

1 Missing Data (Soc 504)

The topic of missing data has gained considerable attention in the last decade, as evidenced by several recent trends. First, most graduating PhDs in statistics now claim “missing data” as an area of interest or expertise. Second, it has become difficult to publish empirical work in sociology without discussion of how missing data was handled. Third, more and more methods for handling missing data have sprouted-up over the last few years.

Although missing data has received a growing amount of attention, there are still some key misunderstandings regarding the problems that missing data generate, as well as acceptable solutions. Missing data are important to consider, because they may lead to substantial biases in analyses. On the other hand, missing data is often harmless beyond reducing statistical power. In these notes, I define various types of missingness and discuss methods for handling it in your research.

2 Types of Missingness

Little and Rubin (1987) define three different classes of missingness. Here, I define the key terms used in discussing missingness in the literature. I will then discuss how they relate to some ways in which we encounter missing data.

- Data missing on Y are *observed at random (OAR)* if missingness on Y is not a function of X . Phrased another way, if X determines missingness on Y , the data are not OAR.
- Data missing on Y are *missing at random (MAR)* if missingness on Y is not a function of Y . Phrased another way, if Y determines missingness on Y , the data are not MAR.
- Data are *Missing Completely at Random (MCAR)* if missingness on Y is unrelated to X or Y . In other words $MCAR = OAR + MAR$. If the data are MCAR or at least MAR, then the missing data mechanism is considered “ignorable.” Otherwise, the missing data mechanism is considered “nonignorable.”

To make these ideas concrete, suppose we are examining the effect of education on income. If missingness on income is a function of education (e.g., highly educated individuals don't report their income), then the data are not OAR. If missingness on income is a function of income (e.g., persons with high income do not report their income), then the data are not MAR.

There are a number of ways in which we may encounter missing data in social science research, including the following:

- Individuals not followed up by design (meets MCAR assumption)
- Item nonresponse (may not meet any assumption)

- Loss due to followup, or attrition (may not meet any assumption)
- Mortality (respondent's, not yours)
- Sample selection (e.g., estimating a model that is only applicable to a subset of the total sample) (may not meet any assumption)

All of these avenues of missingness are common in sociology, and indeed, it is generally more surprising to find you have very little missing data when conducting an analysis than to find you have much missing data. The good news is that the statistical properties of maximum likelihood estimation obtain if the data are MCAR or MAR. That is, the data do not have to be OAR. This is true only when the model is properly specified, however. In that case, whatever “piece” of the data is observed is sufficient to produce unbiased parameters. Why is this true? Below is a plot of the relationship between some variable X and Y.

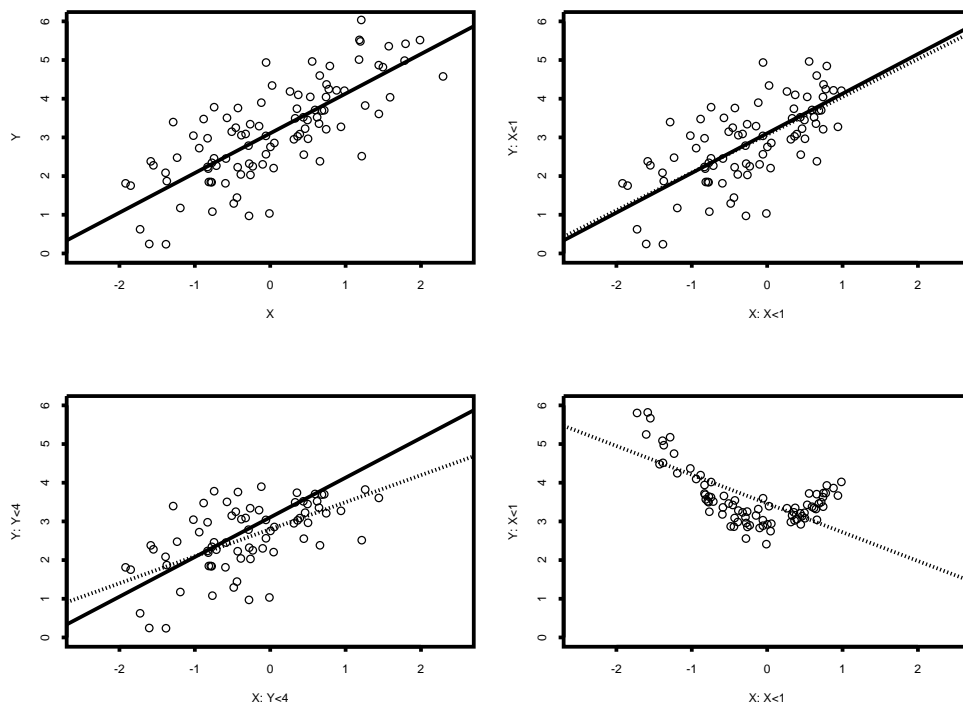


Figure 1. Regression Results for Different Types of Missing Data. Upper left=complete data; upper right=MAR but not OAR; upper left=not MAR; bottom right=incorrect functional form, but data are MAR.

The first plot (upper left corner) shows the fit of the regression model when all the data are observed. The second shows what happens to the fit of the regression model when the data are MAR but not OAR. For this case, I deleted all observations with X values greater than 1. Notice that the regression line is virtually identical to the one obtained with complete data. The bottom left plot shows what happens when the data are neither MAR nor OAR. For that case, I deleted all observations with Y values greater than 4. Observe that the regression line is biased considerably in this case. Finally, the last plot shows what happens

if the data are MAR, but an incorrect functional form is specified for the model. In that case, the model estimates will also be biased.

Although these results are somewhat encouraging in that they indicate that missing data may not always lead to biases, the fact is that it is practically impossible to assess whether data are MAR, exactly because the data are missing on the variable of interest! Furthermore, methods for handling data that are not MAR are statistically difficult and rely on our ability to correctly model the process that generates the missing data. This is also difficult to assess, because the data are missing!

3 Traditional approaches to handling missing data.

A variety of approaches to handling missing data have emerged over the last few years. Below is a list of some of them, along with a brief description. The bad news is that most of them are not useful when the data are not MAR, but rather, only when they are not OAR. This is unfortunate, because, as we discussed above, parameter estimates are not biased when the data are MAR but not OAR.

- **Listwise deletion.** Simply delete an entire observation if it is missing on any item used in the analyses.

Problems: Appropriate only when data are MCAR (or MAR). In that case, the only loss is statistical power due to reduced n . If the data are not MAR, then results will be biased. However, this is often the best method, even when the data are not MAR.

- **Pairwise deletion.** Delete missing variables by pairs. This only works if a model is estimated from covariance or correlation matrices. In that case, the covariances/correlations are estimated pairwise, so you simply don't include an observation that is missing data on one of the items in a pair.

Problems: Seems nice, but poor conceptual statistical foundation. What n should tests be based on? Also leads to bias if the data are not MAR.

- **Dummy variable adjustment.** Set missing values on a variable equal to some arbitrary value. Then construct a dummy variable indicating missingness and include this variable in the regression model.

Problems: This approach simply doesn't work and leads to biases in parameters and standard errors. It also has no sound theoretical justification—it is simply an *ad hoc* method to keep observations in an analysis.

- **Mean Imputation.** Replace a missing observation with the sample mean of the variable. When using longitudinal data, we can instead replace a missing score with the mean of the individual's responses on the other waves. I think this makes more sense than sample mean imputation in this case. An even better approach, I think, is discussed below under regression imputation.

Problems: Sounds reasonable but isn't. Whether the data are OAR or MAR, this approach leads to biases in both the standard errors and the parameters. The main reasons are that it shifts possible extreme values back to the middle of the distribution, and it reduces variance in the variable being imputed.

- **Hotdecking.** Replace a missing observation with the score of a person who matches the individual who is missing on the item on a set of other covariates. If multiple individuals match the individual who is missing on the item, use the mean of the scores of the persons with complete information. Alternatively, a random draw from a distribution can be used.

Problems: Seems reasonable, but reduces standard errors because it (generally) ignores variability in the x . That is, the x are not perfectly correlated, but this method assumes they are. The method also assumes the data are MAR. May be particularly difficult to implement with lots of continuous covariates. Also, the more variables used to match the missing observation, the better, but also the less likely you will be to find a match. Finally, what do you do with multiple missing variables? Impute and impute? Theoretically, you could replace all missing data through multiple passes through the data, but this would definitely produce overconfident and suspect results.

- **Regression-based Imputation.** Estimate a regression model predicting the missing variable of interest for those in the sample with complete information. Then compute predicted scores, using the regression coefficients, for the individuals who are missing on the item. Use these predicted scores to replace the missing data. When longitudinal data are used, and the missing variable is one with a within-individual pattern across time, use an individual-specific regression to predict the missing score.

Problems: This is probably one of the best simple approaches, but it suffers from the same main problem that hotdecking does: it underestimates standard errors by underestimating the variance in x . A simple remedy is to add some random error to the predicted score from the regression, but this begs another question: what distribution should the error follow? This method also assumes the data are MAR.

- **Heckman Selection Modeling.** The classic two-step method for which Heckman won the Nobel Prize involves a) estimating a selection equation ($I(\text{Observed}) = X\beta + e$), b) constructing a new variable—a hazard for sample inclusion—that is a function of predicted scores from this model ($\lambda_i = \frac{\phi(X_i'\beta)}{1 - \Phi(X_i'\beta)}$), c) including this new variable in the structural model of interest ($Y_i = Z_i'\gamma + \delta\lambda_i + u_i$).

Problems: This method is an excellent method for handling data that is *NOT* MAR. However, it has problems. One is that if there is a significant overlap between X and Z , then the method is inconsistent. Theoretically, there should be no overlap between X and Z , but this is often unreasonable: variables related to Y are also relevant to

selection (otherwise, there would be no relationship between observing Y and Y , and the data would thus be MAR). Another is that the standard errors in the structural model are incorrect and must be adjusted. This is difficult. Fortunately, STATA has two procedures that make this adjustment for you: Heckman (continuous structural outcome) and Heckprob (dichotomous structural outcome).

- **Multiple Imputation.** This method involves simulating possible values of the missing data and constructing multiple datasets. We then estimate the model using each new data set and compute the means of parameters across the samples. Standard errors can be obtained via a combination of the between-model variance and the within-model standard errors.

Problems: Although this approach adjusts for the downward bias that some of the previous approaches produce, its key drawbacks include that it assumes the data are MAR, and it is fairly difficult to implement. To my knowledge, there are some standard packages that do this, but they may be tedious to use. Another drawback is that you must assume some distribution for the stochasticity. This can be problematic as well.

- **The EM algorithm.** The EM algorithm is a two-step process for estimating model parameters. It integrates missing data into the estimation process, thus bypassing the need to impute. The basic algorithm consists of two steps: expectation (E step) and maximization (M step). First, separate the data into missing and nonmissing, and establish starting values for the parameters. In the first step, using the parameters, compute the predicted scores for the missing data (the expectation). In the second step, using the predicted scores for the missing data, maximize the likelihood function to obtain new parameter estimates. Repeat the process until convergence is obtained.

Problems: There are two key problems of this approach. One is that there is no standard software (to my knowledge) that makes this accessible to the average user. Second, the algorithm doesn't produce standard errors for the parameters. Other than these problems, this is the best maximum likelihood estimation has to offer. Note that this method assumes the data are MAR.

- **Direct ML Estimation.** Direct estimation involves factoring the likelihood function into components such that the missing data simply don't contribute to estimation of parameters for which the data are missing. This approach retains all observations in the sample and makes full use of the data that are observed.

Problems: Once again, this approach assumes the data are MAR. Not all likelihoods factor easily, and not many standard packages allow such estimation.

- **Bayesian modeling with MCMC methods.** Bayesian estimation is concerned with simulating distributions of parameters so that simple descriptive statistics can be used to summarize knowledge about a parameter. A simple example of a Bayesian model would be a linear regression model (assume for simplicity that σ_e is known). We

already know that the sampling distribution for the regression coefficients in a linear regression model has a mean vector equal to β and a covariance matrix $\sigma_e^2(X'X)^{-1}$. In an OLS model, the OLS estimate for β , $\hat{\beta}$ is found using $(X'X)^{-1}(X'Y)$. The central limit theorem tells us that the sampling distribution is normal, so we can say: $\beta \sim N((X'X)^{-1}(X'Y), \sigma_e^2(X'X)^{-1})$. That being the case, if we simply draw normal variables with this distribution, we will have a sample of parameters, from which inference can be made. (note: If σ_e is not known, its distribution is inverse gamma. The conditional distribution for the regression parameters is still normal, but the marginal distribution for the parameters will be t).

If there are missing data, we can use the assumption that the model is based on, namely that $Y \sim N(X\beta, \sigma_e^2)$, and integrate this into the estimation. Now, rather than simulating β only, we break estimation into two steps. After establishing starting values for the parameters, first, simulate the missing data using the assumption above regarding the distribution for Y . Second, given a complete set of data, simulate the parameters using the formula discussed above regarding the distribution of β . Although this approach seems much like the EM algorithm, it has a couple of distinct advantages. One is that standard errors (technically, the standard deviation of the posterior distribution for parameters) are obtained as a byproduct of the algorithm, so there is no need to come up with some additional method to do so. Another is that, in this process, a better estimate of uncertainty is obtained, given the missing data, than would be obtained using EM.

Problems: Bayesian approaches are not easy to implement, because there is very little packaged software in existence. Also, this method assumes the data are MAR. However, I note that we can modify the model and estimation slightly to adjust for data that aren't MAR. This can become complicated, though.

- **Selection and Pattern Mixture Models.** These approaches are somewhat opposite of each other. We have already dealt with one type of selection model (Heckman). In general, both approaches exploit the conditional probability rule, but they do so in opposite fashions. Pattern mixture models model $p(Y, Observed) = p(Y | Observed)p(Observed)$, while selection models model $p(Y, Observed) = p(Observed | Y)p(Y)$. We have already seen an example of selection. A pattern mixture approach would require us to specify a probability model for Y conditional on being observed multiplied by a model predicting whether an individual is observed. We would simultaneously need to model Y conditional on being unobserved by this model for being observed. This model is underidentified, because, without information on the Y that are missing, we do not know any characteristics of its distribution; thus, some identifying constraints are required.

Problems: The key problem with these approaches include that standard software does not estimate them (with the exception of Heckman's method). However, these are appropriate approaches when the data are not MAR.

4 A Simulation Demonstrating Common Approaches.

I generated 500 samples of size $n = 100$ each, consisting of 3 variables: X_1 , X_2 , and Y . $\rho_{X_1, X_2} = .4$, and the error term, u , was drawn from $N(0, 1)$. Y was computed using: $Y = 5 + 3X_1 + 3X_2 + u$. First, I estimated the regression model on all the samples with complete data. Second, I estimated the regression model on the samples after causing some to be missing. I used 4 different missing data patterns. First, I forced Y to be missing if $X_1 > 1.645$. This induces approximately 5% to be missing in each sample and generates samples in which the data are MAR but not OAR. Second, I forced Y to be missing if $X_1 > 1.037$. This induces approximately 15% to be missing. For comparison, I also estimated the models after making X_1 missing under the same conditions (rather than making Y missing). These results emphasize that the real problem occurs when the *dependent* variable is missing. Third, I forced Y to be missing if $Y > 12.0735$. Fourth, I forced Y to be missing if $Y > 9.4591$. These latter two patterns make the data violate the MAR assumption and generate approximately 5% and 15% missing data, respectively (the mean and variance for Y differs from that for the X variables). I estimated regression models using various approaches to handling the missing data. Below is a table summarizing the results of the simulation.

4.1 Simulation Results

Approach	Int.(Emp/Est S.E.)	X_1 (Emp/Est S.E.)	X_2 (Emp/Est S.E.)
No Missing	5.01(.100/.101)	3.01(.107/.111)	3.00(.112/.111)
X_1 missing if $X_1 > 1.645$			
Listwise	5.01(.103/.104)	3.01(.122/.125)	3.00(.115/.114)
Mean	5.31(.170/.176)	2.88(.157/.215)	3.31(.233/.187)
Dummy	5.01(.103/.105)	3.00(.122/.126)	3.01(.117/.114)
X_1 missing if $X_1 > 1.037$			
Listwise	5.01(.113/.116)	3.01(.140/.147)	3.00(.121/.121)
Mean	5.76(.219/.234)	2.82(.212/.314)	3.57(.271/.230)
Dummy	5.01(.113/.130)	2.99(.141/.163)	3.04(.126/.125)
Y missing if $X_1 > 1.645$			
Listwise	same as above		
Mean	4.57(.230/.210)	2.12(.368/.230)	2.86(.236/.231)
Dummy	5.00(.104/.122)	3.02(.133/.145)	2.88(.153/.130)
Y missing if $X_1 > 1.037$			
Listwise	same as above		
Mean	3.87(.321/.250)	1.28(.342/.275)	2.57(.316/.275)
Dummy	4.95(.130/.173)	2.97(.195/.213)	2.58(.227/.166)
Y missing if $Y > 12.0735$			
Listwise	4.96(.102/.107)	2.97(.119/.122)	2.96(.118/.121)
Mean	4.19(.257/.245)	2.07(.359/.269)	2.08(.341/.268)
Dummy	4.95(.104/.121)	2.90(.143/.134)	2.90(.143/.133)
Heckman	11.41(1.43/.314)	2.95(.128/.134)	lambda=-2.23
Y missing if $Y > 9.4591$			
Listwise	4.90(.118/.122)	2.93(.130/.135)	2.92(.130/.135)
Mean	4.10(.281/.267)	1.97(.391/.293)	1.98(.372/.292)
Dummy	4.75(.139/.164)	2.67(.197/.168)	2.67(.187/.168)
Heckman	9.57(.85/.274)	2.90(.132/.136)	lambda=-2.29

4.2 Summary of Results

The results indicate that, when the data are MAR, only violating the OAR assumption, listwise deletion only costs us efficiency. The dummy variable approach appears to work well, at least with little missing data. As the percent missing increases, the biases in the parameters and (especially) the standard errors begin to become apparent. Mean imputation performs very poorly in all cases. The biases appear to be most problematic when the data that are missing are the outcome data; missingness on the covariate is less troublesome.

In the results for the data that violate the MAR assumption, listwise deletion appears to work about as well as any approach. Once again, mean imputation performs very poorly, as does the dummy variable approach. Heckman selection works well, but the intercept and standard errors are substantially biased.

It is important to remember that this simulation study used better data than are typically

found in real datasets. The variables were all normally distributed, and the regression relationship between the independent and dependent variables was quite strong. These results may be more interesting if the signal/noise ratio were reduced (i.e., the error variance were increased).

5 Recommendations for handling missing data.

The guidelines below are my personal recommendations for handling missing data. They are based on a pragmatic view of the consequences of ignoring missing data, the consequences of handling missing data inappropriately, and the likelihood of publication/rejection using a particular method. These suggestions primarily apply when the outcome variable is missing and not the covariates. When covariates are missing, the consequences of missingness are somewhat less severe.

The standard method for reporting about the extent of missingness in current sociology articles is to a) construct a dummy variable indicating whether an individual is missing on an item of interest, b) conduct logistic regression analyses predicting missingness using covariates/predictors of interest, c) report the results: do any variables predict missingness at a significant level? There are at least two problems with this common approach: 1) what do you do if there is some pattern? Most people ignore it, and this seems to be acceptable. Some people make arguments for why any biases that will be created will be conservative. 2) this approach only demonstrates whether the data are OAR, which, as I said above, doesn't matter!!! I don't recommend this standard approach, but unfortunately, it is fairly standard.

Beyond simply reporting patterns of missingness, I recommend the following for dealing with missing data:

1. The first rule I recommend is to try several approaches in your modeling. If the findings are robust to various approaches, this should be comforting, not just to you, but also to reviewers. So, report in a footnote at least all the various approaches you tried. This will save you the trouble of having to respond to a reviewer by saying you already tried his/her suggestions. If the findings are not robust, this may point you to the source of the problem, or it may give you an idea for altering your research question.
2. If you have less than 5% missing, just listwise delete the missing observations. Any sort of simple imputation or correction may be more likely generate biases, as the simulation showed.
3. If you have between 5% and 10% missing, think about using either a selection model or using some sort of technique like multiple imputation. Here, pay attention to whether the data are missing on X or Y . If the data are missing on X , think about finding a substitute for the X (assuming there is one in particular) that is causing the problem. For example, use education rather than income to measure SES. Also consider whether the data are OAR or MAR. If you have good reason to believe the data are MAR, then just listwise delete but explain why you're doing so. If the data are not MAR, then you should do something beyond listwise deletion, if possible. If nothing else, try to

determine whether the missingness leads to conservative or liberal results. If they should be conservative, then make this argument in the text.

If the data are missing on Y , use either listwise deletion or Heckman selection. If you use listwise deletion, you need to be able to justify it. This becomes more difficult as the missing % increases. Use Heckman selection if you are pretty sure the data are not MAR. Otherwise, a reviewer is likely to question your results. At least use Heckman selection and report the results in a footnote indicating there was no difference between a listwise deletion approach and a Heckman approach (assuming there isn't!) Heckman is sensitive to the choice of variables in the selection equation and structural equation, so try different combinations. Don't have too much overlap.

4. If you have more than 10% missing, you must use a selection model or use some sophisticated technique for handling the missing, unless the data are clearly MAR. If they are, then either listwise delete or use some smart method of imputation (e.g., multiple, hotdecking, EM, etc.).
5. If you have more than 20% missing, find other data, drop problematic variables, get help, or give up.
6. If you have a sample selection issue (either on the independent or dependent variables), use Heckman selection. For example, if you are looking at "happiness with marriage," this item is only applicable to married persons. However, if you don't compensate for differential propensities to marry (and remain so), your parameters will be biased. As proof to the point, our divorce rates are higher now than they were in the past, yet marital happiness is greater now than ever before.

6 Recommended Reading

- Allison. (2002). *Missing Data* . Sage series monograph.
- Little and Rubin. (1987). *Statistical Analysis of Missing Data* .