Missing Data

Brenton Kenkel — PSCI 8357
April 21, 2016

Missing data is a universal problem in the social sciences. Even randomized lab experiments, which bypass many of the inferential problems we've focused on in this course, may suffer from missing data. Missing data can itself pose a threat to inference. We will conclude the semester by talking about these problems and how best to deal with them.

Varieties of Missingness

What you should do with missing data depends on how the data went missing. These definitions are drawn from Rubin (1976) and King et al. (2001).

Missing Completely at Random (MCAR). The probability of missingness does not depend on the values of the data (observed or unobserved).

Under MCAR, the complete observations are in effect a random subsample of the original sample. Any inferential procedure that would have been valid with the full data will therefore still be valid on the subsample of complete observations. The standard errors will of course be larger because there is less data. We call this procedure—running what we wanted to run on the full data on the subsample of observations with no missing data—*listwise deletion* or *complete-case analysis*.

MCAR is not a plausible assumption for most missing data in political science. In surveys, data are missing because respondents lack knowledge or deliberately wish to conceal their answers—these do not occur completely at random. Administrative data are missing when governmental entities lack the capacity or willingness to collect them—again, not a chance process.

The MCAR assumption is testable: you can create a dummy variable for missingness, regress it on the data, and test the composite hypothesis that every variable has zero effect.

Missing at Random (MAR). The probability of missingness does not depend on the missing values themselves. It is a function solely of the observed values. In other words, if we could know the true value of the missing data, we would not learn anything new about its likelihood of going missing.

For example, suppose high-income people are both more likely to be Republicans and to conceal their party affiliation. As long as income is observed, MAR holds. Our data might look like the following table.

Table 1. Illustration of MAR data.

income	party
low	D
low	R
low	D
high	**?**
high	R
high	**?**

But if high-income people were also more likely to conceal their income, then the data would no longer be MAR.

The MAR assumption differs from MCAR in two important ways. The first is that, with MAR data, we cannot *in general* obtain valid inferences by listwise deletion. In the special case of a linear model, though, OLS following listwise deletion produces valid estimates if data are MAR (Little 1992). It is critical that the model be properly specified; otherwise missing data may pose a serious inferential threat (Winship and Radbill 1994).

The second difference from MCAR is that the MAR assumption is not testable. It is a condition on unobservables, namely the values of the missing data. We would need these values to test the MAR assumption, but then again if we had them we wouldn't have to worry about any of this in the first place.

MAR is a necessary condition for missingness to be *ignorable*. (The terminological connection to strong ignorability in causal inference is not a coincidence.) Ignorability means that we can draw valid inferences without knowing the process that generates the missingness. That doesn't mean we can pretend missingness doesn't exist—we have already seen that listwise deletion might

produce invalid inferences—but that we need not model the exact process from which it emerges.[1]

Nonignorable Missingness (NI). The probability of missingness depends on the missing value itself.

NI is the worst-case scenario—though don't take that to mean it's uncommon. To draw valid inferences with nonignorable missing data, we must either (1) know the exact form of the missingness-generating process so that we can estimate a joint model of missingness and the outcome of interest or (2) settle for robust but highly imprecise estimates. Door #1 requires techniques beyond the scope of Stat II. We will briefly venture into door #2 at the end of class.

The figure illustrates hypothetical data with two correlated covariates, X_1 and X_2, in which X_2 is sometimes missing. We consider each of the three assumptions about the missingness process:

1. MCAR: The probability X_2 is missing is constant across observations.
2. MAR: The probability X_2 is missing is a function of X_1.
3. NI: The probability X_2 is missing is a function of X_2.

Multiple Imputation

Even the "good" kinds of missing data, MCAR and MAR, create problems for us. The most pertinent is inefficiency—inflated standard errors. Listwise deletion eliminates every observation for which any covariate is missing. We don't want to throw away $p-1$ perfectly good pieces of information just because one is missing.

The most common way to deal with missing data in political science is *multiple imputation* (Schafer 1999, King et al. (2001)). If the data are MAR (or MCAR), multiple imputation is a path to valid inference. The concept behind multiple imputation is that, under MAR, we can predict the values of the missing observations from the other information in the dataset—*imputation*. These

[1] The other necessary condition for ignorability is a technical one: the parameters of the process that generates missingness must be distinct from the parameters of the process that generates the outcome of interest.

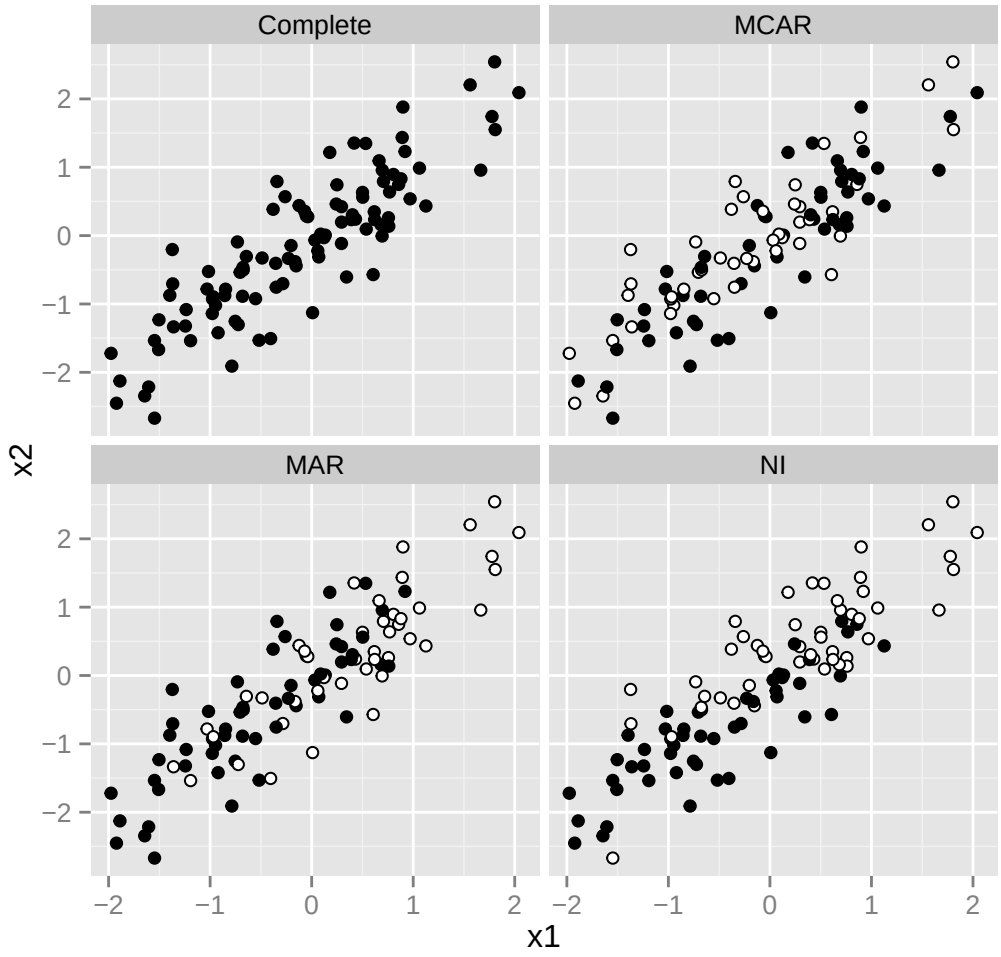

Figure 1. Illustration of data that are complete, missing completely at random (MCAR), missing at random (MAR), and nonignorably missing (NI).

predictions won't be perfect, though.[2] To represent our uncertainty, we will make *multiple* draws from the distribution of predictions.

Multiple imputation consists of the following steps.

1. Use the observed data to construct a model of the joint relationship among the $p+1$ variables $Y, X_1, \ldots, X_p$.
2. Form M "completed" datasets by drawing from the conditional distribution of the missing values given the observed data.
3. Run the statistical procedure you would have used if the data were fully observed on each of the M imputed datasets.
4. Combine the estimates and standard errors from each imputation run to come up with an overall estimate and standard error.

Parts 1 and 2 are the tricky ones, but there's software to take care of that for you.

Let us consider a toy example. Suppose we have the following data on Democratic primary vote choice for ten voters. We want to model vote choice as a function of the voter's gender and whether they do yoga. We have every respondent's gender, but yoga is missing for a single voter.

Table 2. Hypothetical vote choice data.

vote	female	yoga
Hillary	0	0
Bernie	0	1
Bernie	0	0
Bernie	0	0
Bernie	0	**?**
Hillary	1	1
Hillary	1	0
Bernie	1	1
Hillary	1	1
Bernie	1	0

[2] We could only perfectly predict one variable as a linear combination of the others if they were collinear, in which case the OLS estimator is ill-defined.

MAR amounts to assuming that whether you report doing yoga depends only on your gender, not on what the answer is. We'll assume that moving forward.

Based on the other data available, we would probably guess that the missing respondent doesn't do yoga. Of the other men who responded, only 1/4 do yoga. That goes up to 1/3 if we just look at men who responded and voted for Bernie, but that's still less than half. Nonetheless, we wouldn't want to just impute a 0 and move on. That implies a level of certainty we don't have—there's some chance the guy does yoga!

Multiple imputation would consist of creating M datasets, each identical to the original except that the missing value has been randomly drawn from the relevant conditional distribution. Here, we condition on the missing respondent's gender and vote choice,[3] drawing a 0 for yoga with probability 2/3 and a 1 with probability 1/3. With $M = 5$, for example, we might get the following imputations.

Table 3. Imputed values for the observation with missing yoga.

imputation	vote	female	yoga
1	Bernie	0	0
2	Bernie	0	0
3	Bernie	0	1
4	Bernie	0	1
5	Bernie	0	0

In real-world applications, we will have many more observations, covariates, and missing values to deal with. Standard software implementations of multiple imputation, including **Amelia** (King et al. 2001; Honaker and King 2010), for tractability rely on the assumption that the joint distribution of the data is multivariate normal. This means you must take care when your variables are skewed, bimodal, dichotomous, ordinal, categorical, or otherwise not well-described by a normal distribution. You may want to transform your data before imputing and then un-transform it thereafter. In any case, make sure to read the documentation of whatever software you use and set the appropriate

[3] I personally feel a bit squeamish about including post-treatment and response variables in imputation models. King et al. (2001) and Honaker and King (2010), who know more about this than I do, claim that not only is it not a problem, but indeed it is best practices.

options when imputing—the defaults may well produce misleading results.

Each imputation is a completely filled-in dataset. Once you have the imputations in hand, you apply whatever estimator you would have used if you'd had the full data. Let $\hat{\theta}_m$ denote the estimate from the m'th imputation (e.g., a regression coefficient), and let $\hat{V}_m$ denote its estimated variance (squared standard error). The combined parameter estimate is

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m,$$

and the variance estimate is

$$\hat{V} = \frac{1}{M} \sum_{m=1}^{M} \hat{V}_m + \left(1 + \frac{1}{M}\right) \left(\frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta})^2 \right).$$

The variance estimate is a combination of the variance within imputations and the variance between imputations. Our uncertainty about the imputed values is an additional source of variation in the estimates that this formula accounts for.

Standard practice is to use $M = 5$ or $M = 10$ (Schafer 1999; King et al. 2001). Multiple imputation is computationally intensive, and additional imputations beyond the tenth or so don't reduce the standard errors enough to be worth the computational cost.

Nonignorable Missing Outcomes

Nonignorable missingness is the biggest problem for inference when it occurs in the response variable. In the context of the linear model,

$$Y_i = x_i^\top \beta + \epsilon_i,$$

nonignorable missingness occurs when the probability of missingness is a function of the error term—the part of the response that our covariates don't account for.

For example, suppose we want to estimate the relationship between living in the South and holding racial stereotypes. We run a survey to collect data to

run the regression

$$\text{stereotype}_i = \beta_0 + \beta_1 \text{South}_i + \epsilon_i.$$

Suppose that Southerners are less likely to answer the question, regardless of whether they hold stereotypes. In this case, the data are missing at random. As long as some Southerners answer the question, we can use the observed data to estimate the conditional expectations

$$E[\text{stereotype}_i \,|\, \text{South}_i = 1], \quad E[\text{stereotype}_i \,|\, \text{South}_i = 0]$$

without bias. The difference between these will be an unbiased estimate of β_1, assuming the usual conditions hold (Little 1992).

On the other hand, suppose that people who hold stereotypes are less likely to answer the question. Missingness would then be nonignorable, since the probability of missingness depends on the value of the potentially missing variable. In this case, we cannot estimate the prevalence of stereotypes among Southerners and non-Southerners without bias—the within-group sample means will be underestimated, since stereotyped individuals conceal their answers. Therefore, the difference of sample means will not in general be an unbiased estimate of the population difference.

This sounds like a familiar problem ... because it is. Recall the days of yore when we spoke of causal inference. We wanted to estimate the average treatment effect,

$$\tau = E[\tau_i] = E[Y_i(1) - Y_i(0)],$$

but we were stymied by the fact that we only observed one of the two potential outcomes for each observation.

Following Manski (1990), suppose the potential outcomes are bounded below by Y^L and above by Y^U. For example, with a binary response, $Y^L = 0$ and $Y^U = 1$. The individual effect for a treated observation lies within the bounds

$$\underbrace{Y_i - Y^U}_{\tau_i^L} \leq \tau_i \leq \underbrace{Y_i - Y^L}_{\tau_i^H}.$$

Similarly, the individual effect for a control observation lies within

$$\underbrace{Y^L - Y_i}_{\tau_i^L} \leq \tau_i \leq \underbrace{Y^U - Y_i}_{\tau_i^H}.$$

Without making *any additional assumptions* (e.g., about independence or confounders), we can place bounds on the sample average treatment effect:

$$\frac{1}{N}\sum_{i=1}^{N}\tau_i^L \leq \frac{1}{N}\sum_{i=1}^{N}\tau_i \leq \frac{1}{N}\sum_{i=1}^{N}\tau_i^H.$$

The resulting bounds will usually be wider than we would like them to be. In particular, they will always contain zero. (We would need stronger assumptions to rule out the possibility that $Y_i(0) = Y_i(1)$ for all i.) But they still tell us something—they rule out some possibilities, without relying on any questionable or untestable assumptions. They tell us how far the data alone can bring us.

This illustrates one of the two general approaches to nonignorable missing outcomes:

1. Nonparametric bounds analysis. The benefit of this approach is that it does not depend on restrictive assumptions. The most obvious cost is that you yield a range of possibilities instead of a single, specific estimate—though I would argue that this is a benefit if it prevents unwarranted confidence.

   The other cost is sheer difficulty of implementation for anything more complicated than differences of means. For example, the estimators proposed by Manski and Tamer (2002) and Beresteanu and Molinari (2008) for linear regression with interval-bounded missing outcomes are not at all trivial to implement.

2. Directly modeling the process by which data go missing. This will give you a point estimate, but it requires actually knowing the process by which data go missing.

   The most famous example is Heckman's (1979) model of selection bias. If there are unobserved variables that affect both the probability that the outcome is unobserved and the value of the outcome itself, OLS estimation of the outcome equation will be biased. Heckman derives a correction that entails running a first-stage regression to predict the probability of missingness. It requires an exclusion restriction—we need a variable that affects the probability of missingness but not the outcome itself. (If this sounds a lot like instrumental variables, that's because it is.)

The second option is by far the most popular among political scientists, for better or worse.

References

Beresteanu, Arie, and Francesca Molinari. 2008. "Asymptotic Properties for a Class of Partially Identified Models." *Econometrica* 76 (4): 763–814.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–61.

Honaker, James, and Gary King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data." *American Journal of Political Science* 54 (2): 561–81.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95 (1): 49–69.

Little, Roderick J A. 1992. "Regression With Missing X's: A Review." *Journal of the American Statistical Association* 87 (420): 1227.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *The American Economic Review* 80 (2): 319–23.

Manski, Charles F, and Elie Tamer. 2002. "Inference on Regressions with Interval Data on a Regressor or Outcome." *Econometrica* 70 (2): 519–46.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92.

Schafer, Joseph L. 1999. "Multiple imputation: a primer." *Statistical Methods in Medical Research* 8 (1): 3–15.

Winship, Christopher, and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods & Research* 23 (2): 230–57.