

# Regression III

## Lecture 13: Missing Data and Multilevel Models

Dave Armstrong

University of Wisconsin – Milwaukee  
Department of Political Science

e: [armstrod@uwm.edu](mailto:armstrod@uwm.edu)  
w: [www.quantoid.net/ICPSR.php](http://www.quantoid.net/ICPSR.php)

1 / 23

### Why Care about Missing Data?

- Can cause bias, if missingness is systematic.
- Can reduce sample size to a point where reliable inferences are difficult to make, even if missingness is not systematic.
- Systematic missingness can truncate the sample calling into question the generalizability of results.

While studies can get published without handling missing data in an interesting way, this is becoming less viable. Further, dealing with the problem proactively can be a good signal to the reviewer.

2 / 23

### Types of Missing Data

- Unit missing - whole observation is missing (i.e., someone who hangs up the phone before any questions are answered).
- Missing value - when a particular question is not answered by a respondent who is otherwise participating in the survey.
- Missing wave - unit missing for a wave of a longitudinal study.

3 / 23

### Missing Data and the Scientific Process

Missing data causes problems with three steps in the scientific process:

1. Make systematic, structured observations
  - Missing data can affect the quality (reliability and validity) of our observations.
2. Use data to make inferences
  - Missing data can affect the validity of our conclusions about relationships (i.e., our conclusions could be biased or inefficient)
3. Generalize
  - Missing data can limit the representativeness of the sample and thus limit the extent to which our results can be generalized.

4 / 23

## Consequences for Data Analysis

The effects on data analysis are the ones most commonly acknowledged by the literature on missing data.

- Missing data, at a minimum, can pose problems for statistical power (i.e., sample size).
- Statistical procedures also make assumptions about distributions (i.e., error distributions). Missing data can make some of these assumptions less likely to hold, especially if the missingness is not random.
- Missing data can also, as previously stated, reduce reliability, which reduces effect size, which in turn reduces statistical power.

5 / 23

## Consequences for Internal Validity

Internal validity can be defined as - the extent to which a researcher can reasonably claim that a particular factor, usually an intervention of some sort, is responsible for the observed outcome. Confounders and alternative explanations are threats to internal validity. *Selection bias* is an example:

- There can be systematic differences between completers and dropouts in experiments/surveys. These differences could be responsible for the outcome rather than the variable of interest.
- This can lead to a more homogenous sample that is not representative of the population which can cause all sorts of problems for generalizability.
- The same can be true of missingness on particular variables.

6 / 23

## Missing Data in Experiments and Observational Studies

- Randomization in experimental trials is known to, on average, produce groups that are roughly evenly distributed on potential confounders.
  - Missing data reduces the sample size leaving the groups perhaps less likely to equal distributions on confounders.
  - If the two groups are differentially missing, this can cause selection-bias type effects.
- In observational studies, randomization into treatment and control groups is not possible. In these situations, we use statistical controls to attenuate the effect of confounders and alternative explanations.
  - Observations missing information on some of the controls can create the same problem as in the experimental setting.

7 / 23

## Missing Data and Generalization

Missing data can also cause problems at the level of theory building and testing (i.e., global or macro problems).

- The scientific enterprise is one where useful theories live and less useful theories die based on the accumulation of knowledge. This can cause two problems.
  1. In areas where the same flawed research design propagates, the result of continued, iterative improvement is only to find a local maximum.
  2. When studies with missing data show weak relationships, theories might be in danger of being dismissed when complete data would support the theory.

8 / 23

## Rubin's Classification Scheme

Rubin suggested three types of missing data - missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). This classification scheme is based on:

1. The variables with the missing data,
2. Associated variables (i.e., covariates), and
3. a hypothetical mechanism underlying the missingness.

9 / 23

## Background on Rubin's Classification

- Refer to  $\mathbf{R}$  as the matrix of dummy variables that mirrors the data matrix where 1 indicates missing and 0 indicates non-missing.
- Refer to  $\mathbf{Y}$  as the data matrix - the matrix of variables for all observations. Where  $\mathbf{Y}_{obs}$  refers to the observed (i.e., non-missing) values and  $\mathbf{Y}_{miss}$  as the missing values.
- $\phi$  is the relationship of the observed variable matrices  $\mathbf{Y}_{obs}$  and  $\mathbf{Y}_{miss}$  to the dummy variable matrix  $\mathbf{R}$ .  $\phi$  is probabilistic (i.e., theoretical) here because we don't know the values of  $\mathbf{Y}_{miss}$  and thus can never calculate  $\phi$ .
- $\phi$  is the operative piece of information in Rubin's classification scheme.

10 / 23

## Classification of Missing Data Mechanisms

- If  $\phi = \mathbf{0}$  (i.e., there is no relationship between  $\mathbf{R}$  and  $\mathbf{Y}_{obs}$ , and no relationship between  $\mathbf{R}$  and  $\mathbf{Y}_{miss}$ ), then the data are MCAR. Here, randomness is the mechanism that generated the missing data.
- Data are MAR if there is a relationship between  $\mathbf{R}$  and  $\mathbf{Y}_{obs}$ , but there is no relationship between  $\mathbf{R}$  and  $\mathbf{Y}_{miss}$ .
- Data are MNAR if there is a relationship between  $\mathbf{R}$  and  $\mathbf{Y}_{miss}$ . The relationship between  $\mathbf{R}$  and  $\mathbf{Y}_{obs}$  is irrelevant here - it may or may not exist. Note, that this is an impossible distinction for us to make with data.

"Ignorability" is a property of MAR and MCAR data. Here, the mechanism is ignorable if we don't have to model the missing data mechanism. Missingness is ignorable, if we can reasonably recover that information from other observable data.

- If survey respondents fail to respond to some questions because of their positions on the questions, this is MNAR, and thus non-ignorable.

11 / 23

## Dealing with Missing Data

Two ways of dealing with missing data:

- Listwise Deletion
- Imputation
  - Mean imputation
  - Regression Imputation
  - Hot-decking
  - Multiple Imputation

12 / 23

## Listwise Deletion

Listwise deletion involves deleting all observations that have at least one missing data point.

- In the best-case scenario, listwise deletion causes inefficiency.
- In the worst-case scenario, listwise deletion causes bias and inefficiency.

Often times, omitting variables with missing data can be preferable to listwise deletion in MSE terms (though other methods talked about later are better)

13 / 23

## Mean Imputation

Both mean and regression imputation are trying to impute a “best guess” for missing data.

Mean imputation imputes the unconditional mean of the variable for every missing observation.

- Mean imputation reduces variability in the offending  $X$  variable.
- As a result, coefficient estimates will be biased (generally toward 0).
- The variance of the coefficients will also be underestimated.

14 / 23

## Regression Imputation

In regression imputation, the complete cases are used to estimate a regression model and predictions from that model are used as imputations.

That is, we are imputing each observation with its *conditional mean*.

- Impractical/impossible with complicated patterns of missing data.
- Does a reasonably good job of recovering unbiased estimates of parameters in a wide range of situations.
- Drastically underestimates variability in the parameters leading to overconfidence.

15 / 23

## Hot-deck Imputation

Hot-deck imputation, as often done, has a matching (or matching-like) component to it.

- Observations are matched to other observations with similar values on non-missing variables.
- The complete data from the matched case are imputed to the case with missing values.

Hot-decking has some nice properties:

- It is relatively efficient
- It preserves the distribution of observed responses

Hot-decking generally suffers from the same problems of overconfident inferences as the other ad hoc methods (Cranmer and Gill, notwithstanding).

16 / 23

## Multiple Imputation

MI has a similar flavor to regression imputation, but has more appealing properties. Multiple imputation proceeds as follows:

1. Fill in starting values for all observations (random, mean, etc...)
2. Predict  $X_1$  using all other variables in estimation model (including  $y$ ) and fill in the missing observations on  $X_1$  with a draw from the posterior (i.e., sampling distribution) of  $\hat{X}_1$ .
3. Using the previously predicted values for  $X_1$ , move to  $X_2$  and predict its values using all other variables. Fill in the missing values on  $X_2$  with a draw from the posterior of  $\hat{X}_2$ .
4. Move through all the variables similarly and then start over again using the conditionally imputed values from before as starting values.
5. Repeat until convergence.

17 / 23

## More MI

- With multiple imputation, we impute  $m \geq 2$  and usually between 5 and 10 complete datasets.
- We combine the estimates using the following set of equations:

$$\bar{Q} = \frac{1}{m} \sum_{t=1}^m \hat{Q}^{(t)} \quad (1)$$

$$\bar{U} = \frac{1}{m} \sum_{t=1}^m U^{(t)} \quad (2)$$

$$B = \frac{1}{m-1} \sum_{t=1}^m (\hat{Q}^{(t)} - \bar{Q})(\hat{Q}^{(t)} - \bar{Q})' \quad (3)$$

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B \quad (4)$$

18 / 23

## Sampling Distribution of $\bar{Q}$

The distribution of the test statistic is harder than you might think to derive. If we assume that proportion of missing information for every variable is the same, then, we can say:

$$\tilde{T} = (1 - r_1) \bar{U} \quad (5)$$

$$r_1 = \left(1 + \frac{1}{m}\right) \frac{tr(B\bar{U}^{-1})}{k} \quad (6)$$

The test-statistic can be calculated against the null  $Q_0$  as:

$$D_1 = \frac{(\bar{Q} - Q_0)' \tilde{T}^{-1} (\bar{Q} - Q_0)}{k} \quad (7)$$

with a p-value of  $P(F_{k, \nu_1} \geq D_1)$  where:

$$\nu_1 = 4 + (t - 4) \left[1 + \left(1 - t2^{-1}\right) r_1^{-1}\right]^2 \quad (8)$$

19 / 23

## More on the Sampling Distribution

- This assumes that the percentage of missing information for each coefficient is the same.
  - Simulation studies show that the results are relatively robust to deviations from this assumption.
  - Deviations generally result in conservative p-values.
- This also assumes that a  $\chi^2$  approximation for testing a single point estimate is correct (see Schafer 1997, Chapter 4, for a discussion)

20 / 23

## Development of MI Algorithms

- Initially, MI software used a multivariate normal approximation to impute the missing values.
  - This has been shown to work relatively well even if normality is the wrong theoretical model (e.g., when trying to impute dummy variables).
  - If using software that does this, you need to transform variables to be theoretically unbounded and recode ordinal variables to include approximately cardinal information.
  - Gary King and others' Amelia II uses this approximation.
- Other methods (e.g., MICE) respect the level of measurement of the missing variables and use different regression techniques to impute the values.
  - Transformations are still appropriate here because the underlying model for continuous data is still linear.
  - No need to recode ordinal data.
  - mice in R uses this technique.

21 / 23

## Some other Advice

- All variables (including  $y$ ) that are in your analysis model should be in the imputation model (otherwise, potential bias results).
  - Extra variables can be included in the imputation model if they are relevant (results in "super-efficiency", and who wouldn't want that)
  - Also needs to include non-linear trends if they exist (e.g., polynomials).
- It has been shown that somewhere in the neighborhood of 5-10 complete datasets are sufficient to generate the result.

22 / 23

## Missing TSCS Data

- Honaker and King (2010) show that MICE-type procedures tend to systematically miss trends in the data.
- Amelia II has procedures to deal with TSCS missingness.
  - Essentially amounts to putting time-trends in the imputation models, either through polynomials or cubic-splines in time.
  - Can allow for unit-specific time-trends as well by interacting time polynomials with categorical indicator of group membership.
- Still assumes a MAR mechanism.
  - If observations drop out of the sample because of values on the dependent variable (e.g., efficacy of treatment, relapse, etc...), the data are still NI and you will need to build an imputation model (see Schafer sec. 2.5.3 for a brief discussion and citations).

23 / 23