2022 Essex Summer School
3K: Dynamics and Heterogeneity

Robert W. Walker, Ph. D.
Associate Professor of Quantitative Methods
Atkinson Graduate School of Management
Willamette University
Salem, Oregon USA
rwalker@willamette.edu

August 15, 2022

Day VII: On Missing Data

Why do we care about missing data? What problem does an optimal method for missing data accomplish?

Types of Missing Data

- OAR (Observed at Random)
  Missingness on Y is not determined by X.

- MAR (Missing at Random)
  Missingness on Y is not determined by Y.

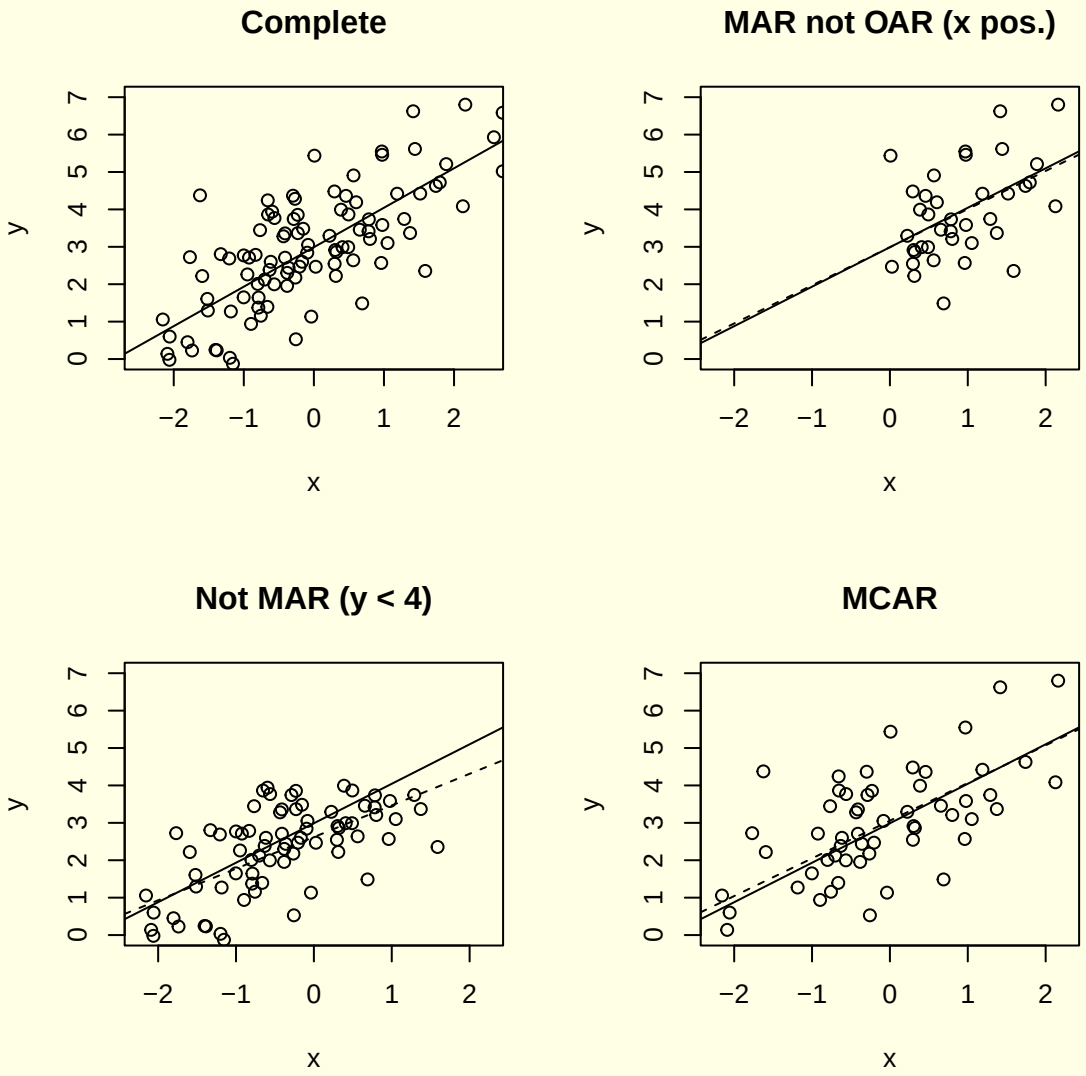
- MCAR (Missing Completely at Random)
  Missingness is both OAR and MAR.

Suppose two variables X and Y.

Illustrating these Points

```
> library(MASS)
> big.X <- mvrnorm(n = 100, c(0, 0, 0), matrix(c(1, 0.8, -0.7, 0.8,
> x <- big.X[, 1]
> y <- 3 + x + rnorm(100)
> df1 <- data.frame(x = x, y = y)
> df2 <- df1
> df2$x[df1$x < 0] <- NA
> df3 <- subset(df1, subset = y < 4)
> df4 <- df1[sample(c(1:100), size=60, replace=FALSE),]
> par(mfrow = c(2, 2))
> plot(df1, main = "Complete", ylim = c(0, 7), xlim = c(-2.5, 2.5))
> abline(lm(y ~ x, data = df1))
> plot(df2, xlim = c(-2.25, 2.25), ylim = c(0, 7), main = "MAR not
> abline(lm(y ~ x, data = df1), lty = 1)
```

```
> abline(lm(y ~ x, data = df2), lty = 2)
> plot(df3, xlim = c(-2.25, 2.25), ylim = c(0, 7), main = "Not MAR
> abline(lm(y ~ x, data = df1), lty = 1)
> abline(lm(y ~ x, data = df3), lty = 2)
> x2 <- big.X[, 2]
> x3 <- big.X[, 3]
> plot(df4, xlim = c(-2.25, 2.25), ylim = c(0, 7), main = "MCAR")
> abline(lm(y ~ x, data = df1), lty = 1)
> abline(lm(y ~ x, data = df4), lty = 2)
```


Complete
MAR not OAR (x pos.)
Not MAR (y < 4)
MCAR
y
x
0 1 2 3 4 5 6 7
−2 −1 0 1 2

What to do?

- Assess missingness?
- Multiple Imputation
- MCMC

Assessing Missingness

- Simple regression techniques (binary GLMs) can serve the useful purpose of identifying the predictors of missingness.
- As a practical matter, we can't really establish the true nature of missingness.
- BUT, we can at least examine missingness as a function of observed covariates. Doing this, we should be careful to engage functional forms above and beyond simple linearity.

Multiple Imputation

- Multiple imputation: the process of "imputing" missing data on multiple occasions to rectangularize a data matrix for analysis.
- Most commonly done with a multivariate normal distribution where the mean vector and variance/covariance matrix form the basis for imputing.
- Other methods are also frequently used including versions of "nearest neighbors" and simple mean imputation. The latter is generally a bad strategy.
- R makes a few methods of imputation readily available.

Imputation with R

- `Amelia`: Performs multiple imputation under a multivariate normal distribution. It contains functions for time series, cross-sectional, and time series cross-sectional data and includes an ability to handle nominal and ordered data.

- `robCompositions`: Multiple imputation for missing data on a simplex or other bounded constraint space.

- `mice`: Multivariate Imputation using Chained Equations. Basically, a Gibbs sampler (a full conditional specification of the imputation equations).

- `mi`: Bayesian specifications of regression models for the imputation of missing data.

Imputation in Stata

Stata was a bit late to the multiple imputation game but has improved considerably in version 12. There is an overarching suite of commands that can be employed with the prefix `mi`. There are some things that Stata cannot (does not) do. For example, there is no method for compositional data; this can be problematic. Of course, the negative correlations can allow unbounded approximations to get close, but that isn't what we really want. It is better to use all of the available information in the data because throwing away information when we are "making it up" guarantees that we make it up less well.

Specifically for TSCS/CSTS data, Honaker and King (2010) in the American Journal of Political Science is on this exact problem. Their software is, however, written in R. You can find the paper from `http://gking.harvard.edu/files/pr.pdf`.

Have Imputations, What do I do?

- The trouble with multiply imputed data come in two forms.
- One is that we want to analyze the (now) complete data but need to account for the presence of predictions.
- The other is that we need some method or methods for the assessment of our imputation.

Analyzing Imputed Data

- Imputed data are predictions from models. Not unlike any predicted variable, the imputations are draws from a sampling distribution and cannot be argued fixed in repeated samples with a straight face. Moreover, the sampling distributions of statistics to which imputations are inputs must reflect the uncertainty of the imputation alongside more conventional uncertainty about the statistic.

- Rubin first gave a formula for combining imputation-based statistics.

- It relies on the asymptotic normality of the statistics; they are a linear combination of normals.

Diagnostics for Imputed Data

- Functions exist for plotting patterns of missingness: `missing.pattern.plot`
- Comparing histograms (after imputation): `mi.hist`, etc.
- Comparing scatterplots: `mi.scatterplot`

The Statistics of it all

We can calculate a few interesting quantities that inform imputation independent of the ultimate goal. For example, let

$$W_\beta = \sum_{k=1}^{K} \frac{V_k}{K}$$

be the average within imputation variance of β_k. Similarly, let the between imputation variance of β be,

$$B_\beta = \sum_{k=1}^{K} \frac{(\beta_k - \overline{\beta})^2}{(K-1)}$$

which yields a total variance of

$$T_\beta = B_\beta(1 + \frac{1}{k}) + W_\beta$$

which gives $\frac{\overline{\beta}}{\sqrt{T_\beta}} \sim t_\nu$ where

$$\nu = (K-1)[1 + WB^{-1}(1 + \frac{1}{K})^{-1}].$$

Furthermore, notice that if the imputations are completely uninformative regarding β, then $\beta_k = \beta^* \; \forall k \in K$ and $T_\beta = W_\beta$. This allows us to construct a ratio, $r = \frac{(1+\frac{1}{K})B_\beta}{W_\beta}$ to measure the increase in variance owing to imputation. Finally, let $\epsilon = \frac{r}{1+r}$ be the proportion of missing information. All of this comes together when the relative efficiency of K imputations relatively to an infinite number is $(1 + \frac{\epsilon}{K})^{-1}$. Take an example of $K = 0.5$ and 5 imputations.

Some General Comments

There is almost no reason not to impute data. Not imputing throws away information. Imputing may not add any but it allows us to retain "real" information. Consider the following scenarios. Suppose that MAR fails. What does imputation really do and what are the properties of the original estimator anyway? What happens with imputation when MAR holds? What is your belief that MCAR ever holds?

The Bayes Approach

Missing data are like anything else that is missing. In the Bayesian framework, we want to sample the missing quantities. We can learn about their distributions, covariances, and the like but we fundamentally want to learn as little as possible from the missing data but maximize the quantity of information recovered from other nonmissing things. Imputation simply becomes another part of the sampler.