CHAPTER 14

Missing Data Problems in Criminological Research

ROBERT BRAME, MICHAEL G. TURNER, AND RAY PATERNOSTER

INTRODUCTION

Missing data problems are a ubiquitous challenge for criminology and criminal justice researchers (Brame and Paternoster 2003). Regardless of whether researchers are working with survey data or data collected from official agency records (or other sources), they will inevitably have to confront data sets with gaps and holes. As a result, researchers who design studies must take whatever steps that are feasibly and ethically possible to maximize the coverage and completeness of the data they will collect. Even with the best planning and implementation, however, nearly all studies will come up short. Records will be incomplete, targeted survey participants will not be located, some who are located will not participate, and some who participate will not participate fully. These common pitfalls result in a problem most researchers face, in that we want to use the data sets we collect to form inferences about an entire target population – not just the subset of that population for whom valid data are available.

Unfortunately, for the field of criminology and criminal justice, there is little advice and guidance on this issue that is available and accessible to applied researchers (see Allison 2002 for a notable exception). Most discussions of missing data issues are presented as statistical problems with examples from a wide range of fields. It is left to criminology and criminal justice researchers to identify appropriate analogies for the problems they encounter. Further complicating matters, most discussions about missing data are pitched in highly technical language and terminology that is outside the training of most criminology and criminal justice researchers. From the vantage point of most criminologists, serious treatments of missing data issues require significant investments of time and effort to identify appropriate analogies and methods (largely from other fields) or technical assistance by individuals who are already acquainted with these issues. To make matters worse, criminologists may be tempted to "solve" their missing data problems by blindly applying missing data routines now available in canned statistical packages.

A.R. Piquero and D. Weisburd (eds.), *Handbook of Quantitative Criminology*,
DOI 10.1007/978-0-387-77650-7_14, © Springer Science + Business Media, LLC 2010

We, of course, will not resolve the problems identified above with a single essay. Instead, our goal for this essay is to provide a systematic primer of the key missing data issues arising in crime research and a conceptual-level discussion of how these issues can be addressed. In the next section, we discuss some different types of missing data problems that typically arise in the kinds of research projects often conducted by criminologists. Next, in section "Analytic Difficulties," we turn our attention to the specific types of analytic difficulties that are presented by various kinds of missing data problems, while section "Methods for Addressing Missing Data Problems" considers some methods that have been devised to deal with these problems. Finally, in section "Conclusions," we offer some closing thoughts and recommendations for how individual researchers and the field as a whole can make systematic progress in dealing with missing data problems.

TYPES OF MISSING DATA PROBLEMS

A wide range of different types of missing data problems routinely arise in criminology and criminal justice research. In this section, we survey the types of problems that most commonly present themselves and provide some examples of each problem. We specifically address six broad categories of missing data problems in this section: (1) missing administrative and official record data, (2) missing survey data, (3) incomplete responses among survey participants, (4) attrition or loss of cases in longitudinal studies, (5) experimental attrition, and (6) planned nonresponse or incomplete data. While we make no claim that these categories are exhaustive, we do think they cover the broad range of missing data problems that are typically encountered by criminologists.

Missing Administrative and Official Record Data

Major data sources such as the FBI Uniform Crime Reports (UCR), the National Incident Based Reporting System (NIBRS), the Supplemental Homicide Reports (SHR), and data sources used in specific research projects such as arrest, court, prison, probation, and parole records, all suffer from a range of difficulties associated with missing data. For example, the UCR, SHR, and NIBRS programs all depend on voluntary participation by law enforcement agencies, and each year – for a variety of reasons – some fraction of agencies choose not to participate in the program (Maltz and Targonski 2002). For example, it is almost always a bad idea to compare UCR-based crime rate estimates for different jurisdictions because the UCR only measures "crimes known to the police." One basic issue that bedevils such comparisons is that victims in some jurisdictions may be more or less likely to report crimes to the police than victims in other jurisdictions, which itself is a missing data problem. In these instances, inferences about crime rate differences among the jurisdictions will be misleading (i.e., whether one city is "safer" than another).

The problem of missing data also arises in court-related research. For example, in a capital punishment study conducted by Paternoster et al. (2003), the authors relied on data maintained in prosecution and court records for each of the death-eligible offenders. Despite a gubernatorial order granting the researchers direct access to the file folders and case records of each offender, the authors reported substantial amounts of missing data on a number of important characteristics, such as the socioeconomic status of the victim.

Correctional studies are not immune from these problems either. For example, the Bureau of Justice Statistics has conducted two major studies of criminal recidivism among offenders released from prisons in over a dozen states in 1983 (Beck and Shipley 1989) and then again in 1994 (Langan and Levin 2002). In each study, the goal was to take the cohort of offenders released from prison and follow each of them for 3 years to measure recidivism rates in arrest, court, and correctional files. Each of these studies reported nontrivial missing data problems in terms of both the background characteristics and post-release experiences of the offenders.

The use of administrative or official record data is often attractive because of the uniformity of data collection procedures and the standardized recording of events and the dates on which those events occurred. Yet, in criminal justice settings, it is almost inevitable that there will be difficult missing data problems lurking within official record databases. Sometimes, these problems are obvious (e.g., an item is left blank on a rap sheet), and sometimes, they are more subtle (e.g., police can only know about crimes if they witness them or if someone reports them). Nevertheless, researchers working with officially recorded data must always be vigilant to try to think about how missing data problems might be affecting their results.

Missing Survey Data

For the last half century, criminologists have increasingly relied on survey research to study a wide range of crime-related phenomena. In fact, some of the earliest interview studies in criminology challenged longstanding assumptions about the relationship between social class and criminality, which themselves were based on officially recorded data (Nye 1958; Williams and Gold 1972; Tittle et al. 1978; Elliott and Ageton 1980; Hindelang et al. 1981). The power of survey research lies in the insights that intensive detailed study of a relatively small group of people can provide about much larger, interesting populations. Yet, survey research – whether focused on samples drawn from the general population or special populations, or on the study of offending or victimization – must confront the problem of non-response.

Researchers must make decisions about the boundaries of the sampling frame (i.e., who has a chance of being selected to receive a survey). Specifically, if the sample is drawn exclusively from certain subpopulations, then consideration must be given to what is lost by excluding other subpopulations. If the sample is drawn by quota, convenience, purposive, or other nonprobability sampling methods, then consideration should be given to the effects of those sampling designs on both external validity (i.e., to what extent do the results generalize to the entire population of interest?) and internal validity (i.e., are the comparisons within the study fair and reasonable?). The basic problem is that studies that systematically exclude or underrepresent individuals in the target population *may* yield invalid results.

Additionally, researchers must obtain voluntary, informed consent from the individuals they seek to interview. Individuals targeted for survey participation have no obligation to make themselves available to researchers or participate in the research. Thus, survey researchers will always have to deal with two sets of difficult nonresponse problems: (1) some individuals who are targeted for the study can be contacted to determine whether they are willing to participate, while some cannot and (2) among those who are successfully contacted, some will agree to participate, and others will not. Virtually, all survey studies in the criminology literature suffer from nonresponse because of both these issues.

Incomplete Responses Among Survey Participants

If the plan for drawing the sample of survey participants is reasonable and there is no systematic omission of individuals from the study either by failure to contact or refusal to participate, the participants may, nevertheless, not answer some of the questions on the survey (Brame and Paternoster 2003). Sometimes, this type of nonresponse is planned and intentional. An example of planned partial nonresponse would be the use of so-called "gateway questions" or "filter questions" wherein the answer to a certain question dictates whether certain subsequent questions will be skipped. An example would be the question, "Have you used marijuana in the past 12 months?" Those replying "No" to this gateway question would not be asked to respond to the follow-up questions pertaining to the location, social circumstances, and perceived effects of marijuana use. Gateway questions are designed to make the survey more sensible for the respondent, and if planned carefully, they can be helpful for survey researchers. But the use of gateway or contingency questions can present risks for the research when people "skip out" of questions wherein they could have provided useful information. Also, in panel studies, there is a risk of so-called "testing effects" when respondents learn that they can shorten the interview if they answer gateway questions in certain ways.

Leaving aside the issue of gateway questions, respondents are routinely told by researchers – for ethical reasons – that they can refrain from answering questions they don't want to answer. So, respondents can answer some questions but not others. For some items in the survey, the respondent may be able to provide valid data, but, for other items, the respondent may be indistinguishable from individuals who could not be contacted or refused to participate in the survey. A related issue arising in survey research is the proper handling of "don't know" responses or Likert-format "neither agree nor disagree" responses. In this case, researchers cannot categorize their responses into meaningful areas, since respondents lack knowledge of how to respond or have not formed an opinion about the particular subject under investigation.

Attrition in Longitudinal Studies

The criminology literature is increasingly populated by various types of studies with a longitudinal component. In some cases, these studies involve the study of a small number of highly aggregated units over time (time-series studies). In other cases, researchers repeatedly follow the same large group of individuals over time (panel studies). Some studies are also able to follow a large number of units over very long periods of time (pooled cross-sectional time-series studies). Regardless of whether the study is focused on changes in crime rates as neighborhood dynamics unfold within a city or whether the study examines how criminal behavior develops and evolves for individuals over the life span, attrition or drop-out is always a potential threat to the study's validity (Brame and Paternoster 2003).

In fact, different kinds of attrition can cause problems for researchers. Cases may be present at the beginning of a study and then drop out; or they may drop out and then drop back in. Individuals die or become incapacitated, new schools open and old schools close, and neighborhoods emerge and then vanish. In other words, one of the hazards of studying cases over time is that the population of cases that is *relevant* for the study and the population of cases *available* for study are both subject to change while a longitudinal study is underway. And, of course, since study participation is voluntary, prior study participants can change their mind about whether they want to continue participating in a study (also generally an ethical requirement). An interesting and useful feature of studies suffering from attrition is that some information is usually known about the cases that drop out – as distinguished from many missing data problems wherein little or nothing may be known about missing cases (see e.g., Brame and Piquero 2003). The advantage of this is that, sometimes, it is possible to make inferences about the missing data based upon information that is available about the cases.

Attrition in Experimental Studies

Attrition is often viewed as a problem with longitudinal studies, but it is an equally important problem in criminal justice field experiments. As an example, consider the experimental studies in the Spouse Assault Replication Program (see e.g., Sherman and Berk 1984; Hirschel and Hutchison 1992). In these studies, misdemeanor domestic violence cases were randomly assigned to arrest and nonarrest treatment conditions. An important outcome measure in these experiments is information about future victimization obtained from victim interviews. The problem is that many of the victims – for a variety of reasons – could not be interviewed. Even though the cases in the different treatment groups are comparable to each other because of the experimental design, it does not follow that the subsample of interviewed cases in the different treatment groups are comparable to each other.

Table 14.1 presents a concrete example of this important problem from the Charlotte Spouse Abuse Experiment (Brame 2000). In this study, only about half of the victims participated in the 6-month follow-up interviews. Within this subsample of interviewed cases, approximately 60% of the victims reported that they had been revictimized. What the observed victim interview data cannot tell us is whether the revictimization rates are higher, lower, or about the same for the noninterviewed cases. Another, more complicated, issue is whether these differences vary in important ways between the different treatment groups. In sum, if the revictimization rates vary in some systematic way for the missing and observed cases, we will have problems obtaining a valid estimate of the treatment effect.

It is fair to point out that this problem is not unique to experimental designs. But the purpose of conducting an experiment is to ensure that comparisons of different groups are, in fact, apples-to-apples comparisons. If interview nonresponse or some other form of attrition leaves us with groups that are not comparable or attrition biases that cause other problems,

TABLE 14.1. Prevalence of victim interview-based recidivism by treatment

Victim interview-based recidivism	Assigned Treatment Group			
	Arrest	Citation	Separate/advise	Total
No subsequent assaults	46	43	41	130
At least one subsequent assault	66	81	61	208
Number interviewed	112	124	102	338
Total number of cases	214	224	212	650
Proportion of cases interviewed	0.500	0.554	0.481	0.520
Proportion of interviewed cases with new victimization	0.589	0.653	0.598	0.615

the additional effort and expense that often go into planning and implementing an experiment might not be worthwhile. In these instances, it may be preferable to conduct a well-controlled observational study.

Planned Nonresponse or Incomplete Data

In some situations, nonresponse or the acquisition of incomplete information is actually formally built into a study's research design. The classic example of this situation is the United States Census Bureau's long and short forms used in conducting the decennial census of the population (which criminologists often analyze) (Gauthier 2002:137). While the Census Bureau attempts to collect certain pieces of information from the entire U.S. population, a more detailed database of information is collected on a probability sample of the population. On the basis of the information provided by the probability sample, it is possible to form inferences about the more detailed information on the long form for the entire population. In this type of situation, planned nonresponse provides a way to make research, data collection, and analysis activities more efficient and economically feasible.

Other types of nonresponse and incomplete data problems, while not built formally into the design, are anticipated with certainty by the research team. For example, a typical recidivism study must pick a date (or dates) on which to conduct follow-up record searches to identify those who recidivated and individuals who did not. One problem is that the status of nonrecidivist is subject to change each and every day. If an individual is a nonrecidivist at the time the records are searched but then recidivates the next day, that individual is counted as a nonrecidivist for purposes of the study. This is also referred to as "right censored data" or "suspended data" (Maltz 1984; Schmidt and Witte 1988). The term "right censored" indicates that the events of interest (i.e., committing new offenses) could have occurred to the right of the last data collection point and therefore go unobserved if subsequent data collections are not administered.

Another type of censoring problem is known as "interval censored data" wherein the analyst does not know the exact time an event occurred, but only that it did, in fact, occur. An example of interval censored data is when individuals are assessed every 6 months and asked to self-report their involvement in criminal activity. At the time of the assessment, it is only determined that the individual did or did not recidivate. If a unit did recidivate, it is not known precisely when the new offense occurred, but only that a failure (a new offense) did indeed occur; instead of an exact time to failure, the analyst would only be in a position to record that the failure occurred between, for example, the third and fourth data collection interval. A special case of interval censored data is referred to as "left censored data" wherein the failure occurred between time zero and a particular inspection time. To continue with the example above, left censoring would be present if a failure occurred prior to the first assessment; the specific time after release before which the event occurred is unknown.

The provisional nature of outcomes, as illustrated above, is not a purely criminological problem. In fact, researchers in other fields such as engineering, medicine, psychology, sociology, and economics often study outcomes that are subject to change after the study has ended (Maltz 1984). As discussed above, generally, researchers refer to this problem as "censoring," and data with provisional outcomes are called "censored." Studies can always be lengthened or extended, so updating can occur, but unless the study is carried out over very long periods of time (sometimes requiring multiple generations of researchers), with very specific documentation in detailing the time in which an event occurred, the problem of censoring will always be present and require attention.

A different but related issue arises when cases are only observed when they exhibit certain characteristics or behaviors and those characteristics or behaviors are actually part of the outcome measure. For example, suppose we wish to study variation in the frequency of offending, only individuals who offend at least one time are observed. In these situations, we say the sample is “truncated.” Truncation is different from censoring. When data are censored, we actually get to see characteristics of individuals who are censored – what we can’t see is their final status; only their provisional status can be seen. When data are truncated, we only get to see the characteristics of individuals who score above the lower limit and below the upper limit of truncation; for other cases, we know nothing about the variables of interest. Thus, censoring is actually more of a problem of incomplete, but partially observed, data, while truncation is actually a problem of completely missing data for cases outside the limits of inclusion in the sample.

Overall, what planned nonresponse, censoring, and truncation all have in common is that they can be anticipated ahead of time by researchers, and analytic methods for addressing them can be planned in advance of data collection efforts (Maddala 1983). Planning for other types of missing data problems will usually be more speculative until the study is underway or concluded.

ANALYTIC DIFFICULTIES

Up to this point, we have been discussing specific kinds of missing data problems without saying much about the difficulties and consequences they create. In this section, we will consider how some of these problems correspond and relate to terminology used by statisticians to describe various missing data problems. We then consider the consequences of different types of missing data problems for the kind of inferential work usually conducted by criminologists. Although the problems discussed in this section are not necessarily mutually exclusive or exhaustive, we think the labels we use and the problems we describe will be recognizable to most criminologists as the most prevalent problems in the field.

Missing Units

This is perhaps the most basic type of missing data problem. It occurs when cases that should have been included in the study (to support the study inferences) are not included. The problem of missing units can occur for a variety of reasons: data were not available, potential participants could not be located or refused to participate, or potential participants were not included in the sampling frame. The researcher may be aware of which cases are missing and know some things about those cases, or the researcher may be oblivious to the existence of those cases. The consequences of missing units can range from simply reducing statistical efficiency of parameter estimates (i.e., increasing the size of the standard errors) to severely biased and misleading parameter estimates.

Missing Items on Otherwise Observed Units

This problem arises when some data are available but other data is missing for certain cases. Items could be missing from administrative data sources because of data recording or data entry problems, individuals may respond "don't know" or prefer not to answer some questions, they may not have seen the question or some variation of planned non-response such as censoring, truncation, or a mixture of simple surveys combined with intensive surveys on a sample may be the source of partially observed data. Attrition from longitudinal studies can also be viewed as a special case of the problem of partially observed data; at some time points, a case supplies data, while at other time points, that case is missing. The range of consequences for this problem is similar to that of the missing units problem: a loss of statistical efficiency and power but no bias to severely biased and misleading parameter estimates.

Missing Completely at Random

According to Little and Rubin (1987), units which are not observed or only partially observed may be missing completely at random. Allison (2002: 3) writes that the data on some measured variable Y are said to be missing completely at random (MCAR) "if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variables in the data set." If $p(x|y) = p(x)$, we say that x and y are independent of each other. Using the notation of Little and Rubin (1987), MCAR implies that

$$p(M|x_\mathrm{o}, x_\mathrm{m}) = p(M) \tag{14.1}$$

where M is a variable which indicates whether individuals have missing data or not and x_o and x_m refer to observed and missing data. If missingness is independent of both the observed and missing data, we say that the data are missing completely at random.

Oftentimes, this is a desirable result because it means that the observed sample is a representative sample of the target population. In other words, on average, missing and observed cases do not differ systematically from each other. A violation of the MCAR assumption would occur if, say, those who fail to provide responses to self-reported offending items on a questionnaire were also on average lower in self-control than those who supplied complete information. However, as we will argue below, even if the missing data on one variable is unrelated to missing data on others, it does not prove that the data are missing completely at random.

The main consequence of MCAR is a loss of efficiency. That is, parameter estimates obtained from MCAR samples will have larger standard errors (owing to smaller sample sizes) than estimates obtained from samples with no missing data at all. But estimator properties of unbiasedness or consistency (bias tends to zero as sample size increases) will still hold in MCAR samples.

Unfortunately, while the MCAR assumption can be called into question by identifying differences between missing and observed cases on characteristics that are observed for all cases, it is a very strong assumption and can never be proven or empirically demonstrated to be true. Even if missing and observed cases look similar to each other on characteristics observed for both groups, there may not be many characteristics on which to compare them; even if there are more than a few characteristics on which to compare them, it is always possible that the groups look different on factors that can only be seen in the observed cases, or that the groups differ with respect to factors that are unobserved for everyone. If a researcher wishes to assert MCAR (which is often a hazardous exercise), the strongest case is showing that observed and missing cases are comparable to each other on a wide range of characteristics (that can be seen for both groups), and a plausible narrative about why the process producing the missing data is likely to operate in a way that will lead to no systematic differences between observed and missing cases.

Missing at Random

According to Little and Rubin's (1987) taxonomy of missing data, cases are missing at random (MAR) when the missing and observed cases are similar to each other after conditioning the analysis on measured variables for the observed cases. They note that the data are missing at random when

$$p(M|x_{\mathrm{o}}, x_{\mathrm{m}}) = p(M|x_{\mathrm{o}}) \tag{14.2}$$

which implies that the missing data patterns are not completely random, but that they are random after conditioning on the observed data. If the probability model for the missing data is $p(M)$ or $p(M|x_{\mathrm{o}})$, then the missing data mechanism is "ignorable."

This assumption is weaker than MCAR, which requires missing and observed cases similar to each other *without conditioning* on any variables – measured or unmeasured. In our previous example, the self-reported offending data would be missing at random if the probability of missing self-report data depended upon a person's level of self-control, but within every level of self-control, the probability of an individual missing data on self-reported offending was unrelated to one's level of self-reported offending.

Similar to MCAR, the validity of the MAR assumption cannot be tested. Clearly, it is not possible to know whether conditioning on measured variables for observed cases is a sufficient adjustment for biases caused by unmeasured variables. In other words, because we do not observe the missing data on self-reported offending, we cannot compare the level of self-control for those with and without self-reported offending data. Thus, the validity of MAR is asserted – but cannot be demonstrated by the researcher, and the assertion is generally accompanied by a narrative about why it is reasonable to think that conditioning on measured variables for the observed cases is sufficient to remove bias caused by missing data. Most formal adjustments for missing data in the criminology and criminal justice literature invoke – either implicitly or explicitly – the MAR assumption.

Nonignorable Nonresponse

According to Little and Rubin (1987) and Allison (2002), the MCAR and MAR assumptions imply that the "missing data mechanism" (i.e., the process which determines which cases are observed and which are missing) is ignorable. In these situations, it will be possible to obtain a point estimate and a measure of uncertainty for that point estimate (i.e., a standard error). However, in some situations, the MCAR and MAR assumptions will not be plausible. For example, it may be that there are simply not many variables measured even for the observed cases, or that variables which could help build a convincing case for MAR are not measured. The basic problem here is that conditioning on measured variables for the observed cases is not sufficient to ensure that the missing and observed cases are equivalent on variables where there is missing data. Little and Rubin (1987) note that nonignorable missing data exists when:

$$p(M|x_{\mathrm{o}}, x_{\mathrm{m}}) \tag{14.3}$$

and is not equal to either $p(M)$ or $p(M|x_{\mathrm{o}})$.

As noted by Little and Rubin (1987), nonignorable missing data models generally imply the use of a sensitivity analysis of some sort. More specifically, nonignorable models require the researcher to formally characterize or specify the process by which some cases have observed data and other cases have missing data in order to arrive at satisfactory parameter estimates. Since the process generating the missing data is not generally well understood and the data contain no information about what is driving the process, it is advisable for the researcher to consider a number of different possible missing data mechanisms. This type of analysis, then, allows the researcher to consider the sensitivity of the final analysis results to variation in untestable assumptions about the missing data mechanism. Little and Rubin (1987) and Rubin (1987) also discuss how this approach can be implemented within a Bayesian framework where the goal is to estimate the posterior distribution of the parameter of interest after specifying appropriate prior distributions for the missing data mechanism and the parameters of substantive interest. In one respect, nonignorable analyses are attractive because they allow us to relax some of the restrictive assumptions on which MCAR and MAR analyses depend. But, these models will also require the researcher to tell a more complicated story about the findings and how sensitive the findings are to different sets of untestable assumptions. These models also tend to be more difficult to estimate and often require specialized computer programming and software development.

Sample Selection Bias

One special class of missing data problems arising in criminology and criminal justice research is the problem of studying individuals who are observed because some event has occurred (i.e., a selected sample) (Glynn et al. 1986). For example, if we study decisions to imprison offenders, we must bear in mind that individuals who could be imprisoned (or placed on probation) must first be convicted of a crime. To obtain valid parameter estimates of the effects of covariates on sentencing outcomes, it would be necessary to properly adjust for the selection process. A widely used approach is the Heckman (1976) two-stage procedure wherein one initially estimates the selection process before obtaining estimates of the substantive process of interest.

Ideally, one would simultaneously estimate models of both the conviction process among all charged offenders and the sentencing process among the convicted offenders. A problem that arises in this literature is the need for valid exclusion restrictions. An exclusion restriction is a variable that is assumed to have important effects on the selection variable (first stage) but no effect on the substantive outcome variable (second stage). If this assumption can be justified theoretically, then it is possible to jointly model the processes governing sample selection and the substantive outcome. If the assumption cannot be justified, then the statistical model depends on untestable assumptions about the bivariate distribution of the error terms of the selection and substantive equations; this is a condition that Smith and Paternoster (1991) and Smith et al. (1989) referred to as “weak identification.”[1] Little and Schenker (1995) go even further, concluding that “the approach cannot be generally recommended.” A recent discussion of this estimator by Bushway et al. (2007) appears in the criminology literature and offers detailed consideration of the advantages, disadvantages and difficulties involved in implementing this estimator.

METHODS FOR ADDRESSING MISSING DATA PROBLEMS

In this section, we discuss the various methodological strategies researchers could employ when faced with missing data problems. In so doing, we offer some thoughts about the kinds of issues researchers should consider when making decisions about how to proceed when missing data is a problem. Inevitably, many of these decisions are predicated on the assumptions about the types of missing data discussed above.

Listwise Deletion of Missing Cases

One simple option is to drop cases on any observations that have missing data for any variable that will be included in the model to be estimated and then proceed with a conventional analysis. When the researcher believes the data are MCAR, no further adjustments are necessary for valid analyses, since the sample based upon the listwise deletion will be a random sample of the original. As noted above, however, MCAR is a strong assumption and should be accompanied by a discussion as to why it is plausible in any particular study. Further, if the data are MAR and not MCAR, then parameter estimates will be biased. Allison (2002: 6) notes that listwise deletion of missing data is the method that is the most robust to violation of the MAR assumption among the independent variables in a regression analysis. In other words, if the probability of missing data on any of the independent variables is not related to the dependent variable, then obtained regression estimates will be unbiased. Cases with missing data are also dropped in some other kinds of analyses – particularly when weights for nonresponse are used. But most efforts to weight do not invoke the MCAR assumption. In these instances, missing cases are dropped, but the assumption underlying the weights is that the data are MAR or that the missing data mechanism is nonignorable.

Pairwise Deletion of Missing Cases

Some multivariate analyses – such as factor analysis, linear regression analysis, and structural equation modeling – are capable of generating solutions and estimates based exclusively on correlation or covariance matrices involving two or more variables (some of which may have missing data). Observations with missing data can be deleted before estimating these matrices (listwise deletion) or each of the individual correlations or covariances can be estimated on the

[1] In effect, a well-justified exclusion restriction allows for the estimation of a MAR model; in the absence of a well-justified exclusion restriction, the only alternative is to impose constraints on the bivariate distribution of the error term. Any constraint imposed would be an example of a nonignorable missing data model.

cases that have valid data on which the correlation can be estimated (pairwise deletion). Some software packages provide researchers with the option of selecting listwise or pairwise (also known as available case analysis) deletion of missing data. As Little and Schenker (1995:45) note, an important problem with correlation or covariance matrices estimated after pairwise deletion is not always positive definite (which implies noninvertability). This problem is more likely to arise when there is a good deal of variation in the rates of missing data for the variables included in the matrix.

Dummy Coded Missing Data Indicators

Sometimes, when missing data exists for an independent or control variable, X, in a regression analysis, researchers can create a dummy variable D, coded 1.0 if the variable has missing data and 0.0 otherwise (Cohen and Cohen 1983). Another variable X^* is also created such that

$$\begin{aligned} X^* &= X \text{ when data are not missing on } X \\ X^* &= Z \text{ when data are missing on } X \end{aligned}$$

where Z can be any constant, though for interpretational purposes Z is usually the mean of X among those cases without missing data. This technically allows each case to contribute information to the analysis. This can be done for any number of right-hand-side variables that have missing data. Then, X^*, D and other variables can be included in the estimated regression model. In this instance, the coefficient for D is the predicted value of the dependent variable for those with missing data on X minus the predicted value of Y for those at the mean of Y, controlling for other variables in the model. The estimated coefficient for X^* is the effect of X among those that have observed data on X.

While this type of analysis will provide parameter estimates, the estimator does have two important shortcomings. First, the coefficient estimates are biased (Allison 2002: 11). Second, the missing dummy variables only allow researchers to see whether rates of missing data vary for different groups within the sample under study. They do not tell us anything about the true unobserved values of the variables that are missing. Other approaches that rely on imputation and weighting attempt to overcome this problem.

Model-Based Imputation

Another method for keeping cases with missing observations in the analysis is to impute or “plug” missing values with some guess about what the value of the variable would have been if it had been observed, and the analysis then proceeds as if there were no missing data (Little and Rubin 1987). The basis for this guess or prediction is the imputation model. In the simplest case, an analyst could simply substitute a mean or median to fill in the missing data. This method, mean substitution, is known to produce biased estimates of the variances and covariances. The analyst might also estimate a regression model on the nonmissing cases wherein the dependent variable, Y, is the variable with missing data and the independent variables are the other relevant variables in the data set thought to be related to Y. Then, the model is used to estimate predicted values of the missing variable for the cases with missing data. These predictions could be plugged in directly or they could be drawn from a probability distribution for that variable. Rubin (1987) has suggested that only three to ten imputations are likely to be needed in most applications. He has shown (1987: 114) that the efficiency of an estimate based on m imputations can be determined as

$$\left(1 + \frac{\gamma}{m}\right)^{-1} \tag{14.4}$$

where γ is the proportion of missing data.

Unless there is a large proportion of missing data, there is little computational advantage in producing more than five imputations. While there are different methods to impute the missing data, all suffer from the same problem that they produce coefficient standard errors that are smaller than what they should be, leading to biased hypothesis tests.

When conducting a final analysis with imputed data, it is important to adjust the standard errors for the uncertainty or sampling variation that is created by the fact that the imputed values are not known. Since they are estimated, they create additional variation that must be taken into consideration. Therefore, standard error formulas that ignore the imputation of the missing data will generally be biased toward zero. Imputations can be based on nonignorable models, or they can assume MAR.

Hot Deck Imputation

Hot deck imputation, used frequently by the U.S. Census Bureau to impute missing values in public-use data sets, differs from model-based imputation insofar as the basis for the imputation is not a statistical model but a "donor" case with valid data (Heitjan and Little 1991). First, the researcher identifies a case with missing data on a variable but has some observed characteristics. Next, the researcher looks for a donor case with valid data and a similar set of observed characteristics. The donor's value on the missing variable is then used as the basis for imputing data for the case with missing information on that variable. Generally, the donor's value would not be directly imputed but would first be perturbed and then imputed. As in model-based imputation, formulas that take the uncertainty of the imputation into account in calculating standard errors must be used in the final analysis.

Weighting

One of the most common methods of adjusting for missing data involves the construction of sampling weights. In general, a sampling weight is the inverse of the probability of being selected for the sample. The problem with missing data is that the self-weighted sample represents only the nonmissing cases. The goal of weighting estimators to adjust for missing data is to ensure that the weighted sample appropriately represents both the missing and the nonmissing cases. One way to construct the sampling weights, then, is to identify the individuals with valid data who look very much like the individuals with missing data on characteristics that are observed for all cases. In the final weighted analysis, these individuals will receive more weight. On the other hand, observed individuals who look less like the missing cases on observed characteristics will tend to receive less weight. Although the goal of weighting the sample is to ensure that both observed and missing cases are adequately represented in the final analysis, the success of that effort cannot be rigorously tested. Thus, weighted estimators share some of the same ambiguities that confront other approaches to missing data; ultimately, if the data are missing, there will be some conjecture about whether corrective measures produce valid inferences.[2]

EM Algorithm

Tanner (1996) is one of many careful discussions about the expectation–maximization (EM) algorithm (see also, McLachlan and Krishnan 1997). The principle behind the algorithm is to iterate to a final solution by imputing or estimating data (either imputation or certain statistics) (the E or expectation step) and then maximizing the log-likelihood function (the M or maximization step). Each of the iterations consists of one E-step and one M-step. Once improvement in maximizing the log of the likelihood function becomes very small, the algorithm is said to have converged. As Tanner (1996:64) notes, the EM algorithm allows researchers to address missing data problems by substituting a single, complex analysis with "a series of simpler analyses." A number of computer programs are now available for estimating models via the EM algorithm. Schafer (1997, 1999) has made available the NORM program that does the EM algorithm and is free to users at: http://www.stat.psu.edu/∼jls/misoftwa.html. SPSS version 16 also contains a missing data procedure that does an EM algorithm. Nevertheless, the EM algorithm is a more advanced topic and expert assistance should be sought when trying to implement it in applied settings.

Mixture Models

Another advanced approach to dealing with missing data is to view the entire population as a mixture of two subpopulations (observed and missing cases). The overall inference is based on averaging the parameters for the two groups where the weight of each group is directly proportional to the percentage of the population in that group (Holland 1986:150). Glynn et al. (1986) and Holland (1986) discuss this model as an alternative to traditional sample selection models. According to these authors, the mixture parameterization is more transparent than the sample selection model. In fact, mixture models show clearly how inferences about the entire population are sensitive to the specification of the model that corresponds to the missing observations. The implication of this work is that the identification of both mixture and selection models depends on strong assumptions that may be too strong for ignorable analyses. Sensitivity analysis may be the best way to proceed.

CONCLUSIONS

Missing data problems are a nearly ubiquitous feature of criminology and criminal justice research projects. Whether the research is focused on the analysis of administrative records within a criminal justice agency or direct interviews with offenders or crime victims, missing data is almost certain to create ambiguity and difficulties for analysts. There are reasons

[2] Additionally, weighted estimators require specialized formulas for calculating variances and standard errors, which are available in some but not all software packages.

to believe that the problem of missing data in criminal justice research may become more pronounced in years to come. For example, in criminal justice field research settings, it is becoming increasingly common for Institutional Review Boards (IRB's) to require researchers to be more comprehensive in their consideration and description of risks to study participants. As researchers begin to emphasize the risks of research to potential study participants, they may begin to withhold their consent at higher rates. Additionally, in an era of increased reliance on cell phones and decreased reliance on landline telephones, it is rapidly becoming more difficult to track individuals down and contact them to participate in research studies. This problem is compounded in longitudinal studies as contact information may change over the course of the study. Despite the regularity with which incomplete data and nonresponse occur, however, the issue has received relatively little attention within the field.

Fortunately for the field, a wide range of tools have been emerging which should lead us to more principled treatments of missing data problems. These treatments range from the use of selection and mixture models to imputations and nonresponse weighting. Commonly used statistical software packages such as SAS, R, and Stata have all seen major improvements to their ability to handle missing data in more rigorous ways in recent years. What is needed now is more intensive training of researchers in the thoughtful use of these tools. This will require criminologists and criminal justice researchers to actively engage with professional statisticians and methodologists about the interesting and difficult missing data problems that commonly arise in our field.

Among the most pressing needs is for criminology and criminal justice researchers to become aware of the distinctions between different types of missing data, how missing data problems threaten the validity of commonly used research designs in the field, and the best ways of addressing the uncertainty missing data creates. In particular, we think the field would benefit by thinking more carefully about including sensitivity analysis in studies with missing data. While sensitivity analysis may not be the most prominent feature of our data analysis efforts, it should probably be present most of the time and be discussed in the appendices and footnotes of our research papers. Ultimately, missing data implies that uncertainty and more systematic, careful consideration of missing data problems will be a step in the right direction.

REFERENCES

Allison PD (2002) Missing data. Sage, Thousand Oaks, CA

Beck AJ, Shipley BE (1989) Recidivism of prisoners released in 1983. Bureau of Justice Statistics, Special Report. U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC

Brame R (2000) Investigating treatment effects in a domestic violence experiment with partially missing outcome data. J Quant Criminol 16:283–314

Brame R, Paternoster R (2003) Missing data problems in criminological research: two case studies. J Quant Criminol 19:55–78

Brame R, Piquero AR (2003) Selective attrition and the age-crime relationship. J Quant Criminol 19:107–127

Bushway SD, Johnson BD, Slocum LA (2007) Is the magic still there? the use of the Heckman two-step correction for selection bias in criminology. J Quant Criminol 23:151–178

Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, New York

Elliott DS, Ageton SS (1980) Reconciling race and class differences in self-reported and official estimates of delinquency. Am Sociol Rev 45:95–110

Gauthier JG (2002) Measuring America: The Decennial Census From 1790 to 2000. U.S. Department of Commerce, Washington, DC

Glynn RJ, Laird NM, Rubin DB (1986) Selection modeling versus mixture modeling with nonignorable nonresponse. In: Wainer H (ed) Drawing inferences from self-selected samples. Springer, New York, pp 115–142

Heckman J (1976) The common structure of statistical models of truncated, sample selection and limited dependent variables, and a simple estimator of such models. Ann Econ Soc Meas 5:475–492

Heitjan DF, Little R (1991) Multiple imputation for the fatal accident reporting system. Appl Stat 40:13–29

Hirschel D, Hutchison IW (1992) Female spouse abuse and the police response: The Charlotte, North Carolina Experiment. Journal of Criminal Law and Criminology 83:73–119

Holland PW (1986) A comment on remarks by Rubin and Hartigan. In: Wainer H (ed) Drawing inferences from self-selected samples. Springer, New York, pp 149–151

Hindelang MJ, Hirschi T, Weis JG (1981) Measuring delinquency. Sage, Beverly Hills, CA

Langan PA, Levin DJ (2002) Recidivism of prisoners released in 1994. Bureau of Justice Statistics Special Report. U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC

Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York

Little RJA, Schenker N (1995) Missing data. In: Arminger G, Clogg CC, Sobel ME (eds) Handbook for statistical modeling in the social and behavioral sciences. Plenum, New York, pp 39–75

Maddala GS (1983) Limited-dependent and qualitative variables in econometrics. Cambridge University Press, New York

Maltz MD (1984) Recidivism. Academic, New York

Maltz MD, Targonski J (2002) A note on the use of county-level UCR data. Journal of Quantitative Criminology 18:297–318

McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley, New York

Nye F (1958) Family relationships and delinquent behavior. Wiley, New York

Paternoster R, Brame R, Bacon S Ditchfield A (2003) An empirical analysis of Maryland's Death Sentence System with respect to the influence of race and legal jurisdiction. Unpublished manuscript. University of Maryland, College Park, MD

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London

Schafer JL (1999) Multiple imputation: a primer. Stat Methods Med Res 8:3–15

Sherman LW, Berk RA (1984) The specific deterrent effects of arrest for domestic assault American Sociological Review 49(2):261–271

Schmidt P, Witte AD (1988) Predicting recidivism using survival models. Springer, New York

Smith DA, Paternoster R (1991) Formal processing and future delinquency: deviance amplification as selection artifact. Law Soc Rev 24:1109–1131

Smith DA, Wish ED, Jarjoura GR (1989) Drug use and pretrial misconduct in New York City. J Quant Criminol 5:101–126

Tanner MA (1996) Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions, 3rd edn. Springer, New York

Tittle CR, Villemez WJ, Smith DA (1978) The myth of social class and criminality: an empirical assessment of the empirical evidence. Am Sociol Rev 43:643–656

Williams JR, Gold M (1972) From delinquent behavior to official delinquency. Soc Probl 20:209–229