Missingness Indicators in Linear Regression

Sara E. Burke

Created 2019-10-05 · Last updated 2024-10-01

Not all variables in a set of research data necessarily have the same number of observations. Some variables have missing values, either because those values were never recorded or because those values do not exist. This document is not intended to review analysis strategies for missing data. Many such reviews already exist. Here, I hope to explain one simple but often misunderstood way of using one variable to indicate the missingness of another variable in a regression model.

When one or more variables have a small amount of missing data, many studies involving convenience samples will simply discard incomplete responses, focusing all statistical analysis on the cases with complete data. This is the default behavior of most statistical software.

In cases where the entire sample is important to the claim being tested, such as cases where researchers are using a systematic sampling procedure as part of an argument for generalizability, omitting cases with missing data is likely to contribute to biased results. For example, people who decline to respond to certain survey questions may differ meaningfully from people who answer those questions, so excluding them from the sample would be similar to excluding people in a particular age bracket or with a particular personality characteristic. In these cases, researchers employ sophisticated imputation techniques informed by careful analysis of the pattern of missing data itself. For example, multiple imputation uses information from available variables to generate several plausible guesses for each missing value, computes several estimates for each parameter of interest, and pools the results to incorporate the uncertainty added by the imputation procedure itself. As with all imputation procedures, multiple imputation is only appropriate if there are unobserved true values (e.g., a participant with a missing value for "age" still necessarily has some true age), not if the values genuinely do not exist (e.g., a participant may have a missing value for "age of first child" due to not having any children). In situations where there are unobserved values in a sample and the statistical argument rests on the sampling procedure, thorough missing data analysis is always necessary and multiple imputation is frequently one of the best analysis options. This document is not about multiple imputation, but if you are faced with a situation that might warrant such a technique, I encourage you to read about it. I would not want to be accused of recommending the use of a missingness indicator in place of multiple imputation.

That said, it is useful for students to know what including a missingness indicator in a linear regression model accomplishes, both because it has some reasonable use cases and because it helps illustrate basic principles of regression.

First, imagine a researcher has a categorical variable, favorite color, with some missing values because participants chose to skip the corresponding question on the survey. She might simply treat "did not respond" as an additional category. If "blue" is her reference category, her regression model might include indicator variables for "green," "purple," "red," "other" (a catchall category for infrequently chosen colors), and "did not respond." This strategy keeps more participants in the analysis than simply omitting cases with missing values, and it makes conceptual sense if she believes that declining to choose a color is a distinct kind of response, informative in its own right.

Now, imagine instead that the researcher has a numeric variable, liking for the color blue on a scale from 0 to 100, with some missing values. Treating missing values as a separate category is also possible in this case, using a strategy that is fundamentally identical to the strategy for the categorical variable. It can be accomplished by replacing the missing values in the blue-liking variable with any constant numeric value, and then including a separate indicator variable distinguishing cases with a missing value from those with a legitimate liking response. The resulting slope for blue-liking will be the same as it would have been if the researcher simply excluded all of the missing values from the model; the standard error will be somewhat smaller due to the larger sample size. If there were other color-liking variables in the same model, each with a different set of missing values, this strategy would permit the whole sample to be included, but would not necessarily permit the researcher to think of the results as fully adjusting for all of the other color-liking variables at once.

Consider the following simple example. There is complete data for the response variable y, but some missing data for the predictor variable x. We will compute a new variable x' that is identical to x but with missing values replaced by some arbitrary constant c, and a new variable m that is 1 when x is missing and 0 otherwise.

$$\begin{array}{ll} \text{if } x \text{ is missing:} & x' = c \\ & m = 1 \\ \\ \text{if } x \text{ is not missing:} & x' = x \\ & m = 0 \end{array}$$

Then we will fit a linear regression model with x' and m as the predictors. In the following equation, b_0, b_1, and b_2 are the intercept and two slopes, respectively, and e is the vector of residuals. As usual, b_0, b_1, and b_2 are computed so as to minimize $\sum e^2$.

$$y = b_0 + b_1 x' + b_2 m + e$$

There are only two possible values for m, 0 and 1. If m is 0, the $b_2 m$ term is 0 and x' is the same as x. If m is 1, x' is always the constant c. Therefore, the regression equation can be rewritten as follows.

$$y = \begin{cases} b_0 + b_1 x + e, & \text{if } m = 0 \\ b_0 + b_1 c + b_2 + e, & \text{if } m = 1 \end{cases}$$

In the case where $m = 1$, all of the terms except the residual are constant. Whatever b_0 and b_1 turn out to be, the residuals will be minimized by identifying the value for b_2 that sets $b_0 + b_1c + b_2$ to the mean value of y for cases with $m = 1$. In other words, b_2 is the difference between the mean of y when x is missing and the fitted value of y when x is the arbitrary constant c. We might purposefully assign a value of c that makes b_2 interesting, but the choice of c does not affect any other part of the results. (For example, if we set c to the mean of the observed x values, then b_2 indicates how far the observed values of y for cases with missing x deviate from what would be expected if those cases had that same mean x.)

In the case where $m = 0$, b_0 and b_1 define the simple linear regression line for the relationship between x and y among cases with no missing data. Again, no matter what values of b_0 and b_1 would minimize the sum of squared residuals just for cases where $m = 0$, those same values will also minimize the overall sum of squared residuals, because when $m = 1$, given any value of the constant $b_0 + b_1c$, some b_2 can be found that minimizes the sum of squared residuals.

In other words, we could fit a linear regression model using only cases with complete data for x and y.

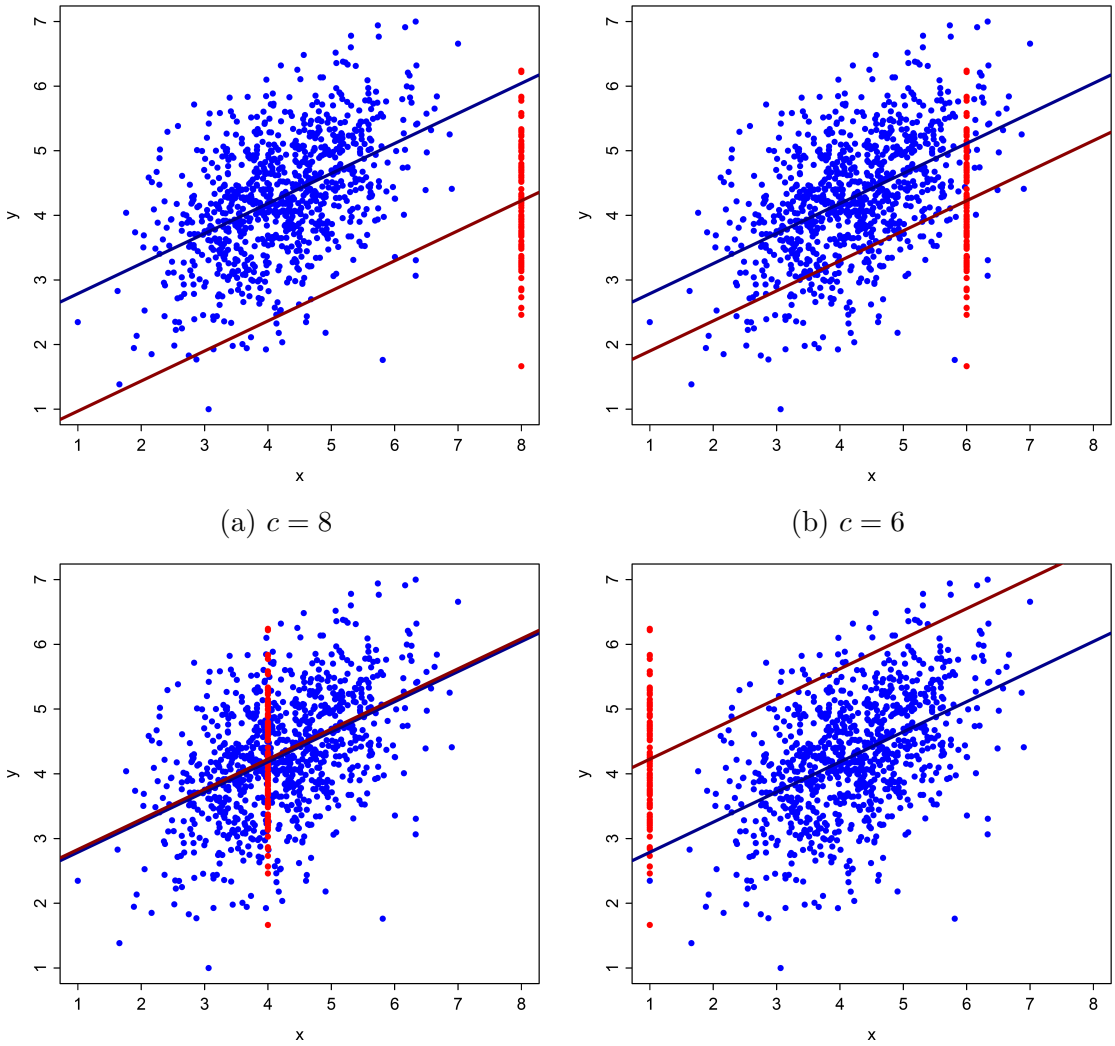
$$y = b_0 + b_1x + e$$

If, instead, we fit a linear regression model using a missingness indicator, we get the same b_0 and b_1, plus an estimate b_2 that describes the mean observed y in cases with missing x (relative to some comparison value of our choosing).

$$y = b_0 + b_1x' + b_2m + e$$

To help demonstrate that the fitted values and the b_0 and b_1 estimates do not depend on the constant c, I have generated 1000 values for y and 900 for x, then fit the above model. Scatterplots and regression lines are shown in Figure 1 on the next page. Regardless of what value c is chosen to replace the missing values in x, the defining features of the model are identical.

In fact, the inclusion of the missingness indicator does nothing more than create a separate category for the cases with missing values.



(a) $c = 8$ (b) $c = 6$

(c) $c = 4$ (d) $c = 1$

Figure 1: Example regression lines representing $\hat{y} = b_0 + b_1 x' + b_2 m$ with all missing values of x replaced by arbitrary constants c. The blue points and line reflect non-missing values of x and the red points and line reflect missing values of x. The blue line is identical regardless of the value chosen for c. The height of the red line indicates the difference between the average y value of the red points and the value of the blue line at the point $x = c$.

Sometimes, a researcher may wish to use a missingness indicator to include a covariate with missing values in a regression model without reducing the available sample. Formal examinations of the missingness indicator approach note that it does not successfully mimic a full adjustment for the covariate in question (Donders, van der Heijden, Stijnen, & Moons, 2006; Groenwold et al., 2012; Jones, 1996; Knol et al., 2010). That is, if the goal of including a covariate is to adjust statistically for the theoretical association between the "true" values of the covariate and those of the predictor(s) of interest, then using a missingness indicator fails because it does not even attempt to imitate the true values of the covariate where they were not observed. The estimate of the main slope of interest will be somewhere between what it would be with a complete case analysis adjusting for the observed values of the covariate and what it would be with a regression run only on the cases with missing values for the covariate (ignoring the covariate). If adjusting for the true values of the covariate is critical to the argument, and there are more than a few missing values, then merely including a missingness indicator will likely undermine rather than strengthen the argument.

In many studies involving convenience samples, there is no chance of achieving results that are truly representative of any specific population. Researchers sometimes use complete case analysis in these situations, given the fact that the (convenience) sample of participants with complete data may be no more or less representative of any specific population than the slightly larger (convenience) sample of participants with any data. Sophisticated imputation strategies could achieve less biased estimates of the population who would have been willing to start the survey in question, regardless of their willingness to finish all of it, but it is unclear that this population is especially informative. In such situations, if the researcher is comfortable with only partially adjusting for a particular covariate, and/or if the decision not to respond to a particular item is likely to be informative in itself, then including a missingness indicator may be a viable strategy.

It bears emphasis that including a missingness indicator for a numeric variable is identical to treating missingness as a separate category for a (nominal) categorical variable. Many researchers seem fully comfortable with the latter, but nervous about the former. We should be equally nervous about both. There are strong reasons that the missingness indicator approach should be avoided in certain cases, such as when fully adjusting for a covariate is essential to the argument, and all those same reasons apply to the "did not respond" category approach for categorical variables. Multiple imputation is often an option to mitigate these problems.

If you have a careful sampling procedure, and part of your statistical argument rests on your careful sampling procedure, then you necessarily must also be careful about missing data. You will want to scrutinize the pattern of missingness and its consequences for your results, and you may want to use some form of multiple imputation for your primary analysis.

In some situations, though, complete case analysis is sufficient. In many of those same cases, a missingness indicator approach may also serve well. Including a missingness indicator may be an especially useful strategy in situations where there is no "true" value to impute, such

as when some childless participants are included in a model with a covariate for "age of first child." Regardless of whether they actually have occasion to use it, I expect that many students will find that understanding the implications of including a missingness indicator helps foster a clearer understanding of regression in general.

References

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*(10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

Groenwold, R. H. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., & Moons, K. G. M. (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. https://doi.org/10.1503/cmaj.110977

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association, 91*(433), 222–230. https://doi.org/10.1080/01621459.1996.10476680

Knol, M. J., Janssen, K. J. M., Donders, A. R. T., Egberts, A. C. G., Heerdink, E. R., Grobbee, D. E., ... Geerlings, M. I. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: An empirical example. *Journal of Clinical Epidemiology, 63*(7), 728–736. https://doi.org/10.1016/j.jclinepi.2009.08.028