Quant A&M Roundtable: Missing Data
Joseph Kush

Why is our data missing?

Missing data is ubiquitous in social science research. Here, we typically think of missing data as item non-response; this is when a unit (observation, person, etc.) fails to respond to some, but not all, items. We differentiate this from unit-non-response, which occurs when a unit does not respond to any items (e.g., Jane Doe was invited to take our survey, but for whatever reason, chose not to take the survey). Consider the following data:

Subject	Age	Gender	Income
1	29	M	$40,000
2	45	M	$36,000
3	81	M	--missing--
4	22	--missing--	$16,000
5	41	M	$98,000
6	33	F	$60,000
7	22	F	$24,000
8	--missing--	F	$81,000
9	33	F	$55,000
10	45	F	$80,000

Here, Subject 3 has a missing value on Income; Subject 4 has a missing value on Gender; and Subject 8 has a missing value on Age.

Our data may be missing for a number of reasons. Consider a longitudinal study, in which individuals complete an in-person test every month for a year. It is possible one subject was out of town on vacation for one of the test dates, or that another individual moved to a new state during that time. Attrition, in which observations drop out prior to the end of the study, is common in longitudinal studies. Data may be missing due to the item or question itself. For example, if an item is confusing or does not contain a response option that corresponds to how the individual would like to answer, the individual may choose not to answer that item. Likewise, if the item asks about sensitive information (e.g., number of times cheated on partner), individuals may be hesitant to respond. In short, data may be missing for a variety of reasons.

Why is missing data problematic?

When data is missing, it may bias our inferences. The observed values we do have may not necessarily be representative of the population we are interested in. For example, imagine only 2/100 people provided their income on a survey, and the average of these two values was $390,000. This may not be representative of the full $n = 100$ sample. Likewise, if the proportion of missing observations is so high

(98% in this example), we ultimately lose our sample size, reducing our statistical power. Some have suggested rules of thumb for when the amount of missingness is so large it becomes worrisome (e.g., < 5%). However, several authors (e.g., Enders, 2010) have demonstrated that the absolute amount of missing data is not the issue; it is the pattern of missing data that is the issue. Enders (2010) gives an example of how < 2% missing can result in 32% or more overall missing data depending on the pattern of missingness.

Types of missing data

Rubin (1976) outlined three types of missing data mechanisms.

1) Missing Completely at Random (MCAR)
2) Missing at Random (MAR)
3) Missing Not at Random (MNAR)

Missing Completely at Random (MCAR)
The missing value is not related to other variables in the model or on the variable itself. Under MCAR, the observed data can be thought of as a random sample of the complete data. If the reason for the missing data is truly random, then the means, variances, and covariances of the observed data do not differ from the corresponding moments of the complete data. Therefore, we can remove the cases with missing data and restrict the data analysis to complete cases. These results will be unbiased. It is important to note that MCAR is testable (Little, 1988). Think of this as a *t*-test between cases with and without missing data on a given variable. Essentially, a *p*-value greater than 0.05 indicates that there is not enough evidence that the data is not missing completely at random (and vice versa, a *p*-value less than 0.05 indicates there is enough evidence that the data does not meet the assumptions of MCAR, ultimately requiring more advanced techniques to handle).

Missing at Random (MAR)
The missing value is related to other variables in the model, but not on the variable itself (e.g., respondents with higher GPAs are less likely to report their income). It is important to note that MAR is not testable.

Missing Not at Random (MNAR)
The missing value is related to the variable itself (e.g., respondents with high income are less likely to report their income). It is important to note that MNAR is not testable.

Missing data methods: Listwise and pairwise deletion (bad)

Listwise deletion excludes an entire observation from analyses if the observation has a missing value on any variable. This may be referred to as complete-case analysis, as only observations that have observed values on all variables are included. Listwise deletion is the default way to handle missing data in most statistical software. Consider the following data:

Subject	Age	Gender	Income
1	29	M	$40,000
2	45	M	$36,000
3	81	M	--missing--
4	22	--missing--	$16,000
5	41	M	$98,000
6	33	F	$60,000
7	22	F	$24,000
8	--missing--	F	$81,000
9	33	F	$55,000
10	45	F	$80,000

If we wanted to calculate the average Age, using listwise deletion, Subjects 3, 4, and 8 would be excluded from the calculation (e.g., [29 + 45 + 41 + 33 + 22 + 33 + 45] / 7 = 35.4).

Pairwise deletion tries to mitigate the loss of data. Pairwise deletion uses any observation that has observed data on the variables you are interested in. For example, if we wanted to calculate the average Age, using pairwise deletion, only Subject 8 would be excluded from the calculation (Subjects 3 and 4 would be included in the calculation of Age, as they are not missing on this variable [though they are missing on other, unrelated variables]. If we wanted to calculate the correlation between Age and Income, using pairwise deletion, Subjects 3 and 8 would be excluded from the calculation (Subject 4 would be included, as there are observed values for the two variables we are interested in).

Under MCAR, listwise deletion is acceptable and will not bias results. However, MCAR is unlikely to be met in practice, indicating that the data are either MAR or MNAR. Again, note that MAR and MNAR are not testable. Under these two scenarios, neither listwise nor pairwise deletion methods are appropriate, and will bias results. So, if we cannot drop or delete our observations that have missing data, what should we do? One answer is we can try and guess what the missing value should be. This is called imputation, in which we plug in an estimated value for the missing value.

Missing data methods: Mean and regression imputation (also bad)

People have suggested ‘mean imputation’ as a strategy for guessing a missing value, in which the mean of all of the observed values on a given variable is computed and plugged in for missing values. The logic here is that if you don’t know what the value is, your best guess would be the mean (i.e., unlikely to be at the tails of the distribution, and, by definition, most likely to be at the mean). Consider the following data:

Subject	Age	Gender	Income
1	29	M	$40,000
2	45	M	$36,000
3	81	M	--missing--
4	22	--missing--	$16,000
5	41	M	$98,000
6	33	F	$60,000
7	22	F	$24,000
8	--missing--	F	$81,000
9	33	F	$55,000
10	45	F	$80,000

For subject 8, our mean imputation would be to calculate the mean age of all observed values (= 39), and plug that value in as our best guess for Subject 8. Mean imputation is not a valid approach, and has been shown to lead to biased estimates (standard error estimates will be downwardly biased, as additional values at the mean reduce overall variance).

Imagine the following two observations with missing income values:

	ID	income	gpa	highest_ed	sector	zip_income	make	model
1	A	.	3.9	Ph.D.	Law	250	Tesla	Model X
2	B	.	1.7	H.S.	Retail	38	Honda	Civic

We see that observation A had a high school GPA = 3.9; the highest education received was obtaining a Ph.D.; works in the sector of law; lives in a ZIP code where the average income is $250,000; and drives a Tesla Model X.

We see that observation B had a high school GPA = 1.7; the highest education received was obtaining completing high school; works in the sector of retail; lives in a ZIP code where the average income is $38,000; and drives a Honda Civic.

Instead of plugging in the mean income value for these two observations, can we somehow use the other information we do have for these two observations to inform our guess as to what the missing value could be? Regression imputation attempts to do just that. Basically, using all of the other observations that have non-missing data, we could estimate the following linear regression model:
income = intercept + b1*gpa + b2*highest_ed + … + b6*model
and use the predicted score from the regression equation as the value to plug in. In essence, if the observed data showed that those with higher GPAs and those who lived in ZIP codes with higher average incomes also had higher incomes, we might predict a larger income value for ID = A than for ID = B. Regression imputation is not as bad as listwise deletion, pairwise deletion, or mean imputation, but it is still not great (overestimates correlations, standard error estimates will be downwardly biased).

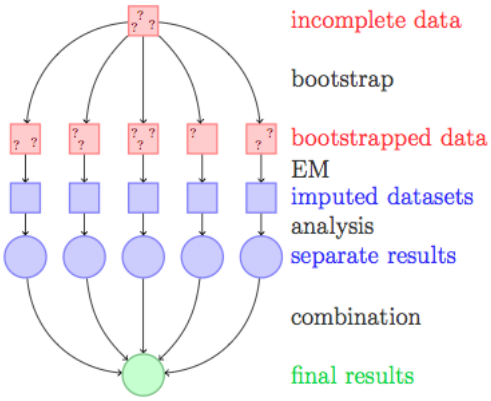
Missing data methods: Multiple imputation (good)

Instead of using regression imputation to predict a (single) value for the missing value, we can think about making numerous guesses for what the missing value should be. Multiple imputation uses the following three steps:

1) Impute *m* (e.g., 5) datasets. Here, each dataset is based on a resampling of the original sample. This mimics random selection of another sample from the population.

2) Run your analyses *m* times. Here, you calculate the desired statistic or parameter estimate for each of the analyses.

3) Pool your parameter estimates and standard errors from the *m* analyses. Here, you report the averages across the *m* analyses.



Two modern techniques used to conduct multiple imputation include multivariate normality and multivariate imputation using chained equations (MICE). The multivariate normality approach assumes that the missing values lie somewhere on a normally distributed multivariate distribution, and likely values are pulled from this distribution and plugged in as guesses. The MICE approach is similar to Bayesian analysis, in which an algorithm (e.g., Gibbs sampler) is used to impute variables sequentially using fully conditional specifications. For more information, you can read: https://www.stata.com/manuals/mi.pdf. Multiple imputation has been shown to produce unbiased estimates under MAR.

Missing data methods: Maximum likelihood estimation (good)

Maximum likelihood estimation (MLE) is a method used to estimate parameters. For example, we typically use ordinary least squares (OLS) to estimate a linear regression model, in which the goal of OLS is to find parameter values that minimize the sum of squared residuals. For this model (or many others), we could also choose to estimate our parameters using maximum likelihood estimation. The goal of MLE is to determine parameter values for which the observed data have the highest (joint) probability (i.e., to maximize the likelihood function). Full information maximum likelihood (FIML) attempts to estimate parameters for all observations, including those with missing values. FIML is often referred to as a model-based approach to missing data, as the missingness is addressed directly in the estimation of the model. To start, sufficient statistics (i.e., mean and variance-covariance matrix) are estimated from the raw, incomplete data via the expectation maximization algorithm. Those estimates then serve as the starting values for the maximum likelihood model estimation. For more

information, you can watch: https://www.youtube.com/watch?v=XepXtl9YKwc. FIML has been shown to produce unbiased estimates under MAR.