Journal Title: Applied regression analysis and generalized linear models /

Volume:
Issue:
Month/Year: 2008
Pages: 548-586

Article Author: John Fox

Article Title: Missing Data in Regression Models

605-646

Call #: HA31.3 .F69 2008

Location: F

CUSTOMER HAS REQUESTED:
Electronic Delivery: Yes
Alternate Delivery Method:
YesHold for pickup

Jason Geller (jg9120)
101 Lassen Court Apt 9
Princeton, NJ 08904

Note

THIRD EDITION

APPLIED REGRESSION ANALYSIS and GENERALIZED LINEAR MODELS

John Fox

McMaster University

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Boston

20 Missing Data in Regression Models

Missing data are a regrettably common feature of data sets in the social sciences. Despite this fact, almost all statistical methods in widespread use, including the methods introduced in the previous chapters of this book, assume that the data in hand are complete.

The current chapter provides a basic introduction to modern methods for handling missing data. The first section of the chapter draws some basic distinctions concerning the processes that generate missing data. The second section briefly describes traditional methods for coping with missing data and explains why they are problematic. The third section shows how the method of maximum likelihood (ML) can be used to estimate the parameters of statistical models in the presence of missing data. The fourth section introduces multiple imputation of missing data—a general, flexible, and convenient method for dealing with missing data that can perform well in certain circumstances. The final section of the chapter introduces methods for handling selection bias and censored data, which are special kinds of missing data.

Data may be missing for a variety of reasons:

- In survey research, for example, certain respondents may be unreachable or may refuse to participate in the survey, giving rise to *global* or *unit nonresponse*.
- Alternatively, again in survey research, some respondents may not know the answers to specific questions or may refuse to respond to them, giving rise to *item nonresponse*.
- Missing data may also be produced by errors in data collection—as when an interviewer fails to ask a question of a survey respondent—or in data processing.
- In some cases, missing data are built into the design of a study, as when particular questions in a survey are asked only of a random subset of respondents.
- It is sometimes the case that data values in a study are *censored*. The most common example of censored data occurs in *survival analysis* (also called *event-history analysis*, *duration analysis*, or *failure-time analysis*), which concerns the timing of events. In a prototypical biomedical application, subjects in a clinical trial are followed for a fixed period of time, and their survival times are recorded at their deaths. Some subjects, however, happily live beyond the termination of the study, and their survival times are therefore censored. Survival analysis is beyond the scope of this book,[1] but censored data can occur in other contexts as well—as, for example, in an exam with a fixed number of questions where it is not possible to score fewer than 0 nor more than the total number of questions correct, no matter how little or much an individual knows.

[1] There are many texts on survival analysis. For example, see Allison (2014) for a brief introduction to survival analysis or Hosmer and Lemeshow (1999) for a more extensive treatment.

Missing data, in the sense that is developed in this chapter, should be distinguished from data that are *conditionally undefined*. A survey respondent who has no children, for example, cannot report their ages. Conditionally undefined data do not threaten the representativeness of a sample as truly missing data do. Sometimes, however, the distinction between missing and conditionally undefined data is not entirely clear-cut: Voters in a postelection survey who did not vote cannot be asked for whom they voted, but they could be (and may not have been) asked whether and for whom they had a preference. Similarly, some respondents asked to state an opinion on an issue may not have an opinion. Are these data missing or simply nonexistent?

It is important to realize at the outset that there is no magic cure for missing data, and it is generally impossible to proceed in a principled manner without making at least partly unverifiable assumptions about the process that gives rise to the missing information. As King Lear said, "Nothing will come of nothing" (although he applied this insight unwisely).

20.1 Missing Data Basics

Rubin (1976) introduced some key distinctions concerning missing data.[2] Let the matrix $\underset{(n\times p)}{\mathbf{X}}$ represent the complete data for a sample of n observations on p variables.[3] Some of the entries of $\mathbf{X}$, denoted by $\mathbf{X}_{\text{mis}}$, are missing, and the remaining entries, $\mathbf{X}_{\text{obs}}$, are observed.[4]

- Missing data are said to be *missing completely at random* (*MCAR*) if the missing data (and hence the observed data) can be regarded as a simple random sample of the complete data. Put alternatively, the probability that a data value is missing, termed *missingness*, is unrelated to the data value itself or to any other value, missing or observed, in the data set.
- If, however, missingness is related to the observed data but—conditioning on the observed data—not to the missing data, then missing data are said to be *missing at random* (*MAR*). In a survey, for example, certain individuals may refuse to report their income, and these people may even differ systematically in income from the sample as a whole. Nevertheless, if the observations are independently sampled, so that one respondent's decision to withhold information about income is independent of others' responses, and if, *conditional on* the information that the respondent does provide (e.g., education, occupation), failure to provide information on income is independent of income itself, then the data are MAR. MCAR is a stronger condition—and a special case—of MAR.
- Finally, if missingness is related to the missing values themselves—that is, if the probability that a data value is missing depends on missing data (including, and indeed usually, the data value itself), even when the information in the observed data is taken into account—then missing data are said to be *missing not at random* (*MNAR*). For example, if conditional on all the observed data, individuals with higher incomes are more likely than others to withhold information about their incomes, then the missing income data are MNAR.

[2]Although Rubin's terminology is potentially confusing, it is in common use and has guided most subsequent work on missing data by statisticians. It would therefore be a mistake, I think, to introduce different terms for these concepts.

[3]If you are unfamiliar with matrix notation, simply think of the matrix $\mathbf{X}$ as a rectangular table of data, with the observations given by the n rows of the table and the variables by the p columns.

[4]Despite the notation, $\mathbf{X}_{\text{mis}}$ and $\mathbf{X}_{\text{obs}}$ are not really matrices; they are, rather, subsets of the complete data matrix $\mathbf{X}$. Together, $\mathbf{X}_{\text{mis}}$ and $\mathbf{X}_{\text{obs}}$ comprise $\mathbf{X}$.

These distinctions are important because they affect the manner in which missing data can be properly handled. In particular, if the data are MCAR or MAR, then it is not necessary to model the process that generates the missing data to accommodate the missing data. When data are MCAR or MAR, the "mechanism" that produces the missing data is therefore *ignorable*. In contrast, when data are MNAR, the missing-data mechanism is *nonignorable*, and it becomes necessary to model this mechanism to deal with the missing data in a valid manner.

Except in some special situations, it is not possible to know whether data are MCAR, MAR, or MNAR. We may be able to show that missingness on some variable in a data set is related to observed data on one or more other variables, in which case we can rule out MCAR, but the converse is not the case—that is, demonstrating that missingness in a variable is not related to observed data in other variables does not *prove* that the missing data are MCAR (because, e.g., nonrespondents in a survey may be differentiated from respondents in some *unobserved* manner). If, on the other hand, a survey question is asked of a random subset of respondents, then data are MCAR by design of the study.

> Missing data are missing completely at random (MCAR) if the missing data can be regarded as a simple random sample of the complete data. If missingness is related to the observed data but not to the missing data (conditional on the observed data), then data are missing at random (MAR). If missingness is related to the missing values themselves, even when the information in the observed data is taken into account, then data are missing not at random (MNAR). When data are MCAR or MAR, the process that produces missing data is ignorable, in the sense that valid methods exist to deal with the missing data without explicitly modeling the process that generates them. In contrast, when data are MNAR, the process producing missing data is nonignorable and must be modeled. Except in special situations, it is not possible to know whether data are MCAR, MAR, or MNAR.

20.1.1 An Illustration

To clarify these distinctions, let us consider the following example (adapted from Little & Rubin, 1990): We have a data set with $n = 250$ observations and two variables. The first variable, X_1, is completely observed, but some of the observations on X_2 are missing. This pattern—where one variable has missing data and all others (in this instance, *one* other variable) are completely observed—is called *univariate missing data*. Univariate missing data are especially easy to handle. For example, while general patterns of missing data may require iterative techniques (as described later in this chapter), univariate missing data do not. Nevertheless, we will get a great deal of mileage out of this simple example.

For concreteness, suppose that the complete data are sampled from a bivariate-normal distribution with means $\mu_1 = 10$, $\mu_2 = 20$, variances $\sigma_1^2 = 9$, $\sigma_2^2 = 16$, and covariance $\sigma_{12} = 8$.[5] The population correlation between X_1 and X_2 is therefore $\rho_{12} = 8/\sqrt{9 \times 16} = 2/3$; the slope for the regression of X_1 on X_2 is $\beta_{12} = 8/16 = 1/2$, and the slope for the regression of X_2 on X_1 is $\beta_{21} = 8/9 \approx 0.889$.

[5]The bivariate-normal distribution is described in online Appendix D on probability and estimation.

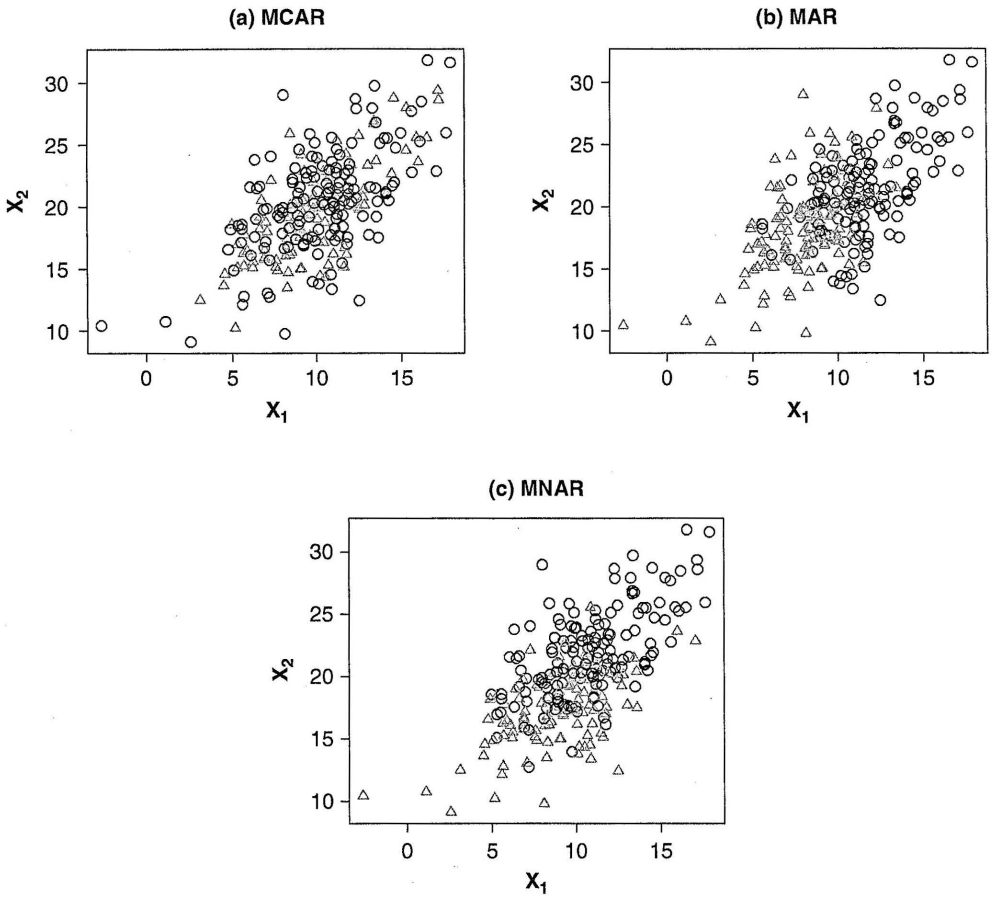

Figure 20.1 The 250 observations in each scatterplot were sampled from a bivariate-normal distribution; in each case, the observations shown as gray triangles have missing data on X_2. In panel (a), the 100 observations with missing data were sampled at random, and the missing data on X_2 are therefore missing completely at random (MCAR). In (b), the probability that an observation has a missing value on X_2 is related to its value on X_1, and so the missing data on X_2 are missing at random (MAR). In (c), the probability that an observation has a missing value on X_2 is related to its value on X_2, and so the missing data on X_2 are missing not at random (MNAR).

Consider the following three mechanisms for generating missing data in the sample of 250 observations:

1. One hundred of the observations on X_2 are selected at random and set to missing. This situation is illustrated in Figure 20.1(a), where the data points represented by black circles are fully observed, and those represented by gray triangles are missing X_2. Here, the missing values of X_2 are MCAR, and the subset of valid observations is a simple random sample of the full data set.

2. In Figure 20.1(b), an observation's missingness on X_2 is related to its (observed) value of X_1:

$$\Pr(X_{i2} \text{ is missing}) = \frac{1}{1 + \exp\left[\frac{1}{2} + \frac{2}{3}(X_{i1} - 10)\right]} \tag{20.1}$$

We recognize Equation 20.1 as a logistic-regression equation,[6] with the probability that X_2 is missing declining as X_1 grows larger. The regression coefficients were calibrated so that approximately 100 observations will have missing data on X_2 (and for the sample in Figure 20.1(b), there are, as it turned out, 109 missing values produced by simulating the missing-data-generating process). X_1 and X_2 are positively correlated, and consequently, there are relatively fewer small values of X_2 in the observed data than in the complete data; moreover, if we look only at the observations with valid data on *both* X_1 *and* X_2, this subset of observations also has relatively few small values of X_1. Because X_1, recall, is fully observed, the missing data on X_2 are MAR.

3. In Figure 20.1(c), an observation's missingness on X_2 is related to the (potentially unobserved) value of X_2 itself:

$$\Pr(X_{i2} \text{ is missing}) = \frac{1}{1 + \exp\left[\frac{1}{2} + \frac{1}{2}(X_{i2} - 20)\right]} \tag{20.2}$$

For our data set, the simulation of this process produced exactly 100 observations with missing data on X_2. Here, too, and indeed more directly, there are relatively few small values of X_2 (and, incidentally, if we exclude the observations with missing data on X_2, of X_1 also). Because missingness on X_2 depends directly on the value of X_2, the missing data are MNAR.

As mentioned, except in those relatively rare instances where missing data are built into the design of a study, it is not possible to verify from the data whether they are MCAR or even MAR—that is, whether the missing-data mechanism is ignorable. Indeed, it is fair to say that missing data are almost always MNAR. Nevertheless, if we can argue plausibly that the departure from MAR is likely small, then dealing with missing data becomes a much more tractable problem. Furthermore, unless we are willing to discard the data, we have to proceed in *some* manner. Rather than requiring perfection, which is probably unattainable, we may have to settle for a solution that simply gets us closer to the truth.

20.2 Traditional Approaches to Missing Data

In evaluating missing-data methods, there are three general questions to answer:

1. Does the method provide *consistent estimates* of population parameters, or does it introduce systematic biases into the results?
2. Does the method provide *valid statistical inferences*, or are confidence intervals and p-values distorted?

[6] See Section 14.1.

3. Does the method use the observed data *efficiently* or does it profligately discard information?

The answers to these questions depend partly on the methods themselves, partly on the nature of the process generating missing data, and partly on the statistics of interest.

There are many ad hoc methods that have been proposed for dealing with missing data; I will briefly describe several of the most common here and will explain why, in which respects, and under what circumstances they are problematic. This discussion is far from complete, however: For example, I have omitted discussion of methods based on reweighting the data.[7]

Complete-case analysis (also called *listwise* or *casewise deletion* of missing data), probably the most widely used approach, simply ignores observations with any missing data on the variables included in the analysis. Complete-case analysis has its advantages: It is simple to implement, provides consistent estimates and valid inferences when the data are missing completely at random, and provides consistent estimates of regression coefficients and valid inferences when missingness on all the variables in a regression does not depend on the response variable (even if data are not MCAR). Because it discards some valid data, however, complete-case analysis generally does not use the information in the data efficiently. This problem can become acute when there are many variables, each with some missing data. For example, suppose each of 10 variables is missing 5% of observations and that missingness in different variables is independent.[8] Then, we would expect only $100 \times .95^{10} \approx 60\%$ of the observations to be completely observed. Furthermore, when data are MAR or MNAR, complete-case analysis usually provides biased results and invalid inferences.

Available-case analysis (also called *pairwise deletion* of missing data) uses all nonmissing observations to compute each statistic of interest. In a least-squares regression analysis, for example, the regression coefficients can be calculated from the means, variances, and covariances of the variables (or, equivalently, from their means, variances, and correlations). To apply available-case analysis to least-squares regression, each mean and variance is calculated from all observations with valid data for a variable and the covariance of two variables from all observations that have valid data for both.[9] Available case analysis appears to use more information than complete-case analysis, but in certain instances, this is an illusion: That is, estimators based on available cases can be *less* efficient than those based on complete cases.[10] Moreover, by basing different statistics on different subsets of the data, available-case analysis can lead to nonsensical results, such as covariances that are inconsistent with one another or correlations outside the range from -1 to $+1$.[11] Finally, except in simple applications, such as linear least-squares regression, it is not obvious how to apply the available-case approach.

---

[7] As a general matter, relatively simple weighting schemes can reduce bias in estimates but do not provide valid inferences. See, for example, Little and Rubin (2002, Section 3.3).

[8] This is not a generally realistic condition: Missingness on different variables is probably positively associated, producing a result not quite as dismal as the one described here. The general point is valid, however: With many variables subject to missing data, there are typically many fewer complete cases than valid observations on individual variables.

[9] This description is slightly ambiguous: In computing the covariance, for example, do we use the means for each variable computed from all valid data for that variable or (as is more common and as I have done in the example reported below) recompute the means for each pair using observations with valid data for both variables in the pair?

[10] An example is estimating the difference between the means of two highly correlated variables (as in a paired t-test): See Little and Rubin (1990, pp. 378–380).

[11] See Exercise 20.1.

Several methods attempt to fill in missing data, replacing missing values with plausible *imputed* values. The resulting completed data set is then analyzed using standard methods. One such approach, termed *unconditional mean imputation* (or *mean substitution*) replaces each missing value with the mean of the observed data for the variable in question. Although mean imputation preserves the means of variables, it makes their distributions less variable and tends to weaken relationships between variables. One consequence is that mean imputation generally yields biased regression coefficients and invalid inferences even when data are MCAR. In addition, by treating the missing data as if they were observed, mean imputation exaggerates the effective size of the data set, further distorting statistical inference—a deficiency that it shares with other simple imputation methods.

A more sophisticated approach, called *conditional-mean imputation*, replaces missing data with predicted values obtained, for example, from a regression equation (in which case the method is also called *regression imputation*). Using available data, we regress each variable with missing data on other variables in the data set; the resulting regression equation is used to produce predicted values that replace the missing data.[12] A problem with regression imputation is that the imputed observations tend to be less variable than real data because they lack residual variation; another problem is that we have failed to account for uncertainty in the estimation of the regression coefficients used to obtain the imputed values. The first of these problems can be addressed, for example, by adding a randomly sampled residual to each filled-in value. The second problem leads naturally to Bayesian multiple imputation of missing values, described below.[13] Regression imputation improves on unconditional mean imputation, but it is far from a perfect technique, generally providing biased estimates and invalid inferences even for missing data that are MCAR.

I applied several methods of handling missing data to the artificial data sets graphed in Figure 20.1 (page 608) and described in the preceding section. The results are shown in Table 20.1. Recall that the data for this example were sampled from a bivariate-normal distribution (with parameters shown at the top of the table). Statistics for the complete data set of $n = 250$ observations are also shown (near the top of the table). Some of the results—for example, the equivalence of complete-case analysis, available-case analysis, and mean imputation for the slope coefficient B_{12} of the regression of X_1 (the completely observed variable) on X_2—are peculiar to univariate missing data.[14] Other characteristics are more general, such as the reasonable results produced by complete-case analysis when missingness does not depend on the response variable (i.e., for the coefficient β_{12} when data are MCAR or, for this example, MNAR, and for the coefficient β_{21} when, again for this example, data are MAR). Note that ML estimation and multiple imputation are the only methods that provide uniformly good results for *all* parameters in *both* the MCAR and MAR data sets.

To illustrate further the properties of the various missing-data methods, I conducted a small simulation study, drawing 1000 samples from the bivariate-normal distribution described above, producing from each sample a data set in which missing data were MAR, and applying complete-case analysis, unconditional-mean imputation, regression imputation, and Bayesian

[12]Because the predictor variables in each of these auxiliary regressions may themselves have missing data, the implementation of regression imputation can be complicated, requiring us to fit different regression equations for different patterns of missing information. The basic idea, however, is straightforward.

[13]See Section 20.4.

[14]See Exercise 20.2.

Table 20.1 Parameter Estimates Obtained by Several Methods of Handling Missing Data Under Different Conditions

	μ_1	μ_2	σ_1^2	σ_2^2	σ_{12}	ρ_{12}	β_{12}	β_{21}
Parameter	10.000	20.000	9.000	16.000	8.000	.667	0.500	0.889
Complete data (n = 250)								
Estimates	10.002	19.976	9.432	16.731	8.114	.646	0.485	0.860
MCAR data set								
Complete cases	10.210	20.400	9.768	17.114	7.673	.593	0.448	0.785
Available cases	10.002	20.400	9.432	17.114	7.673	.604	0.448	0.813
Mean imputation	10.002	20.400	9.432	10.241	4.591	.467	0.448	0.487
Regression imputation	10.002	20.237	9.432	12.454	7.409	.683	0.595	0.785
Maximum likelihood	10.002	20.237	9.394	16.809	7.379	.587	0.439	0.785
Multiple imputation	10.002	20.269	9.432	16.754	7.415	.590	0.443	0.786
MAR data set								
Complete cases	11.615	21.349	6.291	14.247	5.456	.576	0.383	0.867
Available cases	10.002	21.349	9.432	14.247	5.456	.508	0.383	0.578
Mean imputation	10.002	21.385	9.432	8.010	3.068	.353	0.383	0.325
Regression imputation	10.002	19.950	9.432	12.443	8.179	.755	0.657	0.867
Maximum likelihood	10.002	20.000	9.394	17.044	8.103	.640	0.475	0.863
Multiple imputation	10.002	19.914	9.432	17.493	8.342	.649	0.477	0.884
MNAR data set								
Complete cases	10.811	21.833	8.238	12.823	6.389	.622	0.498	0.776
Available cases	10.002	21.833	9.432	12.823	6.389	.581	0.498	0.677
Mean imputation	10.002	21.833	9.432	7.673	3.823	.449	0.498	0.405
Regression imputation	10.002	21.206	9.432	10.381	7.315	.739	0.705	0.776
Maximum likelihood	10.002	17.891	9.394	9.840	5.421	.564	0.551	0.577
Multiple imputation	10.002	21.257	9.432	13.167	7.154	.642	0.543	0.758

NOTES: The data were sampled from a bivariate-normal distribution with means, variances, and covariance as shown. The ML and multiple-imputation methods are described later in the chapter.

multiple imputation to each data set. The results are given in Table 20.2.[15] To simplify the table, I have not shown results for available-case analysis or for ML estimation (which produces results similar to those for multiple imputation). In addition, I have focused on the means and regression coefficients, which are the parameters that are usually of most direct interest.

Table 20.2 shows not only the average parameter estimates for each method (in the upper panel), which are useful for assessing bias, but also the RMSE of each estimator (i.e., the square root of the mean-square error, expressing the efficiency of the estimator), as well as (in the lower panel) the coverage and average interval width of nominally 95% confidence intervals for each method. If a confidence interval is valid, then the coverage should be close to .95. The results generally support the observations that I made above, and in particular, the only method that does uniformly well for all parameters—producing unbiased estimates, valid confidence intervals, and relatively efficient estimates—is multiple imputation.

[15]Similar but more extensive simulations appear in Schafer and Graham (2002). Also see Exercise 20.3.

Table 20.2 Mean Parameter Estimates and Confidence Interval Coverage for a Simulation Experiment With Data Missing at Random (MAR)

Parameter	*Complete Cases*	*Mean Imputation*	*Regression Imputation*	*Multiple Imputation*
Mean parameter estimate (RMSE)				
$\mu_1 = 10$	11.476	10.001	10.001	10.001
	(1.489)	(0.189)	(0.189)	(0.189)
$\mu_2 = 20$	21.222	21.322	20.008	20.008
	(1.355)	(1.355)	(0.326)	(0.344)
$\beta_{12} = 0.5$	0.391	0.391	0.645	0.498
	(0.117)	(0.117)	(0.151)	(0.041)
$\beta_{21} = 0.889$	0.891	0.353	0.891	0.890
	(0.100)	(0.538)	(0.100)	(0.106)
Confidence-interval coverage (mean interval width)				
μ_1	0	.951	.951	.951
	(0.792)	(0.750)	(0.750)	(0.746)
μ_2	.005	0	.823	.947
	(1.194)	(0.711)	(0.881)	(1.451)
β_{12}	.304	.629	.037	.955
	(0.174)	(0.246)	(0.140)	(0.175)
β_{21}	.953	0	.661	.939
	(0.396)	(0.220)	(0.191)	(0.463)

NOTES: The root-mean-square error (RMSE) of the parameter estimates is shown in parentheses below the mean estimates; the mean width of the confidence intervals is shown in parentheses below the coverage. Confidence intervals were constructed at a nominal level of .95.

Traditional methods of handling missing data include complete-case analysis, available-case analysis, and unconditional and conditional mean imputation. Complete-case analysis produces consistent estimates and valid statistical inferences when data are MCAR (and in certain other special circumstances), but even in this advantageous situation, it does not use information in the sample efficiently. The other traditional methods suffer from more serious problems.

20.3 Maximum-Likelihood Estimation for Data Missing at Random*

The method of maximum likelihood can be applied to parameter estimation in the presence of missing data. Doing so requires making assumptions about the distribution of the complete data

and about the process producing missing data. If the assumptions hold, then the resulting ML estimates have their usual optimal properties, such as consistency and asymptotic efficiency.[16]

Let $p(\mathbf{X};\boldsymbol{\theta}) = p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}};\boldsymbol{\theta})$ represent the joint probability density for the complete data $\mathbf{X}$, which as before, is composed of observed and missing components denoted, respectively, as $\mathbf{X}_{\text{obs}}$ and $\mathbf{X}_{\text{mis}}$. The vector $\boldsymbol{\theta}$ contains the unknown parameters on which the complete-data distribution depends. For example, if the variables in $\mathbf{X}$ are multivariately normally distributed (a case that I will examine presently), then $\boldsymbol{\theta}$ includes the population means and covariances among the variables.

In a seminal paper on statistical methods for missing data—the same paper in which he introduced distinctions among data that are MCAR, MAR, and MNAR—Rubin (1976) showed that the ML estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained from the marginal distribution of the observed data, *if missing data are missing at random*. In the general case that I am considering here, we can find the marginal distribution for the observed data by integrating over the missing data, producing

$$p(\mathbf{X}_{\text{obs}};\boldsymbol{\theta}) = \int p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}};\boldsymbol{\theta}) d\mathbf{X}_{\text{mis}}$$

Although it may be difficult to apply this result directly, simplification is possible in certain cases. Moreover, as I will explain shortly, it is, as a practical matter, possible to find $\widehat{\boldsymbol{\theta}}$ in the general case by iterative techniques.[17] As usual, the likelihood function $L(\boldsymbol{\theta};\mathbf{X}_{\text{obs}})$ is the same as the probability density function for the data but treats the observed data as fixed and the unknown parameters as variable. Once we have found the ML parameter estimates $\widehat{\boldsymbol{\theta}}$, we can proceed with statistical inference in the usual manner; for example, we can compute likelihood-ratio tests of nested models and construct Wald tests or confidence intervals for the elements of $\boldsymbol{\theta}$ based on estimated asymptotic variances for $\widehat{\boldsymbol{\theta}}$ obtained from the inverse of the observed information matrix

$$\mathcal{I}(\boldsymbol{\theta};\mathbf{X}_{\text{obs}}) = -\frac{\partial^2 \log_e L(\boldsymbol{\theta};\mathbf{X}_{\text{obs}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$$

Consider, for example, bivariately normally distributed variables X_1 and X_2; as in the previous section, X_1 is completely observed in a sample of n observations, but X_2 has $m < n$ observations missing at random, which for notational convenience, I will take as the first m observations.[18] Then, from the univariate-normal distribution,

$$p_1(x_{i1};\mu_1,\sigma_1^2) = \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left[-\frac{(x_{i1}-\mu_1)^2}{2\sigma_1^2}\right]$$

is the marginal probability density for observation i on variable X_1, and from the bivariate-normal distribution,

$$p_{12}(x_{i1},x_{i2};\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\sigma_{12}) = \frac{1}{2\pi\sqrt{\det\boldsymbol{\Sigma}}}\exp\left[-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right] \tag{20.3}$$

[16]For a general introduction to the method of maximum likelihood, see online Appendix D on probability and estimation.

[17]See Section 20.3.1 on the expectation-maximization (EM) algorithm.

[18]This is the univariate pattern of missing data employed in the examples of the preceding sections.

is the joint probability density for observation i on variables X_1 and X_2. In Equation 20.3, $\mathbf{x}_i \equiv (x_{i1}, x_{i2})'$ is a vector giving a pair of values for X_{i1} and X_{i2}, $\boldsymbol{\mu} \equiv (\mu_1, \mu_2)'$ is the vector of means for the two variables, and

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

is their covariance matrix. Using results in Little and Rubin (1990, pp. 382–383; 2002, chap. 7), the log-likelihood for the observed data is

$$\begin{aligned} \log_e L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12}) = &\sum_{i=1}^{m} \log_e p_1(x_{i1}; \mu_1, \sigma_1^2) \\ &+ \sum_{i=m+1}^{n} \log_e p_{12}(x_{i1}, x_{i2}; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12}) \end{aligned} \tag{20.4}$$

The log-likelihood in Equation 20.4 can easily be maximized numerically, but there is also a simple analytic solution. The statistics

$$\begin{aligned} \overline{X}_1^* &\equiv \frac{\sum_{i=m+1}^{n} X_{i1}}{n-m} \\ \overline{X}_2^* &\equiv \frac{\sum_{i=m+1}^{n} X_{i2}}{n-m} \\ S_1^{2*} &\equiv \frac{\sum_{i=m+1}^{n} (X_{i1} - \overline{X}_1^*)^2}{n-m} \\ S_2^{2*} &\equiv \frac{\sum_{i=m+1}^{n} (X_{i2} - \overline{X}_2^*)^2}{n-m} \\ S_{12}^* &\equiv \frac{\sum_{i=m+1}^{n} (X_{i1} - \overline{X}_1^*)(X_{i2} - \overline{X}_2^*)}{n-m} \end{aligned} \tag{20.5}$$

are the means, variances, and covariance for the two variables computed from the $n - m$ complete cases, and

$$\begin{aligned} \overline{X}_1 &\equiv \frac{\sum_{i=1}^{n} X_{i1}}{n} \\ S_1^2 &\equiv \frac{\sum_{i=1}^{n} (X_{i1} - \overline{X}_1)^2}{n} \end{aligned}$$

are the mean and variance of X_1 computed from all n available cases.[19] The ML estimators of the parameters of the bivariate-normal model are

[19]Note that the denominators for the variances and covariance are the number of observations, $n - m$ or n, rather than degrees of freedom $n - m - 1$ or $n - 1$. Recall that ML estimators of variance are biased but consistent. (See online Appendix D on probability and estimation.)

$$
\begin{aligned}
\widehat{\mu}_1 &= \overline{X}_1 \\
\widehat{\mu}_2 &= \overline{X}_2^* + \frac{S_{12}^*}{S_1^{2*}}\left(\overline{X}_1 - \overline{X}_1^*\right) \\
\widehat{\sigma}_1^2 &= S_1^2 \\
\widehat{\sigma}_2^2 &= S_2^{2*} + \left(\frac{S_{12}^*}{S_1^{2*}}\right)^2 \left(S_1^2 - S_1^{2*}\right) \\
\widehat{\sigma}_{12} &= S_{12}^* \left(\frac{S_1^2}{S_1^{2*}}\right)
\end{aligned}
\tag{20.6}
$$

Thus, the ML estimates combine information from the complete-case and available-case statistics.[20]

The method of ML can be applied to parameter estimation in the presence of missing data. If the assumptions made concerning the distribution of the complete data and the process generating missing data hold, then ML estimates have their usual optimal properties, such as consistency and asymptotic efficiency. When data are MAR, the ML estimate $\widehat{\boldsymbol{\theta}}$ of the parameters $\boldsymbol{\theta}$ of the complete-data distribution can be obtained from the marginal distribution of the observed data, integrating over the missing data:

$$p(\mathbf{X}_{\text{obs}}; \boldsymbol{\theta}) = \int p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}; \boldsymbol{\theta}) d\mathbf{X}_{\text{mis}}$$

Although it may be difficult to apply this result directly, simplification is possible in certain cases. Once we have found the ML parameter estimates, we can proceed with statistical inference in the usual manner, for example, computing likelihood-ratio tests of nested models and constructing Wald tests or confidence intervals.

20.3.1 The EM Algorithm

Arbitrary patterns of missing data do not yield simple expressions for the log-likelihood (such as in Equation 20.4 on page 615 for a univariate missing-data pattern in bivariate-normal data) no closed-form equations for the ML estimates (such as in Equation 20.6). The *expectation-maximization* (*EM*) algorithm, due to Dempster, Laird, and Rubin (1977), is a general iterative method for finding ML estimates in the presence of arbitrary patterns of missing data. Although the EM algorithm is broadly applicable, generally easy to implement, and effective, it has the disadvantage that it does not produce the information matrix and therefore does not yield standard errors for the estimated parameters. The version of the EM algorithm that I will describe here is for ignorable missing data (and is adapted from Little & Rubin, 2002, chaps. 8 and 11). The algorithm can also be applied to problems for which data are MNAR and hence are nonignorable.[21]

[20] See Exercise 20.4 for further interpretation of the ML estimators in Equation 20.6.

[21] See, for example, Little and Rubin (2002, chap. 15).

As before, let $\mathbf{X}$ represent the complete data, composed of the observed data $\mathbf{X}_{\text{obs}}$ and the missing data $\mathbf{X}_{\text{mis}}$. The likelihood based on the complete data is $L(\boldsymbol{\theta}; \mathbf{X})$, where recall, $\boldsymbol{\theta}$ contains the parameters for the distribution of $\mathbf{X}$. Let $\boldsymbol{\theta}^{(l)}$ represent the parameter estimates at the lth iteration of the EM algorithm. Starting values $\boldsymbol{\theta}^{(0)}$ may be obtained from the complete cases, for example. Each iteration of the EM algorithm comprises two steps: an E (*expectation*) *step* and an M (*maximization*) *step*. Hence the name "EM."

- In the E step, we find the expectation of the complete-data log-likelihood, integrating over the missing data, given the observed data and the current estimates of the parameters:

$$E\left[\log_e L(\boldsymbol{\theta}; \mathbf{X}) | \boldsymbol{\theta}^{(l)}\right] = \int \log_e L(\boldsymbol{\theta}; \mathbf{X}) p\left(\mathbf{X}_{\text{mis}} | \mathbf{X}_{\text{obs}}, \boldsymbol{\theta}^{(l)}\right) d\mathbf{X}_{\text{mis}}$$

- In the M step, we find the values $\boldsymbol{\theta}^{(l+1)}$ of $\boldsymbol{\theta}$ that maximize the expected log-likelihood $E\left[\log_e L(\boldsymbol{\theta}; \mathbf{X}) | \boldsymbol{\theta}^{(l)}\right]$; these are the parameter estimates for the next iteration.

When the parameter values stop changing from one iteration to the next (to an acceptable tolerance), they converge to the ML estimates $\widehat{\boldsymbol{\theta}}$.

Suppose, for example, that the complete data $\mathbf{X}$, consisting of n observations on p variables, is multivariately normally distributed, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The sums and sums of squares and cross-products of the variables are a set of sufficient statistics for these parameters:

$$T_j \equiv \sum_{i=1}^{n} X_{ij} \text{ for } j = 1, \ldots, p$$

$$T_{jj'} \equiv \sum_{i=1}^{n} X_{ij} X_{ij'} \text{ for } j, j' = 1, \ldots, p$$

Had we access to the complete data, then the ML estimates of the parameters could be computed from the sufficient statistics:

$$\widehat{\mu}_j = \frac{T_j}{n}$$

$$\widehat{\sigma}_{jj'} = \frac{T_{jj'}}{n} - \widehat{\mu}_j \widehat{\mu}_{j'}$$

(where the estimated variance of X_j is $\widehat{\sigma}_j^2 = \widehat{\sigma}_{jj}$).

Now, imagine that some of the data in $\mathbf{X}$ are MAR but in an arbitrary pattern. Then, in the E step, we find expected sums and sums of products by filling in the missing data with their conditional expected values, given the observed data and current estimates of the parameters. That is,

$$E\left(T_j | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l-1)}, \boldsymbol{\Sigma}^{(l-1)}\right) = \sum_{i=1}^{n} X_{ij}^{(l)}$$

$$E\left(T_{jj'} | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l-1)}, \boldsymbol{\Sigma}^{(l-1)}\right) = \sum_{i=1}^{n} \left(X_{ij}^{(l)} X_{ij'}^{(l)} + C_{ijj'}^{(l)}\right)$$

where

$$X_{ij}^{(l)} = \begin{cases} X_{ij} & \text{if } X_{ij} \text{ is observed} \\ E\left(X_{ij}|\mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l-1)}, \boldsymbol{\Sigma}^{(l-1)}\right) & \text{if } X_{ij} \text{ is missing} \end{cases}$$

and

$$C_{ijj'}^{(l)} = \begin{cases} 0 & \text{if } \textit{either } X_{ij} \textit{ or } X_{ij'} \text{ is observed} \\ C\left(X_{ij}, X_{ij'}|\mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l-1)}, \boldsymbol{\Sigma}^{(l-1)}\right) & \text{if } \textit{both } X_{ij} \textit{ and } X_{ij'} \text{ are missing} \end{cases} \tag{20.7}$$

Finally, $E\left(X_{ij}|\mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)}\right)$ is obtained as the fitted value from the regression of X_j on the other Xs, using the current estimates $\boldsymbol{\mu}^{(l)}$ and $\boldsymbol{\Sigma}^{(l)}$ to obtain the regression coefficients, and $C\left(X_{ij}, X_{ij'}|\mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)}\right)$ is the covariance of the fitted values for X_{ij} and $X_{ij'}$ obtained from the multivariate regression of X_j and $X_{j'}$ on the other Xs, again at current values of the parameters.[22]

Once we have the expected sums and sums of cross-products, the M step is straightforward:

$$\mu_j^{(l)} = \frac{\sum_{i=1}^n X_{ij}^{(l)}}{n}$$

$$\sigma_{jj'}^{(l)} = \frac{\sum_{i=1}^n \left(X_{ij}^{(l)} X_{ij'}^{(l)} + C_{ijj'}^{(l)}\right)}{n} - \mu_j^{(l)} \mu_{j'}^{(l)}$$

Consider the comparatively simple case of bivariate-normal data where the variable X_1 is completely observed and the first m of n observations on X_2 are missing. Take as starting values the means, variances, and covariance computed from the $n - m$ complete cases (given in Equation 20.5 on page 615). Then, because X_1 is completely observed,

$$\begin{aligned} E\left(T_1|\mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}\right) &= \sum_{i=1}^n X_{i1} \\ E\left(T_{11}|\mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}\right) &= \sum_{i=1}^n X_{i1}^2 \end{aligned} \tag{20.8}$$

and, for sums involving X_2, which has m missing values,

$$\begin{aligned} E\left(T_2|\mathbf{X}_{\text{obs}}, \mu^{(0)}, \Sigma^{(0)}\right) &= \sum_{i=1}^m \widehat{X}_{i2} + \sum_{i=m+1}^n X_{i2} \\ E\left(T_{22}|\mathbf{X}_{\text{obs}}, \mu^{(0)}, \Sigma^{(0)}\right) &= \sum_{i=1}^m \left(\widehat{X}_{i2}^2 + S_{2|1}^{2(0)}\right) + \sum_{i=m+1}^n X_{i2}^2 \\ E\left(T_{12}|\mathbf{X}_{\text{obs}}, \mu^{(0)}, \Sigma^{(0)}\right) &= \sum_{i=1}^m \left(X_{i1}\widehat{X}_{i2}^2\right) + \sum_{i=m+1}^n X_{i1}X_{i2} \end{aligned} \tag{20.9}$$

where $\widehat{X}_{i2}$ is the fitted value from the complete-case regression of X_2 on X_1, and $S_{2|1}^{2(0)}$ is the residual variance from this regression. The M-step estimates computed from these expectations are just the ML estimates previously given in Equation 20.6.[23] That is, in the simple case of monotone missing data, the EM algorithm converges to the ML estimates in a single iteration.

[22] In multivariate regression, there is more than one response variable. In the current context, the role of the response variables is played by X_j and $X_{j'}$. See Section 9.5 and Exercise 20.5.

[23] See Exercise 20.6.

The EM algorithm is a general iterative procedure for finding ML estimates—but not their standard errors—in the presence of arbitrary patterns of missing data. When data are MAR, iteration l of the EM algorithm consists of two steps: (1) In the E (expectation) step, we find the expectation of the complete-data log-likelihood, integrating over the missing data, given the observed data and the current estimates of the parameters:

$$E\left[\log_e L(\boldsymbol{\theta}; \mathbf{X}) | \boldsymbol{\theta}^{(l)}\right] = \int \log_e L(\boldsymbol{\theta}; \mathbf{X}) p\left(\mathbf{X}_{\text{mis}} | \mathbf{X}_{\text{obs}}, \boldsymbol{\theta}^{(l)}\right) d\mathbf{X}_{\text{mis}}$$

(2) In the M (maximization) step, we find the values $\boldsymbol{\theta}^{(l+1)}$ of $\boldsymbol{\theta}$ that maximize the expected log-likelihood $E\left[\log_e L(\boldsymbol{\theta}; \mathbf{X}) | \boldsymbol{\theta}^{(l)}\right]$; these are the parameter estimates for the next iteration. At convergence, the EM algorithm produces the ML estimates $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

20.4 Bayesian Multiple Imputation

Bayesian multiple imputation (abbreviated as *MI*) is a flexible and general method for dealing with missing data that are MAR. Like ML estimation, multiple imputation begins with a specification of the distribution of the complete data (assumed to be known except for a set of parameters to be estimated from the data).

The essential idea of multiple imputation is to reflect the uncertainty associated with missing data by imputing *several* values for each missing value, each imputed value drawn from the *predictive distribution* of the missing data and, therefore, producing not one but several completed data sets. Standard methods of statistical analysis are then applied in parallel to the completed data sets. Parameters of interest are estimated along with their standard errors for each imputed data set. Estimated parameters are then averaged across completed data sets; standard errors are also combined across imputed data sets, taking into account the variation among the estimates in the several data sets, thereby capturing the added uncertainty due to having to impute the missing data.

A multivariate-normal model for the complete data is both relatively simple and useful in applications. Indeed, because the model assumed to describe the complete data is used just to obtain imputed values for the missing data, it turns out that the method of multiple imputation is usually not terribly sensitive to the assumption of multivariate normality.[24]

Suppose that X_1 and X_2 are bivariately normally distributed and that, as in previously developed examples, there is a univariate pattern of missing data, with X_1 completely observed and m of the n observations on X_2 MAR. For convenience, and again as before, let us order the data so that the missing observations on X_2 are the first m observations. Let $A^*_{2|1}$ and $B^*_{2|1}$ represent the intercept and slope for the complete-case least-squares regression of X_2 on X_1.[25] In regression imputation, recall, we replace the missing values with the fitted values

[24]See, for example, Schafer (1997, chap. 5). As described in Section 20.4.3, however, there are some pitfalls to be avoided.

[25]The results of the preceding section imply that $A^*_{2|1}$ and $B^*_{2|1}$ are the ML estimators of $\alpha_{2|1}$ and $\beta_{2|1}$. See Exercise 20.6.

$$\widehat{X}_{i2} = A^*_{2|1} + B^*_{2|1}X_{i1} \tag{20.10}$$

Recall as well that a defect of this procedure is that it ignores residual variation in X_2 conditional on X_1. A more sophisticated version of regression imputation adds a randomly generated residual to the fitted value, taking the imputed value as $\widehat{X}_{i2} + E_{i2|1}$, where $E_{i2|1}$ is drawn randomly from the normal distribution $N(0, S^{*2}_{2|1})$, and where

$$S^{*2}_{2|1} \equiv \frac{\sum_{i=m+1}^{n} (X_{i2} - \widehat{X}_{i2})^2}{n - m}$$

is the ML estimator of the residual variance of X_2 given X_1 (based on the $n - m$ complete cases).

There is still a problem, however: The fitted values and generated residuals on which the imputations are based fail to take into account the fact that the regression coefficients $A^*_{2|1}$ and $B^*_{2|1}$ and the residual variance $S^{*2}_{2|1}$ are themselves *estimates* that are subject to sampling variation. MI draws values of the regression parameters and the error variance—let us call these values $\widetilde{\alpha}_{2|1}$, $\widetilde{\beta}_{2|1}$, and $\widetilde{\sigma}^2_{2|1}$—from the *posterior distribution* of the parameters, typically assuming a *noninformative prior distribution*.[26]

As Little and Rubin (1990, pp. 386–387) explain, we may proceed as follows:

1. Given a random draw Z^2 from the chi-square distribution with $n - m - 2$ degrees of freedom, find

$$\widehat{\sigma}^2_{2|1} \equiv \frac{\sum_{i=m+1}^{n} (X_{i2} - \widehat{X}_{i2})^2}{Z^2}$$

2. With $\widetilde{\sigma}^2_{2|1}$ in hand, draw a random slope $\widetilde{\beta}_{2|1}$ from the normal distribution

$$N\left(B^*_{2|1}, \frac{\widetilde{\sigma}^2_{2|1}}{[(n-m)S_1^2]^2}\right)$$

   Here, $S_1^2 \equiv \sum_{i=1}^{n} (X_{i1} - \overline{X}_1)^2/n$ is the ML estimate of the variance of X_1, and $\overline{X}_1 \equiv \sum_{i=1}^{n} X_{i1}/n$ is the ML estimate of the mean of X_1, based on all n cases.

3. Using the previously obtained values of $\widetilde{\sigma}^2_{2|1}$ and $\widetilde{\beta}_{2|1}$, draw a random intercept $\widetilde{\alpha}_{2|1}$ from the normal distribution

$$N\left(\widehat{\mu}_2 - \widetilde{\beta}_{2|1}\overline{X}_1, \frac{\widetilde{\sigma}^2_{2|1}}{(n-m)^2}\right)$$

   where $\widehat{\mu}_2$ is the ML estimate of the mean of X_2 (given in Equation 20.6 on page 616).

4. Finally, replace the missing values in X_2 by

$$\widetilde{X}_{i2} \equiv \widehat{\alpha}_{2|1} + \widetilde{\beta}_{2|1}X_{i1} + \widetilde{E}_i$$

   where $\widetilde{E}_i$ is sampled from $N(0, \widetilde{\sigma}^2_{2|1})$.

[26]Think of the posterior distribution of the parameters as capturing our uncertainty about the values of the parameters. Basic concepts of Bayesian statistical inference, including the notions of prior and posterior distributions, are described in online Appendix D on probability and estimation.

In multiple imputation, this procedure is repeated g times, producing g completed data sets.

More generally, we have a complete data set $\mathbf{X}$ comprising n cases and p multivariately normally distributed variables; some of the entries of $\mathbf{X}$ are MAR in an arbitrary pattern. In this more general case, there is no fully adequate closed-form procedure for sampling from the predictive distribution of the data to impute missing values. Instead, simulation methods must be employed to obtain imputations. Two such methods are data augmentation (described in Schafer, 1997) and importance sampling (described in King, Honaker, Joseph, & Scheve, 2001). General descriptions of these methods are beyond the scope of this chapter.[27]

Raghunathan, Lepkowski, Van Hoewyk, and Solenberger (2001) and van Buuren and Oudshoorn (1999) (also see van Buuren, 2012) suggest a simpler approach that cycles iteratively through a set of regression equations for the variables containing missing data. The formal properties of this approach have not been established, although it appears to work well in practice.[28] Multiple imputation can be extended beyond the multivariate-normal distribution to other models for the complete data, such as the multinomial distribution for a set of categorical variables, and mixed multinomial-normal models for data sets containing both quantitative and categorical data.[29]

20.4.1 Inference for Individual Coefficients

Having obtained g completed data sets, imagine that we have analyzed the data sets in parallel, producing g sets of regression coefficients, $B_0^{(l)}, B_1^{(l)}, \ldots, B_k^{(l)}$ for $l = 1, \ldots, g$ (where, for notational convenience, I have represented the regression constant as B_0). We also find the coefficient standard errors, $\text{SE}(B_0^{(l)})$, $\text{SE}(B_1^{(l)}), \ldots,$ $\text{SE}(B_k^{(l)})$, computed in the usual manner for each completed data set. Rubin (1987) provides simple rules for combining information across multiple imputations of the missing data, rules that are valid as long as the sample size is sufficiently large for the separate estimates to be approximately normally distributed. The context here is quite general: The regression coefficients and their standard errors might be produced by linear least-squares regression, but they might also be produced by ML estimation of a logistic-regression model, by nonlinear least squares, or by *any* parametric method of regression analysis.

Point estimates of the population regression coefficients are obtained by averaging across imputations:

$$\widetilde{\beta}_j \equiv \frac{\sum_{l=1}^{g} B_j^{(l)}}{g} \tag{20.11}$$

The standard errors of the estimated coefficients are obtained by combining information about within- and between-imputation variation in the coefficients:

[27]Multiple imputation by data augmentation is implemented in Schafer's software, available for SAS, S-PLUS, R, and in stand-alone programs. Multiple imputation by importance sampling is implemented in King's software, available for R and in a stand-alone program.

[28]This approach is implemented in the IVEware (imputation and variance estimation) software for SAS, as a stand-alone program; in the MICE (multivariate imputation by chained equations) software for S-PLUS and R, as well as in a stand-alone program; and in the **mi** package for R (Su, Gelman, Hill, & Yajima, 2011). Access to convenient software for multiple imputation is important because the method is computationally intensive.

[29]See, for example, Schafer (1997, chaps. 7–9).

$$\widetilde{\mathrm{SE}}(\widetilde{\beta}_j) \equiv \sqrt{V_j^{(W)} + \frac{g+1}{g} V_j^{(B)}} \tag{20.12}$$

where the within-imputation component is

$$V_j^{(W)} \equiv \frac{\sum_{l=1}^{g} \mathrm{SE}^2\left(B_j^{(l)}\right)}{g}$$

and the between-imputation component is

$$V_j^{(B)} \equiv \frac{\sum_{l=1}^{g} \left(B_j^{(l)} - \widetilde{\beta}_j\right)^2}{g-1}$$

Inference based on $\widetilde{\beta}_j$ and $\widetilde{\mathrm{SE}}(\widetilde{\beta}_j)$ uses the t-distribution, with degrees of freedom determined by

$$df_j = (g-1)\left(1 + \frac{g}{g+1} \times \frac{V_j^{(W)}}{V_j^{(B)}}\right)^2$$

For example, to construct a 95% confidence interval for β_j,

$$\beta_j = \widetilde{\beta}_j \pm t_{.025,df_j} \widetilde{\mathrm{SE}}(\widetilde{\beta}_j)$$

Let γ_j denote the relative amount of information about the parameter β_j that is missing. This is not quite the same as the fraction of observations that are missing on the explanatory variable X_j because, unless X_j is uncorrelated with the other variables in the data set, there will be information in the data relevant to imputing the missing values and because data missing on one variable influence all the regression estimates. The *estimated rate of missing information* is

$$\widehat{\gamma}_j = \frac{R_j}{R_j + 1} \tag{20.13}$$

where

$$R_j \equiv \frac{g+1}{g} \times \frac{V_j^{(B)}}{V_j^{(W)}}$$

The efficiency of the multiple-imputation estimator relative to the maximally efficient ML estimator—that is, the ratio of sampling variances of the ML estimator to the MI estimator—is $\mathrm{RE}(\widetilde{\beta}_j) = g/(g+\gamma_j)$. If the number of imputations g is infinite, MI is therefore as efficient as ML, but even when the rate of missing information is quite high and the number of imputations modest, the relative efficiency of the MI estimator hardly suffers. Suppose, for example, that $\gamma_j = 0.5$ (a high rate of missing information) and that $g = 5$; then $\mathrm{RE}(\widetilde{\beta}_j) = 5/(5+0.5) = 0.91$. Expressed on the scale of the standard error of $\widetilde{\beta}_j$, which is proportional to the length of the confidence interval for β_j, we have $\sqrt{\mathrm{RE}(\widetilde{\beta}_j)} = 0.95$.[30]

[30] See Exercise 20.7.

Bayesian multiple imputation (MI) is a flexible and general method for dealing with data that are missing at random. The essential idea of multiple imputation is to reflect the uncertainty associated with missing data by imputing g values for each missing value, drawing each imputed value from the predictive distribution of the missing data (a process that usually requires simulation) and therefore producing not one but g completed data sets. Standard methods of statistical analysis are then applied in parallel to the completed data sets.

- According to Rubin's rules, MI estimates (e.g., of a population regression coefficient β_j) are obtained by averaging over the imputed data sets:

$$\widetilde{\beta}_j = \frac{\sum_{l=1}^{g} B_j^{(l)}}{g}$$

where $B_j^{(l)}$ is the estimate of β_j from imputed data set l.

- Standard errors of the estimated coefficients are obtained by combining information about within- and between-imputation variation in the coefficients,

$$\widetilde{\mathrm{SE}}(\widetilde{\beta}_j) = \sqrt{V_j^{(W)} + \frac{g+1}{g} V_j^{(B)}}$$

where the within-imputation component is

$$V_j^{(W)} = \frac{\sum_{l=1}^{g} \mathrm{SE}^2\left(B_j^{(l)}\right)}{g}$$

and the between-imputation component is

$$V_j^{(B)} = \frac{\sum_{l=1}^{g} \left(B_j^{(l)} - \widetilde{\beta}_j\right)^2}{g-1}$$

Here, $\mathrm{SE}\left(B_j^{(l)}\right)$ is the standard error of B_j computed in the usual manner for the lth imputed data set.

- Inference based on $\widetilde{\beta}_j$ and $\widetilde{\mathrm{SE}}(\widetilde{\beta}_j)$ uses the t-distribution, with degrees of freedom determined by

$$df_j = (g-1)\left(1 + \frac{g}{g+1} \times \frac{V_j^{(W)}}{V_j^{(B)}}\right)^2$$

Inference for several coefficients proceeds in a similar, if more complex, manner.

20.4.2 Inference for Several Coefficients*

The generalization of Rubin's rules to simultaneous tests or confidence regions for several coefficients entails some complications.[31] Suppose that we wish to test the hypothesis H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$, where $\underset{(s\times 1)}{\boldsymbol{\beta}_1}$ is a subset of $s > 1$ of the $k+1$ elements of the parameter vector $\boldsymbol{\beta}$; typically, this would be the hypothesis H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$. Were it not for the missing data, we could base the hypothesis test on the Wald chi-square statistic,

$$Z_0^2 = (\mathbf{b}_1 - \boldsymbol{\beta}_0)' \widehat{\mathcal{V}}^{-1}(\mathbf{b}_1)(\mathbf{b}_1 - \boldsymbol{\beta}_0)$$

where the vector $\mathbf{b}_1$ contains the estimated coefficients and $\widehat{\mathcal{V}}(\mathbf{b}_1)$ is the estimated asymptotic covariance matrix of $\mathbf{b}_1$.[32]

In the present context, we have estimates for several completed data sets in which the missing data have been imputed, and so we first average the estimates, obtaining

$$\widetilde{\boldsymbol{\beta}}_1 \equiv \frac{1}{g}\sum_{l=1}^{g} \mathbf{b}_1^{(l)}$$

Then we compute the between- and within-imputation components of the covariance matrix of these estimates:

$$\mathbf{V}^{(W)} \equiv \frac{1}{g}\sum_{l=1}^{g} \widehat{\mathcal{V}}\left(\mathbf{b}_1^{(g)}\right)$$

$$\mathbf{V}^{(B)} \equiv \frac{1}{g-1}\sum_{l=1}^{g}\left(\mathbf{b}_1^{(g)} - \widetilde{\boldsymbol{\beta}}_1\right)\left(\mathbf{b}_1^{(g)} - \widetilde{\boldsymbol{\beta}}_1\right)'$$

In analogy to the single-coefficient case, we could compute the total covariance matrix

$$\mathbf{V} \equiv \mathbf{V}^{(W)} + \frac{g+1}{g}\mathbf{V}^{(B)}$$

Basing a test on $\mathbf{V}$, however, turns out to be complicated.

Instead, simplification of the problem leads to the test statistic

$$F_0 \equiv \frac{(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_0)'\left(\mathbf{V}^{(W)}\right)^{-1}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_0)'}{s(1+R)}$$

where

$$R \equiv \frac{g+1}{g} \times \frac{\text{trace}\left[\mathbf{V}^{(B)}\left(\mathbf{V}^{(W)}\right)^{-1}\right]}{s}$$

The test statistic F_0 follows an approximate F-distribution, with s degrees of freedom in the numerator and denominator given by

[31] The results that I give here, and alternative procedures, are explained in greater detail in Rubin (1987, chaps. 3 and 4) and in Schafer (1997, Section 4.3.3).

[32] See, for example, the discussion of Wald tests for generalized linear models in Section 15.3.3.

$$df = \begin{cases} 4 + [s(g-1) - 4]\left[1 + \frac{1}{R} \times \frac{s(g-1) - 2}{s(g-1)}\right] & \text{when } s(g-1) > 4 \\ \frac{1}{2}(g-1)(s+1)\left(1 + \frac{1}{R}\right)^2 & \text{when } s(g-1) \leq 4 \end{cases}$$

20.4.3 Practical Considerations

Although the multivariate-normal model can prove remarkably useful in providing multiple imputations even when the data are not normally distributed, multiple imputation cannot preserve features of the data that are not represented in the imputation model. How essential it is to preserve particular features of the data depends on the statistical model used to analyze the multiply imputed data sets. It is therefore important in formulating an imputation model to ensure that the imputation model is consistent with the intended analysis. The following points should assist in this endeavor:

- *Try to include variables in the imputation model that make the assumption of ignorable missingness reasonable.* Think of imputation as a pure prediction problem, not as a statistical model subject to substantive interpretation. If we are able to do a good job of predicting missing values (and missingness), then the assumption that data are MAR is more credible. Finding variables that are highly correlated with a variable that has missing data, but for which data are available, therefore, will likely improve the quality of imputations, as will variables that are related to missingness. In particular, it is perfectly acceptable, and indeed desirable, to include variables in the imputation model that *are not used* in the subsequent statistical analysis, alongside the variables that are used in the data analysis.[33] There is also nothing wrong with using the variable that is ultimately to be treated as a response to help impute missing data in variables that are to be treated as explanatory variables. To reiterate, the model used for imputation is essentially a prediction model—not a model to be interpreted substantively.
- *If possible, transform variables to approximate normality.*[34] After the imputed data are obtained, the variables can be transformed back to their original scales, if desired, prior to analyzing the completed data sets.
- *Adjust the imputed data to resemble the original data.* For example, imputed values of an integer-valued variable can be rounded to the nearest integer. Ordinal variables can be handled by providing integer codes and then rounding the imputed values to integers. Occasional negative imputed values of a nonnegative variable can be set to 0. Imputed values of a 0/1 dummy variable can be set to 0 if less than or equal to 0.5 and to 1 if greater than 0.5. These steps may not be necessary to analyze the imputed data, but they should not hurt in any event.
- *Make sure that the imputation model captures relevant features of the data.* What is relevant depends on the use to which the imputed data will be put. For example, the multivariate-normal distribution ensures that regressions of one variable on others are linear

[33]See Collins, Schafer, and Kam (2001), who present evidence supporting what they term an *inclusive strategy* for formulating imputation models.

[34]The material in Section 4.2 on transformations for symmetry and in Section 4.6 on Box-Cox transformations for multivariate normality is particularly relevant here.

and additive. Using the multivariate-normal distribution for imputations, therefore, will not preserve *nonlinear* relationships and *interactions* among the variables, unless we make special provision for these features of the data.

Suppose, for example, that we are interested in modeling the potential interaction between gender and education in determining income. Because gender is likely completely observed, but there may well be missing data on both education and income, we could divide the data set into two parts based on gender, obtaining multiply imputed data sets separately for each part and combining them in our analysis of the completed data sets. This approach runs into problems, however, if we find it necessary to divide the data set into too many parts or if the categorical variable or variables used to partition the data are themselves not completely observed.

Allison (2002) suggests forming interaction regressors and polynomial regressors as part of the data set to which the imputation model is applied. The imputed interaction and polynomial regressors are then used in the analysis of the completed data sets. Although such variables are not normally distributed, there is some evidence that multiple imputation based on the multivariate-normal model nevertheless works well in these circumstances.

Although, as explained, the multivariate-normal model can be used to impute a dummy regressor for a dichotomous factor, it is not obvious how to proceed with a polytomous factor. Allison (2002) proposes the following procedure: For an m-category factor, select an arbitrary baseline category (say the last), and code $m-1$ dummy regressors,[35] including these dummy variables in the multiple-imputation process. From the imputed values for the ith observation in the lth imputation, $D_{i1}^{(l)}, D_{i2}^{(l)}, \ldots, D_{i,m-1}^{(l)}$, compute $D_{im}^{(l)} = 1 - \sum_{j=1}^{m-1} D_{ij}^{(l)}$. Assign the ith observation to the category $(1, 2, \ldots, m)$ for which $D_{ij}^{(l)}$ is largest.

> Multiple imputation based on the multivariate-normal distribution can be remarkably effective in practice, even when the data are not normally distributed. To apply multiple imputation effectively, however, it is important to include variables in the imputation model that make the assumption of ignorable missingness reasonable; to transform variables to approximate normality, if possible; to adjust the imputed data so that they resemble the original data; and to make sure that the imputation model captures features of the data, such as nonlinearities and interactions, to be used in subsequent data analysis.

20.4.4 Example: A Regression Model for Infant Mortality

Figure 20.2 (repeating Figure 3.14 from page 45) shows the relationship between infant mortality (number of infant deaths per 1000 live births) and gross domestic product per capita (in U.S. dollars) for 193 nations, part of a larger data set of 207 countries compiled by the United

[35] See Chapter 7 for a general discussion of dummy-variable regressors.

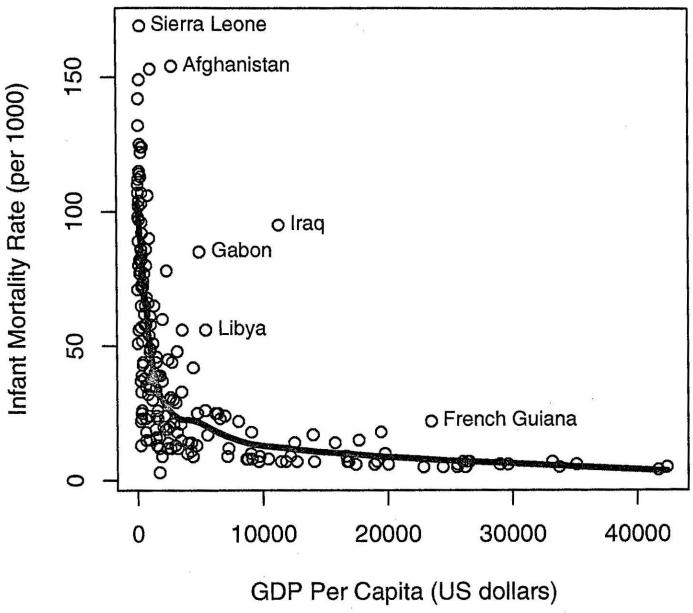

Figure 20.2 Scatterplot for infant mortality and GDP per capita for 193 nations. The line is for a lowess smooth with a span of 1/2. Several nations with high infant mortality for their levels of GDP are identified.

Nations. The amount of missing data in Figure 20.2 is therefore relatively small, comprising only about 7% of the cases.

Let us now consider the regression of infant mortality not only on GDP per capita but also on the percentage of married women practicing contraception and the average number of years of education for women. To linearize the regression, I log-transformed both infant mortality and GDP.[36] A complete-case analysis includes only 62 of the 207 countries and produces the results shown in the upper panel of Table 20.3.

The number of observations with missing data for each of the variables in the analysis is as follows:

Infant Mortality	GDP	Contraception	Female Education
6	10	63	131

There are, however, other variables in the full data set that are highly correlated with contraception and female education, such as the total fertility rate and the illiteracy rate for women. I decided to base imputations on a multivariate-normal model with the four variables in the regression plus the total fertility rate, the expectation of life for women, the percentage of women engaged in economic activity outside the home, and the illiteracy rate for women. Preliminary examination of the data suggested that the multivariate-normal model could be made more appropriate for the data by transforming several of these variables. In particular—as in the regression model in Table 20.3—I log-transformed infant mortality and GDP. I also took the square root of the total fertility rate; cubed female expectation of life, after subtracting

[36] See Exercise 20.8.

Table 20.3 Estimated Coefficients and Standard Errors for the Regression of Infant Mortality on GDP Per Capita, Percentage Using Contraception, and Average Female Education, for 207 Nations (62 Complete Cases)

	Intercept	$\log_e$ *GDP*	*Contraception*	*Female Education*
Complete-case analysis				
Coefficient, B_j	6.88	−0.294	−0.0113	−0.0770
SE(B_j)	(0.29)	(0.058)	(0.0042)	(0.0338)
Multiple-imputation analysis				
Coefficient, $\widetilde{\beta}_j$	6.57	−0.234	−0.00953	−0.105
$\widetilde{\text{SE}}(\widetilde{\beta}_j)$	(0.18)	(0.049)	(0.00294)	(0.033)
Missing Information, $\widehat{\gamma}_j$	0.20	0.61	0.41	0.69

Table 20.4 Means and Standard Deviations of Variables in the Infant Mortality Regression, Complete-Case, and Maximum-Likelihood Estimates

	$\log_e$ *Infant Mortality*	$\log_e$ *GDP*	*Contraception*	*Female Education*
		Estimates based on Complete Cases		
Mean	3.041	8.151	50.90	11.30
SD	(1.051)	(1.703)	(23.17)	(3.55)
		Maximum-Likelihood Estimates		
Mean	3.300	7.586	44.36	10.16
SD	(1.022)	(1.682)	(24.01)	(3.51)

a start of 35 from each value; and took the 1/4 power of female illiteracy. The resulting data set did not look quite multivariate-normal, but several of the variables were more symmetrically distributed than before.

To get a sense of the possible influence of missing data on conclusions drawn from the data, I computed the complete-case estimates of the means and standard deviations of the four variables to be used in the regression, along with ML estimates, obtained by the EM algorithm applied to the eight variables to be used in the imputation model. These results are given in Table 20.4. As one might expect, the means for the complete cases show lower average infant mortality, higher GDP per capita, higher rates of contraception, and a higher level of female education than the ML estimates assuming ignorable missing data; the two sets of standard deviations, however, are quite similar.

Using Schafer's data augmentation method and employing the multivariate-normal model, I obtained imputations for 10 completed data sets.[37] Then, applying Equations 20.11, 20.12, and

[37]Data augmentation employs a *Markov-chain Monte-Carlo* (*MCMC*) method to sample from the predictive distribution of the data. Using Schafer'stextbfnorm package for the R statistical computing environment for these computations, I set the number of steps for the data augmentation algorithm to 20. Technical aspects of the data augmentation algorithm are discussed in Schafer (1997) and, in less detail, in Allison (2002).

20.13 (on pages 621–622), I computed the estimated coefficients, standard errors, and estimated rate of missing information for each coefficient, shown in the lower panel of Table 20.3. With the exception of the female education coefficient, the standard errors from the multiple-imputation analysis are noticeably smaller than those from the complete-case analysis. In addition, the coefficients for GDP and female education differ between the two analyses by about one standard error; the coefficients for contraception, in contrast, are very similar. Finally, the rates of missing information for the three slope coefficients are all large. Because 10 imputations were employed, however, the square-root relative efficiency of the estimated coefficients based on the multiply imputed data is at worst $\sqrt{10/(10+0.69)} = 0.97$.

20.5 Selection Bias and Censoring

When missing data are not ignorable (i.e., MNAR), consistent estimation of regression models requires an explicit auxiliary model for the missingness mechanism. Accommodating nonignorable missing data is an intrinsically risky venture because the resulting regression estimates can be very sensitive to the specifics of the model assumed to generate the missing data.

This section introduces two models in wide use for data that are MNAR: Heckman's model to overcome selection bias in regression and the so-called tobit model (and related models) for a censored response variable in regression. Before examining these models, however, it is useful to develop some basic ideas concerning truncated- and censored-normal distributions.

20.5.1 Truncated- and Censored-Normal Distributions

The distinction between *truncation* and *censoring* is illustrated in Figure 20.3. In each case, there is an unobserved variable ξ that follows the standard-normal distribution, $N(0,1)$. The observed variable Z in panel (a) *truncates* this distribution *on the left* by suppressing all values of ξ below $\xi = -0.75$; that is, there are no observations below the truncation point. The density function $p(z)$ of Z still must enclose an area of 1, and so this density is given by

$$p(z) = \frac{\phi(z)}{1 - \Phi(-0.75)} = \frac{\phi(z)}{\Phi(0.75)} \text{ for } z \geq -0.75$$

where $\phi(\cdot)$ is the density function and $\Phi(\cdot)$ the cumulative distribution function of the standard-normal distribution. In panel (b), where the distribution of ξ is *left-censored* rather than truncated,

$$Z = \begin{cases} -0.75 & \text{for } \xi \leq -0.75 \\ \xi & \text{for } \xi > -0.75 \end{cases}$$

Consequently, $\Pr(Z = -0.75) = \Phi(-0.75)$, that is, the area to the left of -0.75 under the standard-normal density function $\phi(\cdot)$.

It will be useful to have expressions for the mean and variance of a truncated-normal distribution. Suppose now that ξ is normally distributed with an *arbitrary* mean μ and variance

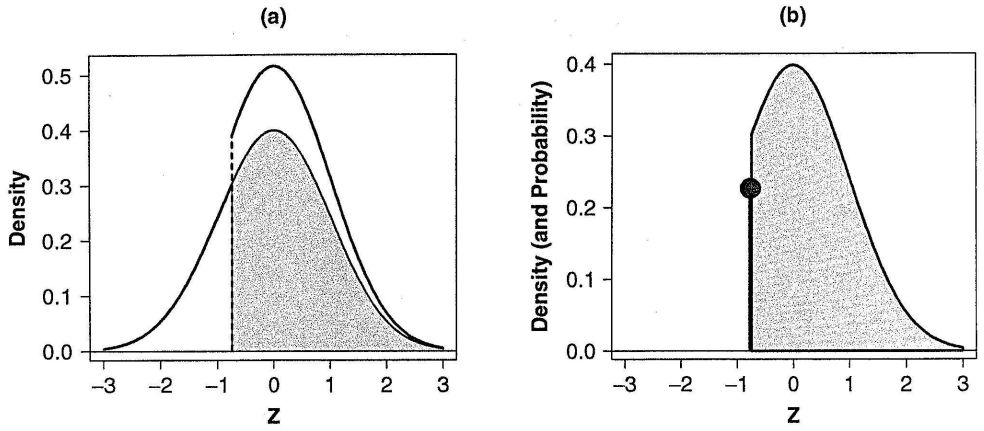

Figure 20.3 (a) Truncated- and (b) censored-normal distributions. In both cases, the underlying distribution is standard normal, $N(0,1)$. In (a), there are no values of Z observed below $Z = -0.75$, and the remaining density is rescaled (see the upper curve) to an area of 1. In (b), values below -0.75 are set to $Z = -0.75$; the probability of observing this value is represented by the "spike" topped by a circle.

σ^2—that is, $\xi \sim N(\mu, \sigma^2)$—and that this distribution is left-truncated at the *threshold* a, giving rise to the observable variable Y. Then, the mean and variance of Y are[38]

$$\begin{aligned} E(Y) &= E(\xi \mid \xi \geq a) = \mu + \sigma m(z_a) \\ V(Y) &= V(\xi \mid \xi \geq a) = \sigma^2[1 - d(z_a)] \end{aligned} \tag{20.14}$$

where

$$\begin{aligned} z_a &\equiv \frac{a - \mu}{\sigma} \\ m(z_a) &\equiv \frac{\phi(z_a)}{1 - \Phi(z_a)} = \frac{\phi(z_a)}{\Phi(-z_a)} \\ d(z_a) &\equiv m(z_a)[m(z_a) - z_a] \end{aligned} \tag{20.15}$$

The quantity $m(z_a)$, called the *inverse Mills ratio*, is a function of the standardized threshold; it will figure prominently in the remainder of this section. As a general matter, the mean of the left-truncated variable Y *exceeds* that of ξ by an amount that depends on the standardized threshold and the standard deviation σ of the untruncated distribution; similarly, the variance of Y is *smaller* than the variance of ξ by a factor dependent on the standardized threshold.[39] The inverse Mills ratio is graphed against z_a in Figure 20.4: As the threshold moves to the right, the relationship between the inverse Mills ratio and z_a becomes more linear.

[38]The derivation of these results, and of some other results in this section, is beyond the level of the text, even in starred material or exercises. See Johnson, Kotz, and Balakrishnan (1994) and Kotz, Balakrishnan, and Johnson (1994).

[39]See Exercise 29.4 for the mean and variance of a right-truncated normal variable.

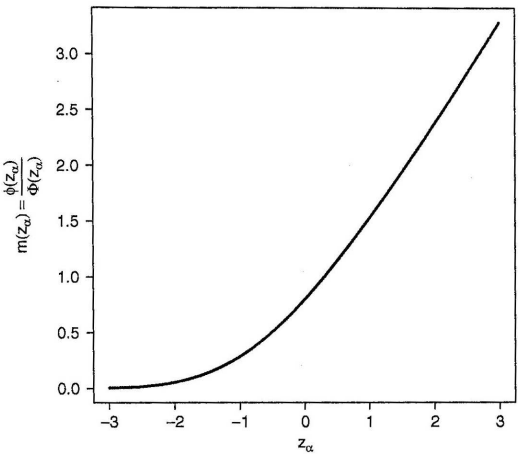

Figure 20.4 The inverse Mills ratio $m(z_\alpha)$ as a function of the standardized threshold z_α.

The expectation and variance of a censored-normal variable follow straightforwardly. Suppose that $\xi \sim N(\mu, \sigma^2)$ is left-censored at $\xi = a$, so that

$$Y = \begin{cases} a & \text{for } \xi \leq a \\ \xi & \text{for } \xi > a \end{cases}$$

Then,[40]

$$\begin{aligned} E(Y) &= a\Phi(z_a) + [\mu + \sigma m(z_a)][1 - \Phi(z_a)] \\ V(Y) &= \sigma^2[1 - \Phi(z_a)]\Big\{1 - d(z_a) + [z_a - m(z_a)]^2\Phi(z_a)\Big\} \end{aligned} \tag{20.16}$$

A variable can be truncated or censored at the right as well as at the left or can be truncated or censored at both ends simultaneously (the latter is termed *interval censoring*). The analysis of right-censored or interval-censored data is essentially similar to the analysis of left-censored data, making adjustments to the formulas in Equations 20.16.[41]

Finally, suppose that the unobservable variables ξ and ζ follow a bivariate-normal distribution, with means μ_ξ and μ_ζ, variances σ_ξ^2 and σ_ζ^2, and correlation ρ (so that the covariance of ξ and ζ is $\sigma_{\xi\zeta} = \rho\sigma_\xi\sigma_\zeta$). Imagine that, as before, Y is a truncated version of ξ, but now the truncation depends *not* on the value of ξ *itself* but rather on that of ζ, so that $Y = \xi$ when $\zeta \geq a$ and Y is unobserved when $\zeta < a$. This process is called *incidental truncation* or *selection*. The mean and variance for the incidentally truncated variable Y are

$$\begin{aligned} E(Y) &= E(\xi \mid \zeta \geq a) = \mu_\xi + \sigma_\xi \rho m(z_a) \\ V(Y) &= V(\xi \mid \zeta \geq a) = \sigma_\xi^2[1 - \rho^2 d(z_a)] \end{aligned} \tag{20.17}$$

where $z_a \equiv (a - \mu_\zeta)/\sigma_\zeta$ and $m(\cdot)$ and $d(\cdot)$ are defined as in Equations 20.15. The effect of incidental truncation depends, therefore, not only on the standardized threshold z_a but also on

[40] See Exercise 20.10.
[41] See Exercise 20.11.

the correlation ρ between the latent variables ξ and ζ. For example, if these variables are positively correlated, then $E(Y) > E(\xi)$ and $V(Y) < V(\xi)$.

The distribution of a variable is truncated when values below or above a threshold (or outside a particular range) are unobserved. The distribution of a variable is censored when values below or above a threshold (or outside a particular range) are set equal to the threshold. The distribution of a variable is incidentally censored if its value is unobserved when another variable is below or above a threshold (or outside a particular range). Simple formulas exist for the mean and variance of truncated- and censored-normal distributions and for the mean and variance of an incidentally truncated variable in a bivariate-normal distribution.

20.5.2 Heckman's Selection-Regression Model

The model and methods of estimation described in this section originated with James Heckman (e.g., Heckman, 1974, 1976), whose work on selection bias won him a Nobel Prize in economics. Heckman's selection-regression model consists of two parts:

1. A *regression equation* for a *latent response variable* ξ:

$$\begin{aligned}\xi_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \\ &= \eta_i + \varepsilon_i\end{aligned} \tag{20.18}$$

2. A *selection equation* that determines whether or not ξ is observed:

$$\begin{aligned}\zeta_i &= \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_p Z_{ip} + \delta_i \\ &= \psi_i + \delta_i\end{aligned} \tag{20.19}$$

where the *observed response variable*

$$Y_i = \begin{cases} \text{missing} & \text{for } \zeta_i \leq 0 \\ \xi_i & \text{for } \zeta_i > 0 \end{cases}$$

The explanatory variables in Equation 20.19 (i.e., the Zs) are intended to predict missingness; they need not be the same as the explanatory variables used in the regression equation of principal interest (Equation 20.18), but in applications, there is usually considerable overlap between the Zs and the Xs. The observed response, for example, might represent earnings for married women, which is missing when they are not in the paid labor force; the latent variable would then represent a notional "potential earnings." An example based on this idea is developed below.

It is assumed that the two error variables ε_i and δ_i follow a bivariate-normal distribution with means $E(\varepsilon_i) = E(\delta_i) = 0$, variances $\sigma_\varepsilon^2 \equiv V(\varepsilon_i)$ and $V(\delta_i) = 1$, and correlation $\rho_{\varepsilon\delta}$. Errors for

different observations are assumed to be independent. Equation 20.19, together with the assumption that the errors δ are normally distributed, specifies a *probit model* for nonmissingness.[42]

As we will see presently, estimating the regression equation (Equation 20.18) just for complete cases—simply omitting observations for which Y is missing—generally produces inconsistent estimates of the regression coefficients. In addition, because of the correlation of the two error variables, the missing data are not ignorable, and so it would be inappropriate, for example, to generate multiple imputations of the missing values of Y as if they were MAR.

Restricting our attention to the complete cases,

$$\begin{aligned} E(Y_i|\zeta_i > 0) &= \eta_i + E(\varepsilon_i \mid \zeta_i > 0) \\ &= \eta_i + E(\varepsilon_i \mid \delta_i > -\psi_i) \end{aligned}$$

The conditional expectation of the error ε_i follows from Equations 20.17 for the incidentally truncated bivariate-normal distribution:[43]

$$E(\varepsilon_i|\delta_i > -\psi_i) = \sigma_\varepsilon \rho_{\varepsilon\delta} m(-\psi_i)$$

Therefore,

$$\begin{aligned} E(Y_i|\zeta_i > 0) &= \eta_i + \sigma_\varepsilon \rho_{\varepsilon\delta} m(-\psi_i) \\ &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \beta_\lambda \lambda_i \end{aligned}$$

where $\beta_\lambda \equiv \sigma_\varepsilon \rho_{\varepsilon\delta}$ and $\lambda_i \equiv m(-\psi_i)$.

Letting $\nu_i \equiv Y_i - E(Y_i|\zeta_i > 0)$, we can write the regression equation for the complete cases as

$$(Y_i|\zeta_i > 0) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \beta_\lambda \lambda_i + \nu_i$$

Regressing Y on the Xs using only the complete cases omits the explanatory variable λ_i, which is the inverse Mills ratio based on the negative of the linear predictor ψ_i from the selection equation (Equation 20.19). Ignoring the missingness mechanism, therefore, can be conceptualized as a kind of omitted-variable specification error.[44] If the errors from the regression and selection equations are uncorrelated (i.e., $\rho_{\varepsilon\delta} = 0$), then $\beta_\lambda = 0$, and ignoring λ_i is inconsequential. Similarly, if λ_i were uncorrelated with the Xs, then we could ignore it without threatening the consistency of the least-squares estimators of the regression coefficients. Uncorrelation of λ_i and the Xs is unlikely, however: The selection and regression equations typically contain many of the same explanatory variables, and unless the degree of selection is low, the inverse Mills ratio is nearly a linear function of the linear predictor (recall Figure 20.4 on page 631). Indeed, *high* correlation between λ_i and the Xs can make consistent estimation of the regression coefficients (by the methods described immediately below) unstable. Note, as well, that the variance of the errors ν_i is not constant.[45]

There are two common strategies for estimating Heckman's regression-selection model: direct application of ML estimation and employing an estimate of λ_i as an auxiliary regressor.

---

[42]For a general treatment of probit regression, see Section 14.1. Recall that we can arbitrarily set the threshold above which Y is observed and below which it is missing to 0 and the error variance δ to 1, to fix the origin and scale of the latent variable.

[43]See Exercise 20.12.

[44]See Sections 6.3 and 9.7.

[45]The variance of ν_i follows from Equations 20.17 for the variance of an incidentally truncated variable in the bivariate-normal distribution: See Exercise 20.13.

- *ML Estimation*:* Let $\boldsymbol{\beta} \equiv (\alpha, \beta_1, \ldots, \beta_k)'$ be the vector of regression coefficients in the regression equation (Equation 20.18); let $\boldsymbol{\gamma} \equiv (\gamma_0, \gamma_1, \ldots, \gamma_p)'$ be the vector of regression coefficients in the selection equation (Equation 20.19); let $\mathbf{x}_i' \equiv (1, X_{i1}, \ldots, X_{ik})$ be the ith row of the model matrix for the regression equation; and let $\mathbf{z}_i' \equiv (1, Z_{i1}, \ldots, Z_{ip})$ be the ith row of the model matrix for the selection equation. For notational convenience, order the data so that the missing observations on Y are the first m of n observations. Then the log-likelihood for Heckman's model can be formulated as follows:[46]

$$\log_e L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2, \rho_{\varepsilon\delta}) = \sum_{i=1}^{m} \log_e \Phi(\mathbf{z}_i'\boldsymbol{\gamma}) + \sum_{i=m+1}^{n} \log_e \left[\frac{1}{\sigma_\varepsilon} \phi\left(\frac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma_\varepsilon}\right) \Phi\left(\frac{\mathbf{z}_i'\boldsymbol{\gamma} + \rho_{\varepsilon\delta}\dfrac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma_\varepsilon}}{\sqrt{\dfrac{1-\rho_{\varepsilon\delta}}{\sigma_\varepsilon}}}\right)\right] \tag{20.20}$$

  This log-likelihood can be maximized numerically.
- *Two-Step Estimation:* Heckman (1979) also proposed a simple and widely used two-step procedure for estimating his regression-selection model.

  Step 1: Define the dichotomous response variable

$$W_i = \begin{cases} 1 & \text{if } Y_i \text{ is observed} \\ 0 & \text{if } Y_i \text{ is missing} \end{cases}$$

  Perform a probit regression of W_i on the Zs, estimating the γs in the usual manner by ML,[47] and finding fitted values on the probit scale,

$$\widehat{\psi}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 Z_{i1} + \widehat{\gamma}_2 Z_{i2} + \cdots + \widehat{\gamma}_p Z_{ip}$$

  $\widehat{\psi}_i$ is simply the estimated linear predictor from the probit model. For each observation, compute the estimated inverse Mills ratio

$$\widehat{\lambda}_i = m(-\widehat{\psi}_i) = \frac{\phi(-\widehat{\psi}_i)}{1 - \Phi(-\widehat{\psi}_i)} = \frac{\phi(\widehat{\psi}_i)}{\Phi(\widehat{\psi}_i)}$$

  Step 2: Use $\widehat{\lambda}$ as an auxiliary regressor in the least-squares regression of Y_i on the Xs for the complete cases,

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \beta_{\widehat{\lambda}} \widehat{\lambda}_i + \nu_i^*, \quad \text{for } i = m+1, \ldots, n \tag{20.21}$$

  This least-squares regression provides consistent estimates of the regression coefficients, $\alpha, \beta_1, \beta_2, \ldots, \beta_k$. The heteroscedasticity of the errors, however, requires an adjustment to the usual OLS standard errors.[48]

[46] See Exercise 20.14.
[47] As described in Chapters 14 and 15.
[48] See Exercise 20.15.

To illustrate the application of Heckman's selection-regression model, I will return to the Canadian Survey of Labour and Income Dynamics (the SLID),[49] examining the relationship between women's earnings and their education, age, and the region in which they reside, restricting attention to married women between the ages of 18 and 65. Earnings is represented by the women's composite hourly wage rate, which is missing if they are not in the paid labor force. Preliminary examination of the data suggested regressing the log of composite hourly wages on the square of years of education and a quadratic in age, along with four dummy regressors for five regions of Canada (the Atlantic provinces, Quebec, Ontario, the prairie provinces, and British Columbia, taking the Atlantic provinces as the baseline category).

Of the 6427 women in the sample, 3936 were in the paid labor force.[50] Because women whose potential earnings are relatively low may very well be less likely to work outside the home, there is a potential for selection bias if we simply ignore the 2491 women who are not in the labor force, causing us to underestimate the effects of the explanatory variables on (potential) earnings.[51]

I formulated a selection model in which labor-force participation is regressed on region dummy variables; dummy regressors for the presence in the household of children 0 to 4 and 5 to 9 years of age; family income less the woman's own income, if any (in thousands of dollars); education (in years); and a quadratic in age (in years). The results are shown in Table 20.5. At the left of the table are ordinary least-squares estimates *ignoring* the selection process. The table also shows two-step and ML estimates for the Heckman model, both for the regression equation and for the selection equation. For the two-step estimation procedure, the selection equation was estimated in a preliminary probit regression.

In this application, the two-step/probit and ML estimates are very similar and are not terribly different from the OLS estimates based on the complete cases. Moreover, the ML estimate of the correlation between the errors of the regression and selection equations is fairly small: $\widehat{\rho}_{\varepsilon\delta} = .320$. The degree of collinearity induced by the introduction of the inverse Mills ratio regressor in the second step of the two-step procedure is not serious, as shown in Table 20.6, which compares generalized variance inflation factors for the model as estimated by OLS and Heckman's two-step procedure.[52]

[49]In Chapter 12, I used the SLID for a regression of earnings on sex, age, and education. In Chapter 14, the SLID provided data for a logistic regression of young married women's labor force participation on region, presence of children, family income, and education.

[50]The complete SLID sample of married women between 18 and 65 years of age consists of 6900 respondents. I omitted the relatively small number of observations (comprising about 7% of the sample) with missing data on variables other than earnings.

[51]If, however, our goal is to *describe* the regression of earnings on the explanatory variables for those who are in the paid labor force, then an analysis based on women who have earnings should be perfectly fine, as long as we are careful to ensure the descriptive accuracy of the model—for example, by using component-plus-residual plots to check for nonlinearity (see Chapter 12).

[52]Generalized variance-inflation factors (GVIFs), introduced in Section 13.1.2, are appropriate for terms in a model that have more than 1 degree of freedom, such as the region and age terms in this model. When a term has 1 degree of freedom, the GVIF reduces to the usual variance inflation factor (VIF, also discussed in Chapter 13). Taking the $1/(2df)$ power of the GVIF makes values roughly comparable across different degrees of freedom. Treating the linear and quadratic components of the age term as a set is important here because otherwise there would be artifactual collinearity induced by the high correlation between Age and Age^2.

Table 20.5 Least-Squares, Heckman Two-Step, and Heckman ML Estimates for the Regression of Women's Composite Hourly Wages on Region, Education, and Age

	OLS		*Two-Step/Probit*		*ML*	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Coefficient			*Regression Equation*			
Constant	1.10	0.15	0.442	0.227	0.755	0.177
Quebec	0.223	0.031	0.205	0.033	0.214	0.032
Ontario	0.303	0.026	0.332	0.028	0.319	0.027
Prairies	0.126	0.027	0.147	0.029	0.137	0.027
B.C.	0.371	0.036	0.392	0.038	0.382	0.037
Education^2	0.00442	0.00013	0.00492	0.00018	0.00469	0.00014
Age	0.0687	0.0074	0.0917	0.0096	0.0807	0.0081
Age^2	−0.000717	0.000088	−0.00105	0.00012	−0.000892	0.000099
Inv. Mills Ratio			0.361	0.088		
			Selection Equation			
Constant			−1.46	0.30	−1.44	0.30
Quebec			−0.0665	0.0533	−0.0674	0.0533
Ontario			0.193	0.048	0.194	0.048
Prairies			0.117	0.049	0.117	0.049
B.C.			0.145	0.067	0.150	0.067
Children 0–4			−0.414	0.050	−0.439	0.049
Children 5–9			−0.261	0.043	−0.251	0.042
Family Income			−0.00399	0.00097	−0.00475	0.00097
Education			0.0815	0.0061	0.0817	0.0060
Age			0.0878	0.0135	0.0882	0.0134
Age^2			−0.00145	0.00015	−0.00145	0.00015

As is seldom a bad idea, I will leave the last word on Heckman-type adjustments for selection bias to John Tukey (1986), who states,[53]

> I think that an important point that we have to come back to at intervals is that knowledge always comes from a combination of data and assumptions. If the assumptions are too important, many of us get unhappy. I think one thing we were told in this last discussion was that all the formal ways that have been found for attacking this problem ended up being very dependent upon these assumptions. Therefore, people like me have to be very uncomfortable about the results. (p. 58)

[53]Tukey made these comments about a paper delivered by Heckman and Robb (1986) to a symposium on statistical methods for self-selected samples (collected in a volume edited by Wainer, 1986). The models introduced by Heckman and Robb are not the same as Heckman's selection-regression model discussed in this section, but they are similarly motivated and structured.

Table 20.6 Generalized Variance Inflation Factors for Terms in the OLS and Heckman Two-Step Regression of Log Hourly Wages on Region, Education, and Age

		$GVIF^{1/(2df)}$	
Term	*df*	*OLS Estimates*	*Heckman Two-Step Estimates*
Region	4	1.003	1.024
Education^2	1	1.025	1.402
Age (quadratic)	2	1.012	1.347
Inverse Mills Ratio	1	–	2.202

Heckman's regression model consists of two parts:

1. A regression equation for a latent response variable ξ:

$$\xi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

2. A selection equation that determines whether or not ξ is observed:

$$\zeta_i = \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_p Z_{ip} + \delta_i = \psi_i + \delta_i$$

where the observed response variable

$$Y_i = \begin{cases} \text{missing} & \text{for } \zeta_i \leq 0 \\ \xi_i & \text{for } \zeta_i > 0 \end{cases}$$

It is assumed that the two error variables ε_i and δ_i follow a bivariate-normal distribution with means $E(\varepsilon_i) = E(\delta_i) = 0$, variances $V(\varepsilon_i) = \sigma_\varepsilon^2$ and $V(\delta_i) = 1$, and correlation $\rho_{\varepsilon\delta}$ and that errors for different observations are independent. Heckman's model can be consistently estimated by ML or by a two-step procedure. In the first step of the two-step procedure, the selection equation is estimated as a probit model; in the second step, the regression equation is estimated by OLS after incorporating the auxiliary regressor $\widehat{\lambda}_i = \phi(\widehat{\psi}_i)/\Phi(\widehat{\psi}_i)$, where $\widehat{\psi}_i$ is the fitted value from the first-step probit equation, $\phi(\cdot)$ is the density function of the standard-normal distribution, and $\Phi(\cdot)$ is the distribution function of the standard-normal distribution.

20.5.3 Censored-Regression Models

When the response variable Y in a regression is censored, values of Y cannot be observed outside a certain range—say, the interval (a, b). We can detect, however, whether an observation falls below the lower threshold a or above the upper threshold b, and consequently, we have *some* information about the censored values.

Let us assume, in particular, that the *latent response variable* ξ is linearly related to the regressors $X_1, X_2, \ldots, X_k$, so that

$$\xi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{20.22}$$

and that the other assumptions of the normal-regression model hold: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and $\varepsilon_i, \varepsilon_{i'}$ are independent for $i \neq i'$. We cannot observe ξ directly, however, but instead we collect data on the *censored response variable* Y, where

$$Y_i = \begin{cases} a & \text{for } \xi_i \leq a \\ \xi_i & \text{for } a < \xi_i < b \\ b & \text{for } \xi_i \geq b \end{cases} \tag{20.23}$$

Equations 20.22 and 20.23 define the *censored-regression model.* A model of this type was first proposed by James Tobin (1958), for data censored to the left at 0—that is, for $a = 0$ and $b = \infty$. Left-censored regression models are called *tobit models*, in honor of Tobin (another Nobel Prize winner in economics). The censored-regression model can be estimated by the method of ML.[54]

*Rewriting the regression equation in vector form for compactness as $\xi_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, the log-likelihood for the censored-regression model in Equations 20.22 and 2.23 is

$$\log_e L(\boldsymbol{\beta}, \sigma_\varepsilon^2) = \sum_{Y_i = a} \log_e \Phi\left(\frac{a - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma_\varepsilon}\right) + \sum_{a < Y_i < b} \log_e \left[\frac{1}{\sigma_\varepsilon}\phi\left(\frac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma_\varepsilon}\right)\right] + \sum_{Y_i = b} \log_e \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta} - b}{\sigma_\varepsilon}\right)$$

The log-likelihood therefore comprises terms for left-censored, fully observed, and right-censored observations.[55]

For an example of censored regression, I turn once again to the Canadian SLID data. We last encountered the SLID in the previous section, where the earnings of married women were regressed on region, education, and age. I employed Heckman's selection-regression model because earnings were unavailable for women who were not in the paid labor force. I will now develop a similar example in which the response variable is hours worked in the year preceding the survey. This variable is left-censored at the value 0, producing a classic tobit regression model.[56] The explanatory variables are region, the presence in the household of children 0 to 4 and 5 to 9 years old, family income less the woman's own income (if any), education, and a quadratic in age. The SLID data set includes 6340 respondents with valid data on the variables employed in this example.

Preliminary examination of the data suggested a square-root transformation of hours worked. This transformation does not, of course, serve to spread out the values of the response variable for the 31% of respondents who reported 0 hours worked—that is, the transformed response for all censored observations is $\sqrt{0} = 0$. OLS and ML tobit estimates for the regression model are shown in Table 20.7. The OLS estimates are consistently smaller in magnitude than the corresponding tobit estimates.[57]

[54] An alternative is to employ Heckman's two-step procedure, described in the preceding section.

[55] Estimation is facilitated by reparameterization. See, for example, Greene (2003, Section 22.3.3).

[56] The latent response variable therefore represents "propensity" to work outside the home; presumably, if that propensity is above the threshold 0, we observe positive hours worked.

[57] See Exercise 20.16.

Table 20.7 OLS and ML Tobit Estimates for the Regression of Square-Root Hours Worked on Several Explanatory Variables

	OLS		*Tobit*	
Coefficient	*Estimate*	*SE*	*Estimate*	*SE*
Constant	−20.3	3.8	−58.7	5.4
Quebec	−0.745	0.710	−1.58	1.01
Ontario	3.55	0.63	5.02	0.89
Prairies	3.64	0.65	5.36	0.91
B.C.	2.09	0.88	3.73	1.23
Children 0–4 (present)	−6.56	0.65	−8.63	0.91
Children 5–9 (present)	−5.05	0.56	−6.91	0.79
Family Income ($1000s)	−0.0977	0.0128	−0.139	0.018
Education (years)	1.29	0.08	1.87	0.11
Age (years)	2.32	0.17	3.84	0.25
Age^2	−0.0321	0.0019	−0.0529	0.0028

In the censored-regression model, the latent response variable ξ is linearly related to the regressors $X_1, X_2, \ldots, X_k$:

$$\xi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and $\varepsilon_i, \varepsilon_{i'}$ are independent for $i \neq i'$. We cannot observe ξ directly but instead collect data on the censored response variable Y:

$$Y_i = \begin{cases} a & \text{for } \xi_i \leq a \\ \xi_i & \text{for } a < \xi_i < b \\ b & \text{for } \xi_i \geq b \end{cases}$$

When Y is left-censored at 0 (i.e., $a = 0$ and $b = \infty$), the censored-regression model is called a tobit model in honor of James Tobin. The censored-regression model can be estimated by maximum likelihood.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 20.1. Consider the following contrived data set for the variables X_1, X_2, and X_3, where the question marks indicate missing data:

X_1	X_2	X_3
1	1	?
1	?	1
−1	−1	?
−1	?	−1
?	1	−1
?	−1	1
5	?	?

(a) Using available cases (and recomputing the means and standard deviations for each pair of variables), find the pairwise correlations among the three variables and explain why the correlations are not consistent with each other.

(b) Compute the correlation between X_1 and X_2 using means and standard deviations computed *separately* from the valid observations for each variable. What do you find?

(c) *Show that the available-case correlation matrix among the variables X_1, X_2, and X_3 is not positive semidefinite.

Exercise 20.2. *In univariate missing data, where there are missing values for only one variable in a data set, some of the apparently distinct methods for handling missing data produce identical results for certain statistics. Consider Table 20.1 on page 612, for example, where data are missing on the variable X_2 but not on X_1. Note that the complete-case, available-case, and mean-imputation estimates of the slope β_{12} for the regression of X_1 on X_2 are identical. Prove that this is no accident. Are there are any other apparent agreements between or among methods in the table? If so, can you determine whether they are coincidences?

Exercise 20.3. *Duplicate the small simulation study reported in Table 20.2 on page 613, comparing several methods of handling univariate missing data that are MAR. Then repeat the study for missing data that are MCAR and for missing data that are MNAR (generated as in Figure 20.1 on page 608). What do you conclude? *Note:* This is not a *conceptually* difficult project, but it is potentially time-consuming; it also requires some programming skills and statistical software that can generate and analyze simulated data.

Exercise 20.4. *Equation 20.6 (on page 616) gives the ML estimators for the parameters μ_1, μ_2, σ_1^2, σ_2^2, and σ_{12} in the bivariate-normal model with some observations on X_2 missing at random but X_1 completely observed. The interpretation of $\widehat{\mu}_1$ and $\widehat{\sigma}_1^2$ is straightforward: They are the available-case mean and variance for X_1. Noting that S_{12}^*/S_1^{2*} is the complete-case slope for the regression of X_2 and X_1, offer interpretations for the other ML estimators.

Exercise 20.5. **Multivariate* linear regression fits the model

$$\underset{(n\times m)}{\mathbf{Y}} = \underset{(n\times k+1)}{\mathbf{X}}\ \underset{(k+1\times m)}{\mathbf{B}} + \underset{(n\times m)}{\mathbf{E}}$$

where $\mathbf{Y}$ is a matrix of response variables; $\mathbf{X}$ is a model matrix (just as in the *univariate* linear model); $\mathbf{B}$ is a matrix of regression coefficients, one column per response variable; and $\mathbf{E}$ is a

matrix of errors. The least-squares estimator of $\mathbf{B}$ is $\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ (equivalent to what one would get from separate least squares regressions of each Y on the Xs). See Section 9.5 for a discussion of the multivariate linear model.

(a) Show how $\widehat{\mathbf{B}}$ can be computed from the means of the variables, $\widehat{\mu}_Y$ and $\widehat{\mu}_X$, and from their covariances, $\widehat{\mathbf{\Sigma}}_{XX}$ and $\widehat{\mathbf{\Sigma}}_{XY}$ (among the Xs and between the Xs and Ys, respectively).

(b) The fitted values from the multivariate regression are $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}}$. It follows that the fitted values $\widehat{Y}_{ij}$ and $\widehat{Y}_{ij'}$ for the ith observation on response variables j and j' are both linear combinations of the ith row of the model matrix, $\mathbf{x}_i'$. Use this fact to find an expression for the covariance of $\widehat{Y}_{ij}$ and $\widehat{Y}_{ij'}$.

(c) Show how this result can be used in Equation 20.7 (on page 618), which applies the EM algorithm to multivariate-normal data with missing values.

Exercise 20.6. *Consider once again the case of univariate missing data MAR for two bivariately normal variables, where the first variable, X_1, is completely observed, and m observations (for convenience, the first m) on the second variable, X_2, are missing.

(a) Let $A^*_{2|1}$ and $B^*_{2|1}$ represent the intercept and slope for the complete-case least-squares regression of X_2 on X_1. Show that $A^*_{2|1}$ and $B^*_{2|1}$ are the ML estimators of $\alpha_{2|1}$ and $\beta_{2|1}$. (*Hint:* Use Equations 20.6 giving the ML estimators of μ_1, μ_2, σ_1^2, σ_2^2, and σ_{12}.)

(b) Show that the M step from the first iteration of the EM algorithm (see Equations 20.8 and 20.9 on page 618 for the E step) produces the ML estimates (given in Equations 20.6 on page 616). That is, demonstrate that the EM algorithm converges in a single iteration.

Exercise 20.7. As explained in Section 20.4.1, the efficiency of the multiple-imputation estimator of a coefficient $\widetilde{\beta}_j$ relative to the ML estimator $\widehat{\beta}_j$ is $\text{RE}(\widetilde{\beta}_j) = g/(g + \gamma_j)$, where g is the number of imputations employed and γ_j is the rate of missing information for coefficient β_j. The square root of $\text{RE}(\widetilde{\beta}_j)$ expresses relative efficiency on the coefficient standard-error scale. Compute $\text{RE}(\widetilde{\beta}_j)$ and $\sqrt{RE(\widetilde{\beta}_j)}$ for combinations of values of $g = 1, 2, 3, 5, 10, 20$, and 100, and $\gamma_j = .05, .1, .2, .5, .9$, and .99. What do you conclude about the number of imputations required for efficient inference?

Exercise 20.8. Examine the United Nations data on infant mortality and other variables for 207 countries, discussed in Section 20.4.4.

(a) Perform a complete-case linear least-squares regression of infant mortality on GDP per capita, percentage using contraception, and female education. Does it appear reasonable to log-transform infant mortality and GDP to linearize this regression? What about contraception and education?

(b) *Examine a scatterplot matrix (Section 3.3.1) for the variables used in the imputation example. What do you find? Then apply the multivariate Box-Cox procedure described in Section 4.6 to these variables. Remember first to subtract 35 from female expectation of life (why?). Do the results that you obtain support the transformations

employed in the text? Apply the transformations and reexamine the data. Do they appear more nearly normal?

Exercise 20.9. Truncated normal distributions:

(a) Suppose that $\xi \sim N(0, 1)$. Using Equations 20.14 (page 630) for the mean and variance of a left-truncated normal distribution, calculate the mean and variance of $\xi \mid \xi > a$ for each of $a = -2, -1, 0, 1$, and 2.

(b) *Find similar formulas for the mean and variance of a *right*-truncated normal distribution. What happens to the mean and variance as the threshold moves to the left?

Exercise 20.10. *Suppose that $\xi \sim N(\mu, \sigma^2)$ is left-censored at $\xi = a$, so that

$$Y = \begin{cases} a & \text{for } \xi \leq a \\ \xi & \text{for } \xi > a \end{cases}$$

Using Equations 20.14 (on page 630) for the *truncated* normal distribution, show that (repeating Equations 20.16 on page 631)

$$E(Y) = a\Phi(z_a) + [\mu + \sigma m(z_a)][1 - \Phi(z_a)]$$
$$V(Y) = \sigma^2[1 - \Phi(z_a)]\left\{1 - d(z_a) + [z_a - m(z_a)]^2\Phi(z_a)\right\}$$

Exercise 20.11. *Equations 20.16 (on page 631) give formulas for the mean and variance of a left-censored normally distributed variable. (These formulas are also shown in the preceding exercise.) Derive similar formulas for (a) a right-censored and (b) an interval-censored normally distributed variable.

Exercise 20.12. *Using Equations 20.17 (page 631) for the incidentally truncated bivariate-normal distribution, show that the expectation of the error ε_i in the Heckman regression model (Equations 20.18 and 20.19 on page 632) conditional on Y being observed is

$$E(\varepsilon_i \mid \zeta_i > 0) = E(\varepsilon_i | \delta_i > -\psi_i) = \sigma_\varepsilon \rho_{\varepsilon\delta} m(-\psi_i)$$

Exercise 20.13. *As explained in the text, the Heckman regression model (Equations 20.18 and 20.19, page 632) implies that

$$(Y_i | \zeta_i > 0) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \beta_\lambda \lambda_i + \nu_i$$

where $\beta_\lambda \equiv \sigma_\varepsilon \rho_{\varepsilon\delta}$, $\lambda_i \equiv m(-\psi_i)$, and

$$\psi_i = \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_p Z_{ip}$$

Show that the errors ν_i are heteroscedastic, with variance

$$V(\nu_i) = \sigma_\varepsilon^2[1 - \rho_{\varepsilon\delta}^2 \lambda_i(\lambda_i + \psi_i)]$$

where σ_ε^2 is the error variance in the regression equation (Equation 20.18), and $\rho_{\varepsilon\delta}$ is the correlation between the errors of the regression and selection equations. (*Hint:* See Equations 20.17

on page 631 for the variance of an incidentally truncated variable in a bivariate-normal distribution.)

Exercise 20.14. *The log-likelihood for the Heckman regression-selection model is given in Equation 20.20 (page 634). Derive this expression. (*Hint:* The first sum in the log-likelihood, for the observations for which Y is missing, is of the log-probability that each such Y_i is missing; the second sum is of the log of the probability density at the observed values of Y_i times the probability that each such value is observed.)

Exercise 20.15. *Explain how White's coefficient-variance estimator (see Section 12.2.3), which is used to correct the covariance matrix of OLS regression coefficients for heteroscedasticity, can be employed to obtain consistent coefficient standard errors for the two-step estimator of Heckman's regression-selection model—the second step of which entails an OLS regression with heteroscedastic errors (Equation 20.21 on page 634). (*Hint:* Refer to Exercise 20.13 for the variance of the errors in the second-step OLS regression.)

Exercise 20.16. Greene (2003 p. 768) remarks that the ML estimates $\widehat{\beta}_j$ of the regression coefficients in a censored-regression model are often approximately equal to the OLS estimates B_j divided by the proportion P of *uncensored* observations; that is, $\widehat{\beta}_j \approx B_j/P$. Does this pattern hold for the hours-worked regression in Table 20.7 (page 639), where $P = .69$?

Summary

- Missing data are missing completely at random (MCAR) if they can be regarded as a simple random sample of the complete data. If missingness is related to the observed data but not to the missing data (conditional on the observed data), then data are missing at random (MAR). If missingness is related to the missing values themselves, even when the information in the observed data is taken into account, then data are missing not at random (MNAR). When data are MCAR or MAR, the process that produces missing data is ignorable, in the sense that valid methods exist to deal with the missing data without explicitly modeling the process that generates them. In contrast, when data are MNAR, the process producing missing data is nonignorable and must be modeled. Except in special situations, it is not possible to know whether data are MCAR, MAR, or MNAR.
- Traditional methods of handling missing data include complete-case analysis, available-case analysis, and unconditional and conditional mean imputation. Complete-case analysis produces consistent estimates and valid statistical inferences when data are MCAR (and in certain other special circumstances), but even in this advantageous situation, it does not use information in the sample efficiently. The other traditional methods suffer from more serious problems.
- The method of maximum likelihood (ML) can be applied to parameter estimation in the presence of missing data. If the assumptions made concerning the distribution of the complete data and the process generating missing data hold, then ML estimates have their usual optimal properties, such as consistency and asymptotic efficiency. When data are MAR, the ML estimate $\widehat{\theta}$ of the parameters θ of the complete-data distribution can be obtained from the marginal distribution of the observed data, by integrating over the missing data,

$$p(\mathbf{X}_{\text{obs}};\boldsymbol{\theta}) = \int p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}};\boldsymbol{\theta})d\mathbf{X}_{\text{mis}}$$

Although it may be difficult to apply this result directly, simplification is possible in certain cases. Once we have found the ML parameter estimates, we can proceed with statistical inference in the usual manner, for example, by computing likelihood-ratio tests of nested models and constructing Wald tests or confidence intervals.

- The EM algorithm is a general iterative procedure for finding ML estimates—but not their standard errors—in the presence of arbitrary patterns of missing data. When data are MAR, iteration l of the EM algorithm consists of two steps: (1) In the E (expectation) step, we find the expectation of the complete-data log-likelihood, integrating over the missing data, given the observed data and the current estimates of the parameters:

$$E\left[\log_e L(\boldsymbol{\theta};\mathbf{X})|\boldsymbol{\theta}^{(l)}\right] = \int \log_e L(\boldsymbol{\theta};\mathbf{X})p\left(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}},\boldsymbol{\theta}^{(l)}\right)d\mathbf{X}_{\text{mis}}$$

(2) In the M (maximization) step, we find the values $\boldsymbol{\theta}^{(l+1)}$ of $\boldsymbol{\theta}$ that maximize the expected log-likelihood $E\left[\log_e L(\boldsymbol{\theta};\mathbf{X})|\boldsymbol{\theta}^{(l)}\right]$; these are the parameter estimates for the next iteration. At convergence, the EM algorithm produces the ML estimates $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

- Bayesian multiple imputation (MI) is a flexible and general method for dealing with data that are missing at random. The essential idea of multiple imputation is to reflect the uncertainty associated with missing data by imputing g values for each missing value, drawing each imputed value from the predictive distribution of the missing data (a process that usually requires simulation), and therefore producing not one but g completed data sets. Standard methods of statistical analysis are then applied in parallel to the completed data sets.

  - According to Rubin's rules, MI estimates (e.g., of a population regression coefficient β_j) are obtained by averaging over the imputed data sets:

$$\widetilde{\beta}_j = \frac{\sum_{l=1}^{g} B_j^{(l)}}{g}$$

    where $B_j^{(l)}$ is the estimate of β_j from imputed data set l.
  - Standard errors of the estimated coefficients are obtained by combining information about within- and between-imputation variation in the coefficients,

$$\widetilde{SE}(\widetilde{\beta}_j) = \sqrt{V_j^{(W)} + \frac{g+1}{g}V_j^{(B)}}$$

    where the within-imputation component is

$$V_j^{(W)} = \frac{\sum_{l=1}^{g} \text{SE}^2\left(B_j^{(l)}\right)}{g}$$

    and the between-imputation component is

$$V_j^{(B)} = \frac{\sum_{l=1}^{g} \left(B_j^{(l)} - \widetilde{\beta}_j\right)^2}{g-1}$$

Here, $\mathrm{SE}\left(B_j^{(l)}\right)$ is the standard error of B_j computed in the usual manner for the lth imputed data set.

– Inference based on $\widehat{\beta}_j$ and $\widetilde{\mathrm{SE}}(\widetilde{\beta}_j)$ uses the t-distribution, with degrees of freedom determined by

$$df_j = (g-1)\left(1 + \frac{g}{g+1} \times \frac{V_j^{(W)}}{V_j^{(B)}}\right)^2$$

Inference for several coefficients proceeds in a similar, if more complex, manner.

- Multiple imputation based on the multivariate-normal distribution can be remarkably effective in practice, even when the data are not normally distributed. To apply multiple imputation effectively, however, it is important to include variables in the imputation model that make the assumption of ignorable missingness reasonable; to transform variables to approximate normality, if possible; to adjust the imputed data so that they resemble the original data; and to make sure that the imputation model captures features of the data, such as nonlinearities and interactions, to be used in subsequent data analysis.
- The distribution of a variable is truncated when values below or above a threshold (or outside a particular range) are unobserved. The distribution of a variable is censored when values below or above a threshold (or outside a particular range) are set equal to the threshold. The distribution of a variable is incidentally censored if its value is unobserved when another variable is below or above a threshold (or outside a particular range). Simple formulas exist for the mean and variance of truncated- and censored-normal distributions and for the mean and variance of an incidentally truncated variable in a bivariate-normal distribution.
- Heckman's regression model consists of two parts:

1. A regression equation for a latent response variable ξ,

$$\xi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

2. A selection equation that determines whether or not ξ is observed,

$$\zeta_i = \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_p Z_{ip} + \delta_i = \psi_i + \delta_i$$

where the observed response variable

$$Y_i = \begin{cases} \text{missing} & \text{for } \zeta_i \leq 0 \\ \xi_i & \text{for } \zeta_i > 0 \end{cases}$$

It is assumed that the two error variables ε_i and δ_i follow a bivariate-normal distribution with means $E(\varepsilon_i) = E(\delta_i) = 0$, variances $V(\varepsilon_i) = \sigma_\varepsilon^2$ and $V(\delta_i) = 1$, and correlation $\rho_{\varepsilon\delta}$, and that errors for different observations are independent.

Heckman's model can be consistently estimated by ML or by a two-step procedure. In the first step of the two-step procedure, the selection equation is estimated as a

probit model; in the second step, the regression equation is estimated by OLS after incorporating the auxiliary regressor $\widehat{\lambda}_i = \phi(\widehat{\psi}_i)/\Phi(\widehat{\psi}_i)$, where $\widehat{\psi}_i$ is the fitted value from the first-step probit equation, $\phi(\cdot)$ is the density function of the standard-normal distribution, and $\Phi(\cdot)$ is the distribution function of the standard-normal distribution.

- In the censored-regression model, the latent response variable ξ is linearly related to the regressors $X_1, X_2, \ldots, X_k$:

$$\xi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and $\varepsilon_i, \varepsilon_{i'}$ are independent for $i \neq i'$. We cannot observe ξ directly but instead collect data on the censored response variable Y,

$$Y_i = \begin{cases} a & \text{for } \xi_i \leq a \\ \xi_i & \text{for } a < \xi_i < b \\ b & \text{for } \xi_i \geq b \end{cases}$$

When Y is left-censored at 0 (i.e., $a = 0$ and $b = \infty$), the censored-regression model is called a tobit model in honor of James Tobin. The censored-regression model can be estimated by ML.

Recommended Reading

- Little and Rubin (2002), central figures in the recent development of more adequate methods for handling missing data, present a wide-ranging and largely accessible overview of the field. A briefer treatment by the same authors appears in Little and Rubin (1990).
- Another fine, if mathematically more demanding, book on handling missing data is Schafer (1997). Also see the overview paper by Schafer and Graham (2002).
- van Buuren (2012) is an accessible, book-length presentation of the simpler chained-equations approach to multiple imputation of missing data.
- Allison's (2002) monograph on missing data is clear, comprehensive, and directed to social scientists (as is the paper by King et al., 2001).
- The edited volume by Wainer (1986) on sample-selection issues contrasts the points of view of statisticians and econometricians—in particular in an exchange between John Tukey and James Heckman. Also see the paper by Stolzenberg and Relles (1997) and the review paper by Winship and Mare (1992).