Introduction to Missing Data Part 2

Jarvis T. Chen (jarvis@hsph.harvard.edu)

13 April 2023

Department of Social and Behavioral Sciences
Harvard T. H. Chan School of Public Health

Last time...

We discussed:

- Missing data mechanisms:
  - Missing Completely At Random (MCAR)
  - Missing At Random (MAR)
  - Missing Not At Random (MNAR or NMAR)
- Recall that MCAR is a special case of MAR.
- We generally do not have the information in a given dataset to distinguish MAR from MNAR.

Last time...

We started to build some intuition about missing data:

- Bias can occur because we only get to observe a selected sample of the population of interest, and conditioning on being observed induces correlation among common causes of being missing ("collider stratification").
- When the subjects with missing covariate values differ systematically from those with complete data with respect to the outcome of interest, results from a traditional data analysis omitting the missing cases may be biased.
- If subjects with any missing covariates are excluded from the analysis (complete case analysis), analyses can be **biased** and **inefficient**.
- However, note that if the prevalence of missing data is $< 5\% - 10\%$, the amount of bias is limited and may be negligible.

Ad hoc missing data handling methods

- Complete case analysis (valid for MCAR but usually not for MAR or MNAR)
    - whether or not there will be bias depends on the target estimand of interest and the estimator
- Missing indicator method (**not recommended!**)
- Last Observation Carried Forward (**not recommended!**)

Statistically Principled Methods

- Incorporate statistical (stochastic) information about the missing values and/or the missingness mechanism, e.g.
  - **Weighting** – weight the observed data by the inverse of the probability of being observed to "recover" the population we would have observed if missingness had not occurred.
  - **Multiple imputation (MI)** – generate $K > 1$ imputed values for missing observations from appropriate probability distribution.

An additional note on weighting

- Robins et al. (1994) show that there is a class of estimators that involves *augmenting* the simple inverse probability weighted complete case estimating equations. The augmented inverse probability weighted estimator is relatively more efficient than the simple IPW estimator and also has the property of **double robustness**.
    - Doubly robust estimators are consistent estimators if either the model for missingness is correctly specified OR the model for the outcome is correctly specified (or both). Where doubly robust, augmented inverse probability weighted estimators are feasible in practice, these offer protection against misspecification of these models. However, in all but the simplest settings, they can be challenging to implement.

Here is an accessible reference: Vansteelandt, S., Carpenter, J., & Kenward, M. G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6(1), 37–48. https://doi-org.ezp-prod1.hul.harvard.edu/10.1027/1614-2241/a000005

Multiple Imputation

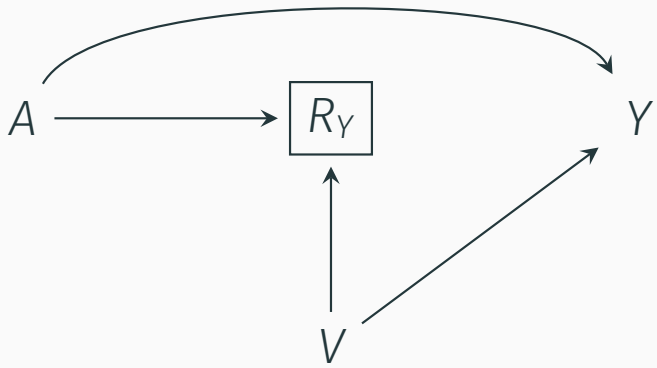
- Overview of the method
- How to combine results over imputations
- What should be included in the imputation model?
- How many imputations?
- How to do the imputations?

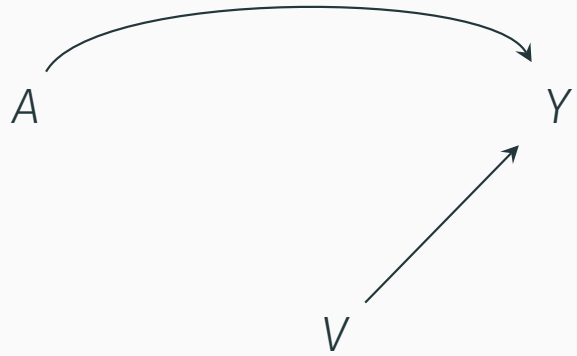
Multiple imputation

- Probably the most common **statistically principled** method for dealing with missing data
- Replace the unknown missing data Z_{mis} with **multiple** simulated values, $Z_{mis}^{(1)}, Z_{mis}^{(2)}, \ldots, Z_{mis}^{(m)}$.
- Each of the m completed datasets is analyzed by standard complete-data methods.
- The variability among the results of the m analyses provides a measure of the uncertainty due to missing data
- Combined with within-imputation measures of uncertainty from ordinary sample variation, allows us to make inferential statements about the parameters of interest.

Why multiple imputation?

Recall the statistician's concern: "Can I recover the **joint distribution** of all of the variables that we would have observed in a population where missingness never happened?"

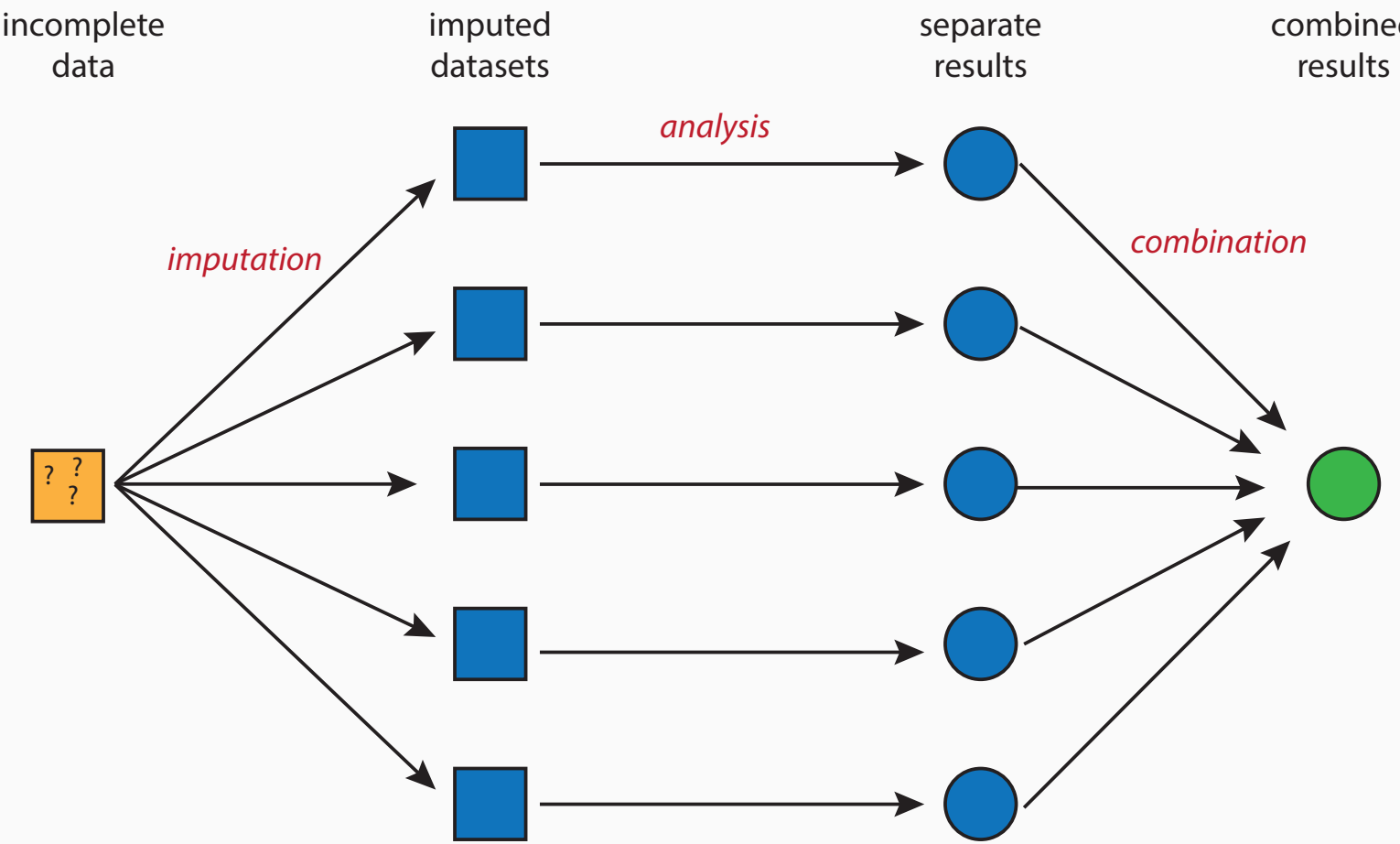




Multiple imputation: steps

1. Construct m "complete" datasets by imputing the missing values (usually under MAR).
2. Perform a standard complete data analysis within each of the m datasets and obtain an estimate of interest, e.g. $\hat{\beta}^{(k)}$, $k = 1, \ldots, m$.
3. Combine the estimates over the imputed datasets and compute the variance taking into account the between-imputation variability.

Multiple imputation: overview



Combining complete data inferences

Let $\hat{\beta}$ be the complete data point estimate for β, the estimate of our parameter of interest that we would use if no data were missing. Let $\hat{V}_\beta$ be the variance estimate associated with $\hat{\beta}$, so that $\sqrt{\hat{V}_\beta}$ is the complete-data standard error. With m imputations, we can calculate m different versions of $\hat{\beta}$ and $\hat{V}_\beta$.

Let

$$\hat{\beta}^{(k)} = \hat{\beta}(Y_{obs}, Y_{mis}^{(k)})$$

and

$$\hat{V}_\beta^{(k)} = \hat{V}_\beta(Y_{obs}, Y_{mis}^{(k)})$$

be the point and variance estimates using the kth set of imputed data, $k = 1, \ldots, m$. Rubin (1987, Chapter 3) gives the following rule for combining them.

Rubin's Rules

The multiple-imputation point estimate for β is simply the **average** of the complete-data point estimates,

$$\bar{\beta} = \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}^{(k)}$$

The variance estimate associated with $\bar{\beta}$ has two components. The **within-imputation variance** is the average of the complete-data variance estimates,

$$\bar{V}_\beta = \frac{1}{m} \sum_{k=1}^{m} \hat{V}_\beta^{(k)}$$

The **between-imputation variance** is the variance of the complete-data point estimates,

$$B = \frac{1}{m-1} \sum_{k=1}^{m} (\hat{\beta}^{(k)} - \bar{\beta})^2$$

Rubin's Rules

The **total variance** is defined as

$$T = \bar{V}_\beta + (1 + m^{-1})B$$

Note that this expression for the total variance combines two sources of variability: the within-imputation variability (V_β) and the between imputation variability (B).

Inferences are based on the t-distribution

$$\frac{\beta - \bar{\beta}}{\sqrt{T}} \sim t_\nu$$

where the degrees of freedom are given by

$$\nu = (m-1)\left[1 + \frac{\bar{V}_\beta}{(1+m^{-1})B}\right]^2$$

Thus, a $100(1-\alpha)\%$ interval estimate for $\bar{\beta}$ is

$$\bar{\beta} \pm t_{\nu,1-\alpha/2}\sqrt{T}$$

This is of course implemented in software, but it's worth knowing how to do it by hand if you are ever doing multiple imputation on a desert island!

Missing information

Note that the degrees of freedom depend not only on m, but also on the ratio

$$r = \frac{(1 + m^{-1})B}{\bar{V}_\beta}$$

- Rubin (1987) calls r the *relative increase in variance due to non-response*, because $\bar{V}_\beta$ represents the estimated total variance when there is no missing information about β (i.e. when $B = 0$).
- When m is large and/or r is small, the degrees of freedom will be large and $\frac{\beta - \bar{\beta}}{\sqrt{T}} \sim t_\nu$ will be approximately normal.

Missing information

We can also estimate the fraction of missing information about β as

$$\begin{aligned}\hat{\lambda} &= (\bar{V}_\beta^{-1} - (\nu+1)(\nu+3)^{-1}T^{-1})\bar{V}_\beta \\ &= \frac{r + 2/(\nu+3)}{r+1}\end{aligned}$$

Schafer (1997) recommends calculating r and $\hat{\lambda}$ as "interesting and useful diagnostics for assessing how the missing data contribute to inferential uncertainty about β."

What to include in the imputation model?

Multiple imputation was originally conceived in the survey setting where there is a distinction between the imputer (database constructor) and the analyst (data user). Often the imputer has access to more detailed (potentially confidential) information with which to predict the missing values.

In many non-survey settings, the imputer and the analyst are the same person.

What to include in the imputation model?

Meng (1994) found that so long as the imputation model includes all the variables (and information) in the analysis model, no bias is introduced. Nominal confidence interval coverage will be at least as great as actual coverage and equal when the two models coincide.

The key idea is that the **imputation model should be at least as rich as the analytic model**.

Why only a few imputations are needed

- Rubin proposed that a relatively small number of imputations ($m = 5$ or $m = 10$) is sufficient to obtain relatively efficient estimates in a multiple imputation analysis.
- If the fraction of missing information about a scalar estimand is λ, the relative efficiency (in units of standard errors) of a point estimate based on m imputations to one based on an infinite number of imputations is approximately $(1 + \lambda/m)^{-1/2}$ (Rubin, 1987, p. 114).
- E.g. if $\lambda = 0.2$, the relative efficency of an estimate based on $m = 3$ imputations will tend to have a standard error only $\sqrt{1 + 0.2/3} = 1.033$ times as large as the estimate with $m = \infty$. With $\lambda = 0.5$, an estimate based on $m = 5$ imputations will tend to have a standard error only $\sqrt{1 + 0.5/5} = 1.049$ as large.

Why only a few imputations are needed

Table 1: Large sample relative efficiency when using a finite number of proper imputations, m, rather than an infinite number, as a function of the fraction of missing information (Rubin 1987, p. 114)

	Fraction of missing information								
m	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.95	0.91	0.88	0.85	0.82	0.79	0.77	0.75	0.73
2	0.98	0.95	0.93	0.91	0.89	0.88	0.86	0.85	0.83
3	0.98	0.97	0.95	0.94	0.93	0.91	0.90	0.89	0.88
4	0.99	0.98	0.96	0.95	0.94	0.93	0.92	0.91	0.90
5	0.99	0.98	0.97	0.96	0.95	0.94	0.94	0.93	0.92
$\vdots$									
∞	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Number of imputations

- More recently, Graham et al. (2007) have shown that effects of m on **statistical power for detecting a small effect size** can be strikingly different from what is observed for **relative efficiency**.
- They showed that if statistical power is the main consideration, the number of imputations typically must be much higher than previously thought.
- For example, Graham et al. (2007) recommend that at least $m = 40$ imputations are need with 50% missing information to guarantee less than a 1% power fall off compared to the comparable FIML analysis.
- Unless your dataset is really enormous, creating and storing $m = 40$ imputed datasets should not be a huge burden.

How to get the imputations

- There are a number of different methods for imputing the missing values. The choice of method is often determined by the patterns of missingness (e.g. monotone versus non-monotone missingness) in the dataset at hand as well as the types of variables (continuous, dichotomous, categorical, mixed).
- The overarching principle is that "proper" imputations should be drawn at random from the conditional (predictive) distribution of the missing data given the observed data.
- In general, a proper imputation of Z^{mis} is obtained by randomly drawing values from $f(Z^{mis}|Z^{obs})$. By choosing values from $f(Z^{mis}|Z^{obs})$, we are implicitly assuming that missingness is MAR.
- To help us build some intuition about how to generate imputations for the missing values, let's consider the very simple situation where we have one continuous variable that is partially missing.

Regression imputation

Regression imputation replaces missing values by predicted values from a regression of the missing item on items observed for the unit, usually calculated from units with both observed and missing variables present.

For example, using the Six Cities example, let's assume that we have fully observed data on $Y = log.FEV1$ and $X_2 = age$ and partially observed data on $X_1 = height$. We fit an imputation model for height given age and log.FEV1 among the fully observed data,

$$\mathbb{E}(X_1) = \gamma_0 + \gamma_1 X_2 + \gamma_2 Y$$

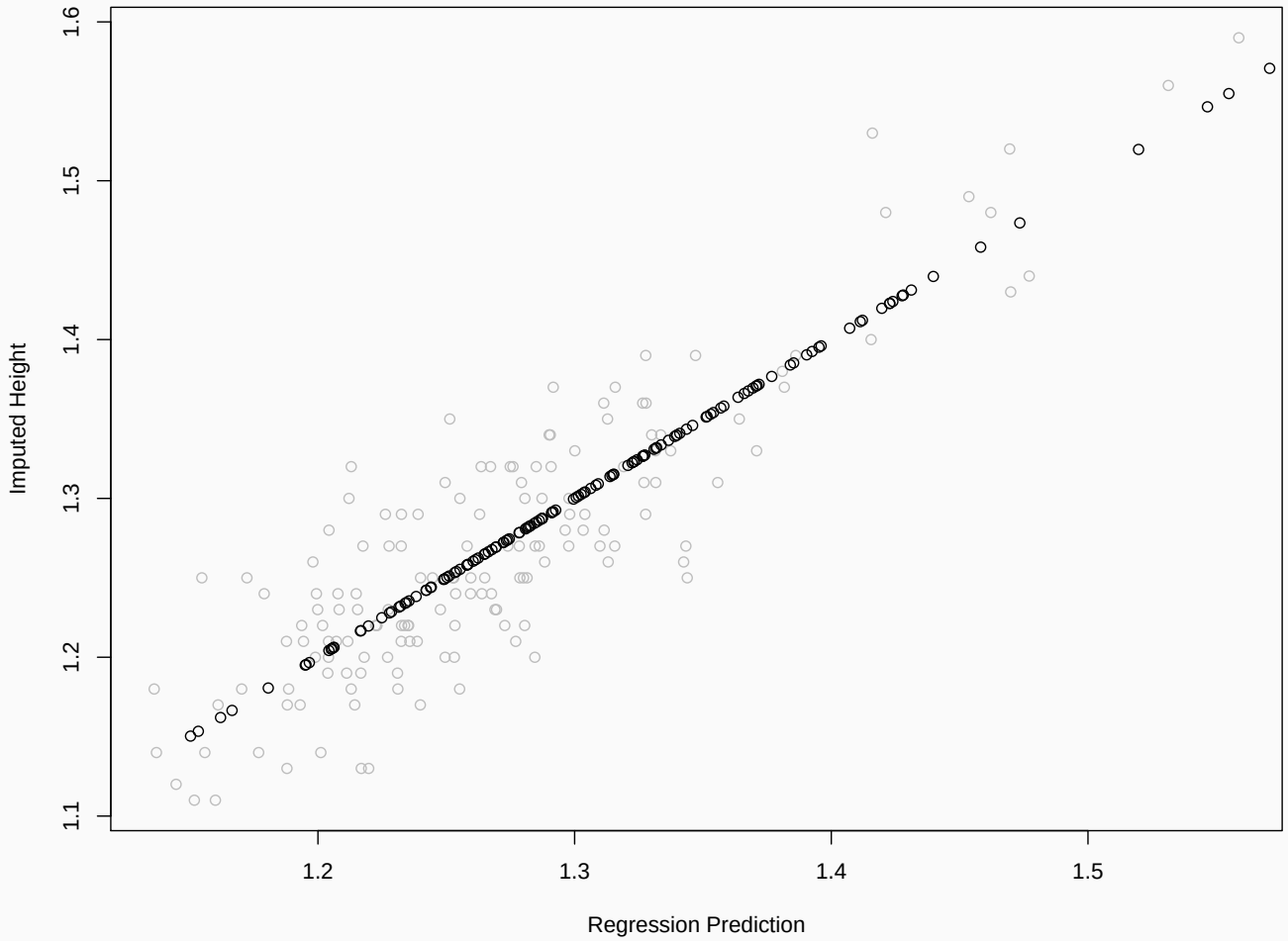
and we get

$$\hat{X_1} = 0.9 + 0.04 * age + 0.19 * log.FEV1$$

We use this to get predicted values of height for all those missing X_1.

Regression imputation: example

The variance of X_1 in the completed dataset is too low, because the imputed values all come from the fitted regression line.



Stochastic regression imputation

- Stochastic regression imputation replaces missing values by a value predicted by regression imputation plus a residual, drawn to reflect uncertainty in the predicted value.
- With normal linear regression models, the residual will naturally be normal with zero mean and variance equal to the residual variance in the regression.
- With a binary outcome, as in logistic regression, the predicted value is a probability of 1 versus 0, and the imputed value is a 1 or 0 drawn with that probability.

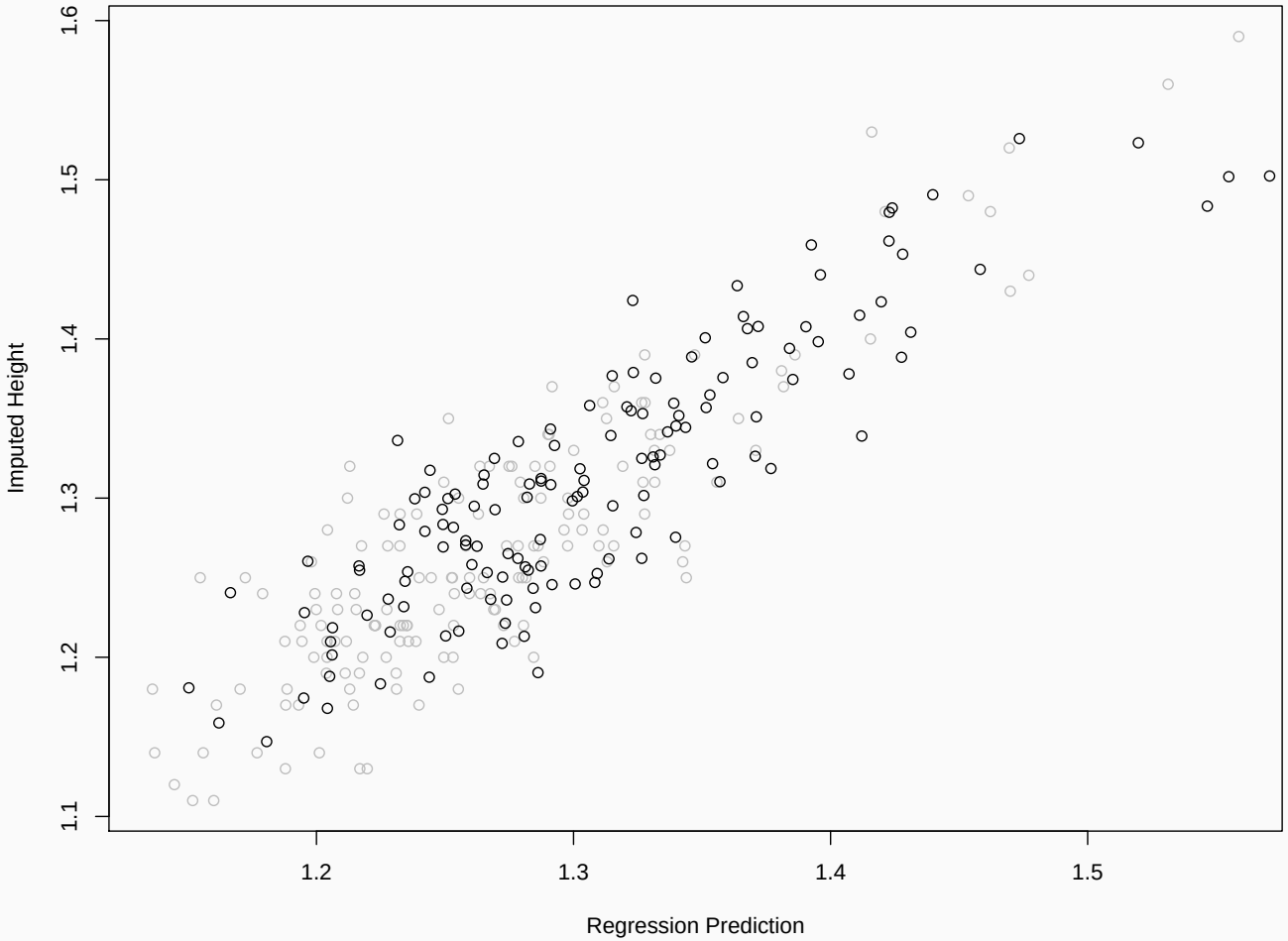
Stochastic regression imputation: example

In our Six Cities example, if we want to add some random error into our imputation model, we can generate predicted values of X_1 with

$$\hat{X_1} = 0.9 + 0.04 * age + 0.19 * log.FEV1 + 0.04 * \epsilon$$

where ϵ is a random draw from a standard normal $\sim N(0, 1)$ and 0.04 is the square root of the error variance from the imputation model.

Stochastic regression imputation: example



‘Proper’ imputation

- There is an additional source of uncertainty that needs to be reflected!
- We have not taken into account the fact that the parameters in the imputation model $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\sigma}^2)$ are only estimates with their own sampling variability.
- Because we do not know precisely what the population values are, when we act as if we have this knowledge we underestimate variability.
- To incorporate this additional source of variation, each imputation should be based on different (i.e. randomly perturbed) values for the regression coefficients and σ^2. Specifically, these values should be random draws from their posterior distribution. (Fortunately, this is implemented in software packages, so we don’t actually have to do this by hand).

Predictive mean matching (PMM)

Suppose there is a single variable X that has some cases with missing data, and a set of variables Z (with no missing data) that are used to impute X.

1. For cases with no missing data, estimate a linear regression of X on Z, producing a set of estimated coefficients $\widehat{\beta}$.
2. Make a random draw from the "posterior predictive distribution" of β, producing a new set of coefficients β^*.
    - e.g. take a random draw from a multivariate normal distribution with mean β and estimated covariance matrix of β (with an additional random draw for the residual variance).
3. Using β^*, generate predicted values for X for **all** cases (both missing and non-missing on X)

Predictive mean matching (PMM)

4. For each case with missing X, identify a set of cases with observed X whose predicted values are close to the predicted value for the case with missing data.
5. From among those close cases, randomly choose one and assign its observed value to substitute for the missing value. Repeat steps 2 through 5 for each completed data set.

Note that unlike regression imputation, the purpose of the linear regression is not to actually generate imputed values. Rather, it serves to construct a metric for matching cases with missing data to similar cases with data present.

Imputation in the real world

Typical problems that may arise include:

- For a given Z_j, predictors Z_{-j} used in the imputation model may themselves be incomplete.
- Circular dependence can occur, where Z_1 depends on Z_2 and Z_2 depends on Z_1 because in general Z_1 and Z_2 are correlated, given other variables.
- Especially for large numbers of variables and small n, collinearity and empty cells may occur.
- Rows and columns can be ordered, e.g. as with longitudinal data.
- Variables may be of different types (e.g. binary, unordered, ordered, continuous)

Imputation in the real world

- The relationship between Z_j and Z_{-j} could be complex, e.g. non-linear, or subject to a censoring process.
- Imputation can create impossible combinations or destroy deterministic relations in the data
    - e.g. suppose there is a section of a survey that only non-smokers answer about their preferred brand of cigarettes. Now suppose that some non-smokers have NAs for those questions. If you are not careful, you can end up with imputed values for favorite brands of cigarettes for the non-smokers!
- Imputations can be nonsensical.
- Models that will be applied to the imputed data may not (yet) be known.

Monotone vs. non-monotone

Recall that if we can order the variables such that the amount of missing data on Z_1 is less than Z_2, which is less than Z_3 and so on, we have a **monotone missing data pattern**. (This arises in longitudinal studies when missingness occurs only through dropout).

In this situation, missing values can be imputed by

- First fitting an appropriate model for Z_1 given fully observed covariates X, and then randomly sampling from this model to impute the missing values for Z_1.
- Next, fit an appropriate model for Z_2 given the fully observed X and the *observed* and *imputed* values of Z_1, and impute values for Z_2.
- Imputation proceeds in this way until all of the missing values have been filled in.

The resulting set of imputed values is a proper imputation of Z^{mis} from $f(Z^{mis}|Z^{obs})$ when the MAR assumption holds.

Monotone vs. non-monotone

When the missing data patterns are non-monotone or intermittent, iterative and more computationally demanding methods are usually required because it is no longer straightforward to sample randomly from $f(Z^{mis}|Z^{obs})$.

Imputation

Broadly speaking, methods for imputing the missing values fall into two categories.

- Methods based on the multivariate normal distribution
  - e.g. Amelia II package (King et al. 2001) in R, which uses a bootstrapping-based EM algorithm that is both fast and robust.
  - See Supplemental Slides for more information.
- Methods based on a fully conditional specification.
  - e.g. the mice package in R

Fully Conditional Specification (FCS)

- In order to address the issues posed by real-life complexities of the data, it may be convenient to specify the imputation model separately for each column in the data.
- The **joint distribution** that we care about is obtained by specifying a set of **conditional** densities $P(Z_j|Z_{-j}, \theta_j)$
- Multiple Imputation by Chained Equations (MICE) allows the user to impute the missing data as a series of **conditional univariate imputations** (Van Buuren et al. 2006).
- This is implemented in the `mice` package in R.

Stef Van Buuren's book, *Flexible Imputation of Missing Data* `https://stefvanbuuren.name/fimd/` is an excellent resource available online!

Multiple Imputation with Chained Equations (MICE)

- Let the hypothetically complete data $\mathbf{Z}$ be a partially observed random sample from the p-variate multivariate distribution $P(\mathbf{Z}|\boldsymbol{\theta})$.
- We assume that the multivariate distribution of $\mathbf{Z}$ is completely specified by $\boldsymbol{\theta}$, a vector of unknown parameters.
- The problem is how to get the multivariate distribution of $\boldsymbol{\theta}$, either explicitly or implicitly. The MICE algorithm obtains the posterior distribution of $\boldsymbol{\theta}$ by sampling iteratively from conditional distributions of the form

$$
\begin{aligned}
&P(Z_1|Z_{-1}, \theta_1) \\
&\vdots \\
&P(Z_p|Z_{-p}, \theta_p)
\end{aligned}
$$

Multiple Imputation with Chained Equations (MICE)

The parameters $\theta_1, \ldots, \theta_p$ are specific to the respective conditional densities and are not necessarily the product of a factorization of the 'true' joint distribution $P(\mathbf{Z}|\boldsymbol{\theta})$. Starting from a simple draw from observed marginal distributions, the ith iteration of chained equations is a Gibbs sampler that successively draws

$$\begin{aligned}
\theta_1^{*(t)} &\sim P(\theta_1|Z_1^{obs}, Z_2^{(t-1)}, \ldots, Z_p^{(t-1)}) \\
Z_1^{*(t)} &\sim P(Z_1|Z_1^{obs}, Z_2^{(t-1)}, \ldots, Z_p^{(t-1)}, \theta_1^{*(t)}) \\
&\vdots \\
\theta_p^{*(t)} &\sim P(\theta_p|Z_p^{obs}, Z_1^{(t)}, \ldots, Z_{p-1}^{(t)}) \\
Z_p^{*(t)} &\sim P(Z_p|Z_p^{obs}, Z_1^{(t)}, \ldots, Z_p^{(t)}, \theta_p^{*(t)})
\end{aligned}$$

where $Z_j^{(t)} = (Z_j^{obs}, Z_j^{*(t)})$ is the jth imputed variable at iteration t. Note that previous imputations $Z_j^{*(t-1)}$ only enter $Z_j^{*(t)}$ through its relation with other variables, and not directly. Convergence can be quite fast, often requiring only a small number of iterations.

Multiple Imputation with Chained Equations (MICE)

- The method has been found to work well in a variety of simulation studies.
- However, note that it is possible to specify models for which no known joint distribution exists! For example, two linear regressions specify a joint multivariate normal given specific regularity conditions. However, the joint distribution of one linear and, say, one proportional odds regression model is unknown, yet very easy to specify with the MICE framework.
- The consequences of incompatibility on the quality of the imputations is an area of active research (see Van Buuren 2018, Section 4.5.3 for discussion of this).
- One strategy for avoiding potential problems of incompatibility is to specify a sensible order for the conditionally specified models.

Some practical advice

- Sometimes we have **derived variables** that are a deterministic function of other variables in the dataset. E.g. BMI is defined as $weight/height(kg/m^2)$.
- Unless we specify otherwise, the imputation model is unaware of the relationship between weight, height, and BMI, and will produce imputations that are inconsistent with the deterministic rule.
- One way to deal with this is to **Impute, then Transform**: we leave BMI outside of the imputation process and impute any missing height and weight data. Then we create BMI from the imputed values of weight and height afterwards.
- However, the tendency will be that estimates of parameters related to BMI will be biased towards zero.

Some practical advice

- Another possibility is to create BMI before imputation, and then impute BMI as **Just Another Variable (JAV)**. **JAV** means that we **Transform, then Impute**, but as noted, this can result in inconsistency between imputed values of weight and height and BMI.
- A third approach is **Passive Imputation**, in which the transformation is done on-the-fly within the imputation algorithm. Since the transformed variable is available for imputation, the hope is that passive imputation removes the bias of **Impute, then Transform**, while restoring consistency among the imputations that was broken in **JAV**.
- Some simulation studies favor the use of **JAV**, but Van Buuren (2018) urges caution when using any of the three methods for ratio variables.

Some practical advice

- Bartlett et al. (2015) proposed a novel rejection sampling method that creates imputations that are congenial in the sense of Meng (1994) with the substantive (complete-data) model.
- The method has been implemented in the `smcfcs` package.
- Van Buuren (2018, Section 6.4.1) shows that the `scmfcs` method is far better than **Impute, then Transform**, **JAV**, or **Passive Imputation** for deriving ratio variables.

See this helpful blog post on `scmfcs`:
`https://thestatsgeek.com/2014/05/10/multiple-imputation-with-interactions-and-non-linear-terms/`

Some practical advice

- In general, if your analytic model will include interactions between variables that are partially missing, those interactions should also be included in the imputation model.
- This is related to the notion of **congeniality** (Meng 1994): the imputation model should be at least as richly parameterized as the analytic model.
- As with derived variables, **Impute, then Transform**, **JAV**, and **Passive Imputation** are all options for imputing interactions.
- As with the derived ratio variables, `smcfcs` appears to perform best.

Some practical advice

- It's always a good idea to understand the patterns of missing data in your dataset before you start imputing!
- If you have two variables with substantial missingness on each, it is not really too useful to include either of these when imputing the other.
- If you have a lot of highly correlated variables, the imputation model may be unstable and potentially lead to wide confidence intervals. It may be helpful to reduce the set of variables included in the imputation models to a set of weakly correlated predictors.
- Consider what kind of checks you can perform after imputation to see that the imputed values are reasonable and "make sense."

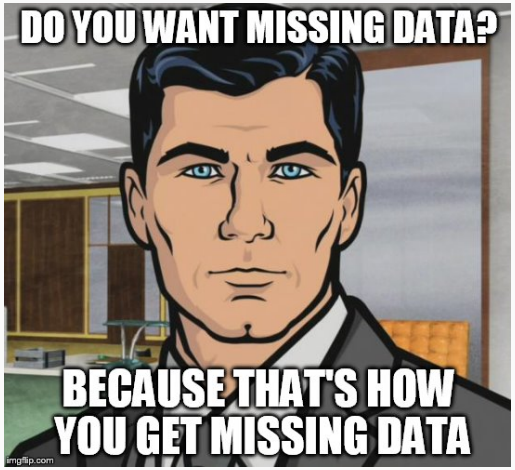
Burton and Altman (BJC 2004) proposed guidelines for reporting analyses with missing data

1. Quantification of completeness of covariate data
    - if availability of data is an exclusion criterion, specify the number of cases excluded for this reason
    - provide the total number of eligible cases and the number with complete data
    - report the frequency of missing data for every variable
2. Exploration of the missing data
    - discuss any known reasons for missing covariate data
    - present the results of any comparisons of characteristics between the cases with or without missing data

Burton and Altman (BJC 2004) proposed guidelines (cont)

3. Approaches for handling missing covariate data
    - provide sufficient details of the methods adopted
    - give appropriate references for any imputation method used
    - for each analysis, specify the number of cases included and the associated number of events.

Recommendations



1. Practice good data collection procedures to try to minimize missing data as much as possible.
2. Collect data on as many variables possible that are predictive of missingness in order to justify Missing At Random (MAR) assumptions.
3. Think carefully about what is the likely missingness mechanism (MCAR, MAR, MNAR).

Recommendations

4. Consider the pattern of missingness (monotone/non-monotone; unit missingness vs. item missingness)
5. For dropout in longitudinal settings, inverse probability weighting is an attractive approach, under the assumption of MAR.
6. For intermittent missingness, non-monotone missingness, and multiple missing covariates, multiple imputation under the assumption of MAR is attractive and increasingly easy to implement using standard software.

Recommendations

7. Model based approaches are powerful and can sometimes even address MNAR, but likely will need expert statistical guidance to implement.
8. **Be thorough in documenting how you handled missing data and what assumptions were made**.
9. Consider sensitivity analyses to explore sensitivity to assumptions made.

References

Some important textbooks:

1. Little RJA, Rubin DB. (1987) *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
2. Rubin DB. (1987) *Multiple Imputation for Non-Response in Surveys*. New York: John Wiley & Sons.
3. Schafer JL. (1997) *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman & Hall/CRC.
4. Van Buuren S (2018) *Flexible Imputation of Missing Data, 2nd Edition*, Boca Raton: Chapman & Hall/CRC. Available online at: `https://stefvanbuuren.name/fimd/`

References

Some helpful papers:

1. Graham JW, Olchowski AE, Gilreath TD. 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* 8:206-13
2. Graham, JW. (2009) Missing Data: Making It Work in the Real World. *Annu Rev Psychol* 60:549-576.
3. Horton NJ, Kleinman KP. (2007) Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *The American Statistician* 61:79-90.
4. King G, Honaker J, Joseph A, Scheve K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review* 95: 49-69.

References

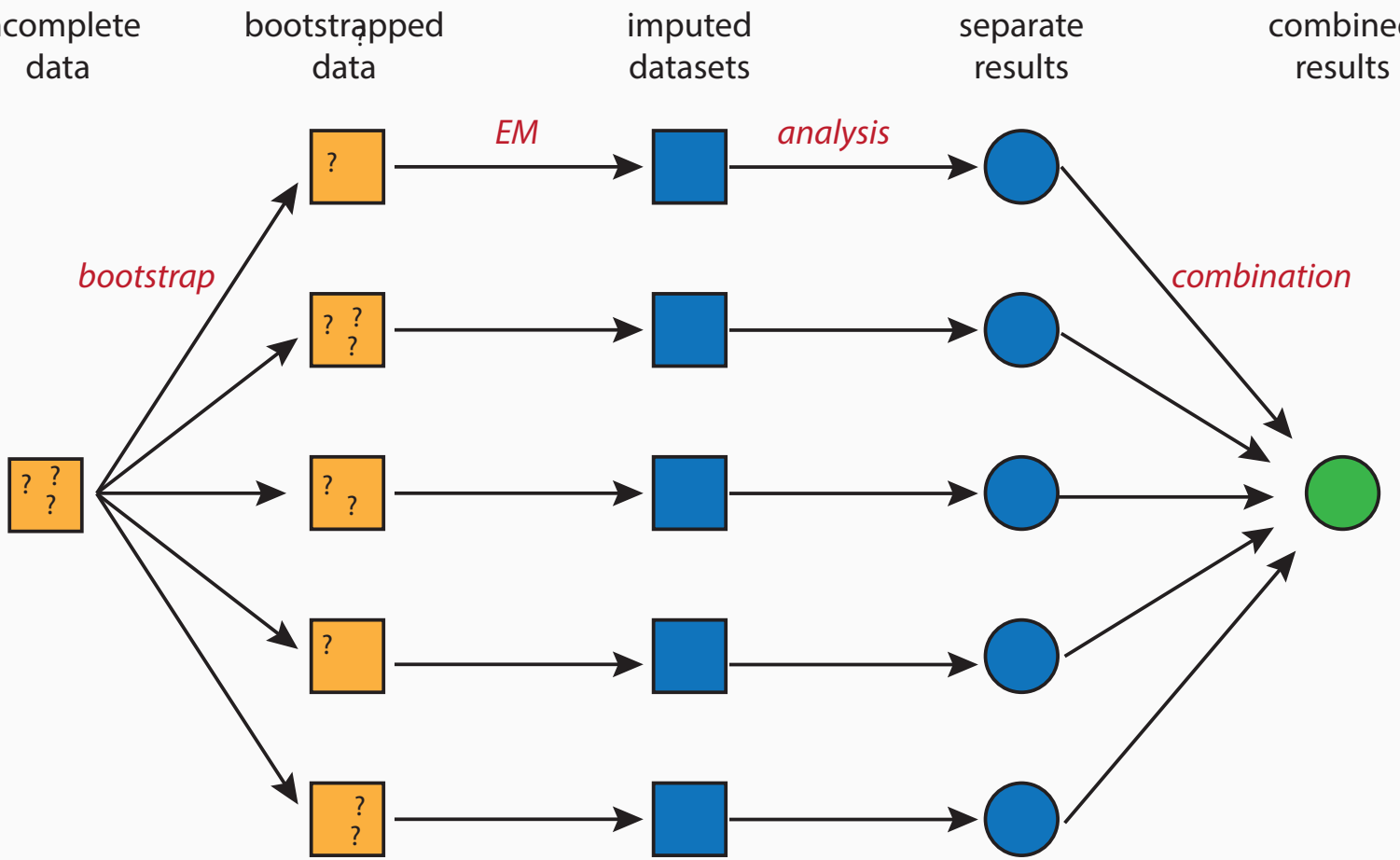
5. Meng, X. L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science* 9 (4): 538-73.
6. Rubin DB (1997). Muliple imputation after 18+ Years. *Journal of the American Statistical Association* 91(434): 473-489.
7. Van Buuren, S., J.P.L. Brand, C.G.M. Groothuis-Oudshoorn and D.B. Rubin (2006) Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 76: 1046-1064.
8. Vansteelandt, S., Carpenter, J., & Kenward, M. G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6(1), 37–48. `https://doi-org.ezp-prod1.hul.harvard.edu/10.1027/1614-2241/a000005`

Supplemental Slides

Multiple imputation based on the multivariate normal distribution

- The **multivariate normal** model has often been used for multiple imputation as it is computationally tractable (since only the mean vector and the variance-covariance matrix needs to be estimated).
- This model has been used even when some of the variables are not Gaussian
- Interestingly, numerous authors suggest that the assumption of multivariate normality works well, even when some of the variables are binary or categorical.
- Appropriate data transformations can be used to try to make the data more normal.
- Implemented in the Amelia II package in R.
  `https://gking.harvard.edu/amelia`

Amelia II


incomplete data
bootstrapped data
imputed datasets
separate results
combined results
EM
analysis
bootstrap
combination

Amelia II

- Draw m samples of size n with replacement from the data.
- In each sample, run the EM algorithm to produce estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (the mean and covariance of the multivariate normal distribution of the data).
- Then for each set of estimates, use the original sample units to impute the missing observations in their original positions.
- The result is m multiply imputed data sets that can be used for complete-data analysis and combined using the usual rules.

Some useful data transformations

- log or square root transform continuous variables with skewed distributions
- binary variables entered as is
- for a categorical variable with p categories, include $p - 1$ binary dummies.
- do not round imputed values.