α's are the regression weights, ε is the random error, and W = baseline score if $M = 0$ and $W = 0$ if $M = 1$. Models (1) and (2) are not identical so that the estimate of α_1 is not necessarily a correct estimate (i.e., unbiased) for β_1. To see this, let us rewrite model (2) per missingness pattern:

$$\text{Pattern 1 (where } M = 0) \text{ :Post Score} = \alpha_0 + \alpha_1 \text{Treatment} + \alpha_2 \text{Baseline score} + \varepsilon \tag{3}$$

$$\text{Pattern 2 (where } M = 1) \text{ :Post Score} = \alpha_0 + \alpha_1 \text{Treatment} + \alpha_3 + \varepsilon \tag{4}$$

The former equation is identical to model (1) and so the treatment effect is correctly estimated as long as the missingness in the baseline score is unrelated to the post score. However, the treatment effect is incorrectly estimated in pattern 2 because this model does not correct for the baseline score. This is true particularly in observational studies and is not related to any cause of missing data. Hence, the missing indicator method would very likely produce biased results in general (Donders et al., 2006).

There are, however, situations where the missing indicator method produces valid results. In randomized controlled trials (RCTs), both sub-models deliver correct estimates of the treatment effect, and so the missing indicator method produces an unbiased estimate of the treatment effect (White & Thompson, 2005). Specifically, the treatment effect is correctly estimated in pattern 1 because it is very highly unlikely that the post score (which happens in the future) causes missingness in the baseline score (which has already happened in the past). In pattern 2, the treatment effect is also correctly estimated because randomization implies that the baseline covariates (or their missingness) are independent of the treatment. It should be emphasized that the missing indicator method may result in biased estimates, even in randomized studies when the analysis model (1) is not of interest (e.g., the analysis model includes the interaction between the treatment and a covariate), or when the baseline score becomes a confounder within the second pattern. This can happen if the missingness in the baseline score depends on the treatment allocation and the baseline score itself or their interaction (see, for details, Kayembe et al., 2020). In the intensity of depression example, for instance, there might be more missing baseline scores among those with a high intensity score in the control group. This may happen when randomization is not blind to the participants. In this case, those patients (who are assigned to the control group and have a high intensity score) might be less motivated to provide their baseline score because they don't expect any potential benefit from the treatment as they are allocated to the non-treated (i.e., control) group.

Table 3 ■ Treatment effect estimates, standard errors, 95% confidence intervals, and sample sizes related to the missing indicator method (MIM), complete case analysis (CCA), and unadjusted analysis (UA) with missing data in the baseline negative affect score

Methods	β	Std. Error	lo.95	hi.95	N
UA 0.312	1.241	-2.153	2.778	91	
CCA 0.293	0.907	-1.510	2.097	88	
MIM 0.552	0.928	-1.292	0.848	91	

Table 4 ■ Treatment effect estimates, standard errors, 95% confidence intervals, and sample sizes related to the missing indicator method (MIM), complete case analysis (CCA), and unadjusted analysis (UA) with missing data in two independent variables.

Methods	β	Std. Error	lo.95	hi.95	N
UA	0.312	1.241	-2.153	2.778	91
CCA	0.287	0.939	-1.582	2.156	82
MIM	0.613	0.913	-1.202	2.429	91

Activities

The following activities aim at exemplifying how the missing indicator method should be implemented in practice. Further, we compare it with two other alternatives: (1) unadjusted analysis, where the analysis model is not adjusted for any baseline covariates, and (2) complete case analysis, where complete cases are used in the analysis only. This will be done by using the data from Kole-Snijders et al. (1999). In the original study, 149 participants with chronic-lower-back-pain (CLBP) were randomly allocated to three groups: Operant behavioral treatment with coping skills training (OPCO), operant behavioral treatment with group-discussion (OPDI) and a waiting-list control group (WLC). Various outcome measures were collected before and after randomization. Here, we focus on the negative affect score as the outcome of interest. To simplify the presentation, we exclude the WLC group and remove the missing post measurements of the negative affect score. This resulted in a dataset of N=91 participants.

Activity 1

The first activity deals with missing data in a continuous baseline variable. Students need to create the variable M, indicating whether a participant has a missing value in the baseline variable. Subsequently, students will compute a new variable (W), which equals the value of the baseline variable (i.e., the negative affect score at baseline - denoted by negpre) if that variable is observed. Otherwise, it takes the value of zero (i.e., $W = 0$). This results in the regression equation, which is comparable to equation 2.

$$\text{negpost} = \beta_0 + \beta_1 \text{group} + \beta_3 M + \beta_4 W + \varepsilon.$$

Students will now run the analysis based on three different methods: (a) Unadjusted analysis (UA) (b) Complete case analysis (CCA) (c) Missing indicator method (MIM) CCA can be completed by excluding cases with missing data in the covariate (or independent variable). UA is an analysis with no adjustment, which means performing a regression analysis with the treatment indicator as the only independent variable.

Activity 2

The second activity aims to demonstrate how the missing indicator method can be extended to two independent variables with missing data. Here, a binary incomplete variable (social security for working disabled (SSWD) coded as 1 = yes and 0 = no) is added to the analysis model of Activity 1. First, students should look at the previous regression equation and come up with ideas. What could be changed if a second (incomplete) independent variable is added to the model? What does the model look like now?

Similar to Activity 1, students will need to create an indicator variable for each independent variable with missing data. Instruct the students to create a second indicator variable, which shows whether SSWD includes missing data or not; call it the indicator variable M2, where M2=0 if SSWD is observed and M2=1 if SSWD is missing. Next, a new variable W2 is computed, which equals the observed value of SSWD if it is not missing but zero (or any other fixed value) if the value of SSWD is missing. This will result in the following regression equation:

$$\text{negpost} = \beta_0 + \beta_1 \text{group} + \beta_2 M_1 + \beta_3 W_1 + \beta_3 M_2 + \beta_4 W_2 + \varepsilon.$$

As in activity 1, students will now run the analysis based on MIM, CCA and UA.

Strategy to assess the activities

Tables 3 and 4 report the results of Activities 1 and 2, respectively. As expected, all methods leaded to the same conclusion. This is not surprising because these methods are all valid in randomized studies, provided proper randomization. Of course, the point estimates are not identical because each method uses a different statistical model. Nevertheless, the same conclusion can be drawn from these methods concerning the treatment effect.

One may then ask what is the added value of using the missing indicator method at all if there exist simpler alternatives? It can be argued that adjustment for covariates in randomized studies is advantageous because it leads to greater precision and increased statistical power (Lingsma et al., 2010). Moreover, the MIM approach is expected to provide more accurate estimates of effect sizes because it increases both the number of covariates and the sample size. In contrast, UA and CCA use, respectively, fewer variables and less data. Therefore, it is always desirable to adjust for relevant covariates in randomized studies.

Conclusion

The goal of this vignette was to present an alternative method of dealing with missing data in covariates. The exercises are used to elaborate that the missing indicator method leads to equally valid results in randomized studies. It is important to emphasize two points:

(1) Contrary to conventional texts on missing data, the reasons for missingness (known as the missing data mechanisms, see, among others, Tan and Jolani, 2022, Chap. 1) do not play a crucial role in the validity of the missing indicator method in randomized studies. Due to proper randomization, the analysis model within each pattern (i.e., equations 3 and 4) is correct and delivers an unbiased estimate of the treatment effect irrespective of any missingness mechanisms.

(2) The missing indicator method is generally invalid in non-randomized studies such as observational or cohort studies. The main reason is that both equations 3 and 4 lead to biased estimates of the coefficient of interest (e.g., β_1). In equation 3, the missingness mechanism can play a role if the missingness in the independent variable relates to the dependent variable (while this was not the case in randomized studies because it is implausible that the outcome causes the missing baseline values). Moreover, model (4) delivers biased estimates because this model is not adjusted for covariates and, hence it is incorrect in general.

Moreover, it should be noted that the missing indicator method (and mean imputation) are not recommended for handling missing outcome data due to their tendency to underestimate SEs. They are, nevertheless, more appropriate for managing incomplete covariates in RCTs (see, among others, Kayembe et al., 2022, 2023).

Several suggestions can be taken into consideration for further study. Students have already experienced inclusion of multiple covariates in a linear regression analysis. To explore further, similar analyses can be repeated when the analysis model is a generalized linear model (e.g., a logistic regression model). Moreover, students could inspect how the results change, depending on the rate of missing data.

References

Donders, A. R., van der Heijden, G. J., Stijnen, T., et al. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*, 1087–1091.

Groenwold, R. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., & Moons, K. G. (2012). Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. *Canadian Medical Association Journal*, *184*(11), 1265–1269.

Kayembe, M. T., Jolani, S., Tan, F. E., & van Breukelen, G. J. (2020). Imputation of missing covariate in randomized controlled trials with a continuous outcome: Scoping review and new results. *Pharmaceutical Statistics*, *19*(6), 840–860.

Kayembe, M. T., Jolani, S., Tan, F. E. S., & van Breukelen, G. J. P. (2022). Imputation of missing covariates in randomized con- trolled trials with continuous outcomes: Simple, unbiased and efficient methods. *Journal of Biopharmaceutical Statistics*, *32*(5), 717–739.

Kayembe, M. T., Tan, F. E. S., van Breukelen, G. J. P., & Jolani, S. (2023). Dealing with missing observations in the outcome and covariates in randomized controlled trials. *Journal of Statistical Computation and Simulation*, *94*(7), 1513–1542.

Kole-Snijders, A., Vlaeyen, J. W., Goossens, M. E., Rutten-van Mölken, M. P., Heuts, P. H., van Breukelen, G., & van Eek, H. (1999). Chronic low-back pain: What does cognitive coping skills training add to operant behavioral treatment? Results of a randomized clinical trial. *Journal of Consulting and Clinical Psychology*, *67*, 931–939.

Lingsma, H., Roozenbeek, B., & Steyerberg, E. (2010). Covariate adjustment increases statistical power in randomized controlled trials. *Journal of Clinical Epidemiology*, *63*, 1391–1393.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd). John Wiley & Sons.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Tan, F. E. S., & Jolani, S. (2022). *Applied linear regression for longitudinal data with an emphasis on missing observations*. Chapman & Hall/CRC Press.

White, I. R., & Thompson, S. G. (2005). Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine*, *24*(7), 993–1007.

Appendix: Source codes

Activity 1: SPSS Syntax

```
* Create the indicator variable M1.
RECODE negpre (SYSMIS=1) (ELSE=0) INTO M1.
* Recode the covariate 'negpre' into W1.
RECODE negpre (SYSMIS=0) (ELSE=Copy) INTO W1.
EXECUTE.
* Regression analysis using MIM.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT negpost
  /METHOD=ENTER group M1 W1.
* Regression analysis using CCA.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT negpost
  /METHOD=ENTER group negpre.
* Regression analysis using UA.
REGRESSION
  /MISSING PAIRWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT negpost
  /METHOD=ENTER group.
```

Activity 1: R code

```
# "Haven" is a R package that is necessary to import and read SPSS files.
# Packages need to be first installed , and then loaded in every time R is restarted.
install.packages("haven")
library(haven)
# This command reads the SPSS file 'data.sav' into R.
# In the quote marks the full file path must be dedicated for R to be able to read the file.
data <- read_sav("data.sav")
# Create the indicator variable M1
data$M1 <- ifelse(is.na(data$negpre), 1, 0)
# Recode the covariate 'negpre' into W1
data$W1 <- ifelse(is.na(data$negpre), 0, data$negpre)
# Regression analysis using MIM
model_mim <- lm(negpost ~ group + M1 + W1, data = data)
```

```r
summary(model_mim)
# Regression analysis using CCA
# Note that the lm function automatically uses complete cases
model_cca <- lm(negpost ~ group + negpre, data = data)
summary(model_cca)
# Regression analysis using UA
model_ua <- lm(negpost ~ group, data = data)
summary(model_ua)
```

Activity 2: SPSS Syntax

```
* Create the indicator variable M2.
RECODE sswd (SYSMIS=1) (ELSE=0) INTO M2.
EXECUTE.
* Recode the covariate 'sswd' into W2.
RECODE sswd (SYSMIS=0) (ELSE=copy) INTO W2.
EXECUTE.
* Regression analysis using MIM.
REGRESSION
  /MISSING PAIRWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT negpost
  /METHOD=ENTER group M1 W1 M2 W2.
* Regression analysis using CCA.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT negpost
  /METHOD=ENTER group sswd negpre.
* Regression analysis using UA.
REGRESSION
  /MISSING PAIRWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT negpost
  /METHOD=ENTER group.
```

Activity 2: R code

```r
# Create the indicator variable M2
data$M2 <- ifelse(is.na(data$sswd), 1, 0)
View(data)
# Recode the covariate 'sswd' into W2
data$W2 <- ifelse(is.na(data$sswd), 0, data$sswd)
# Regression analysis using MIM
model_mim2 <- lm(negpost ~ group + M1 + W1 + M2 + W2, data = data)
summary(model_mim2)
```

```
# Regression analysis using CCA
# Note that the lm function automatically uses complete cases
model_cca2 <- lm(negpost ~ group + negpre + sswd, data = data)
summary(model_cca2)
# Regression analysis using UA
model_ua2 <- lm(negpost ~ group, data = data)
summary(model_ua2)
```

Citation

Jolani, S., & Weinstein, P. (2024). Dealing with missing data in covariates: The missing indicator method. *The Quantitative Methods for Psychology*, *20*(3), v32–v38. doi: 10.20982/tqmp.20.3.p032.

Received: 25/01/2024 ~ Accepted: 10/06/2024

Extended activity metadata

Concept illustrated	Missing data in covariates	*Type of activity*	in-class with instructor
Prerequisite	Simple linear regression	*Types of data*	Experimental data, randomized trials
Co-requisite	Multiple linear regression	*Computation by*	R and SPSS
Suitable class size	medium (30 postgraduate students)	*Duration*	1 hour