Univariate Multiple Imputation

Utrecht University Winter School: Missing Data in R

Utrecht University

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Single Imputation

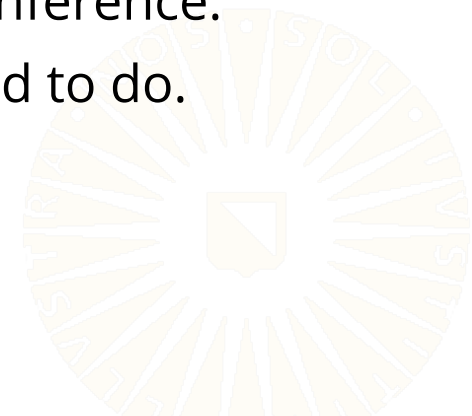
Imputation is Just Prediction*

Imputation is nothing more than a type of prediction.

1. Train a model on the observed parts of the data, Y_{obs}.
    - Train the imputation model.
2. Predict the missing values, Y_{mis}.
    - Generate imputations.
3. Replace the missing values with these predictions.
    - Impute the missing data.

Imputation can be used to support either prediction or inference.

- Our goals will dictate what type of imputation we need to do.



*Levels of Uncertainty Modeling

van Buuren (2018) provides a very useful classification of different imputation methods:

1. Simple Prediction
   - The missing data are naively filled with predicted values from some regression equation.
   - All uncertainty is ignored.

2. Prediction + Noise
   - A random residual error is added to each predicted value to create the imputations.
   - Only uncertainty in the predicted values is modeled.
   - The imputation model itself is assumed to be correct and error-free.

3. Prediction + Noise + Model Error
   - Uncertainty in the imputation model itself is also modeled.
   - Only way to get fully proper imputations in the sense of Rubin (1987).

Do we really need to worry?

The arguments against single imputation can seem archaic and petty. Do we really need to worry about this stuff?

Do we really need to worry?

The arguments against single imputation can seem archaic and petty. Do we really need to worry about this stuff?

- YES!!! (At least if you care about inference)

The following are results from a simple Monte Carlo simulation:

	Complete Data	Conditional Mean	Stochastic	MI
cor(X, Y)	0.500	0.563	0.498	0.497
Type I Error	0.052	0.138	0.120	0.054

Mean Correlation Coefficients and Type I Error Rates

Do we really need to worry?

The arguments against single imputation can seem archaic and petty. Do we really need to worry about this stuff?

- YES!!! (At least if you care about inference)

The following are results from a simple Monte Carlo simulation:

	Complete Data	Conditional Mean	Stochastic	MI
cor(X, Y)	0.500	0.563	0.498	0.497
Type I Error	0.052	0.138	0.120	0.054

Mean Correlation Coefficients and Type I Error Rates

- Conditional mean substitution overestimates the correlation effect.
- Both single imputation methods inflate Type I error rates.
- MI provides unbiased point estimates and accurate Type I error rates.

Simulate Some Toy Data

```r
library(mvtnorm)
library(dplyr)

nObs <- 1000 # Sample Size
pm   <- 0.3  # Proportion Missing

sigma <- matrix(c(1.0, 0.5, 0.5, 1.0), ncol = 2)

dat0 <- rmvnorm(nObs, c(0, 0), sigma) %>% as.data.frame()
colnames(dat0) <- c("y", "x")
```

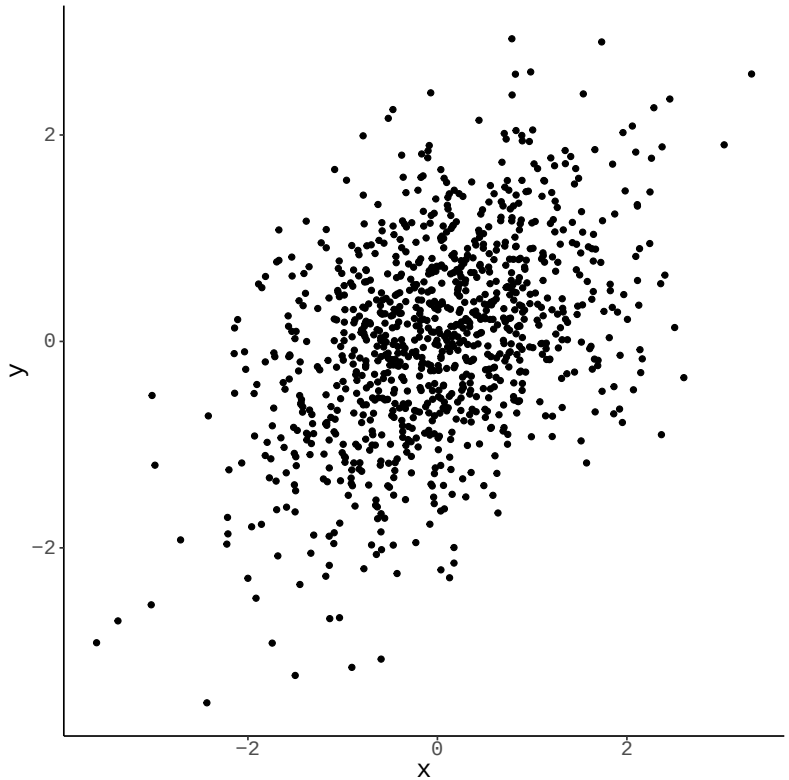
Simulate Some Toy Data

```r
## Impose MAR Nonresponse:
dat1 <- dat0
mVec <- with(dat1, x < quantile(x, probs = pm))

dat1[mVec, "y"] <- NA

## Subset the data:
yMis <- dat1[mVec, ]
yObs <- dat1[!mVec, ]
```

Look at the Data

```
head(dat0, n = 5) %>% round(3)
```

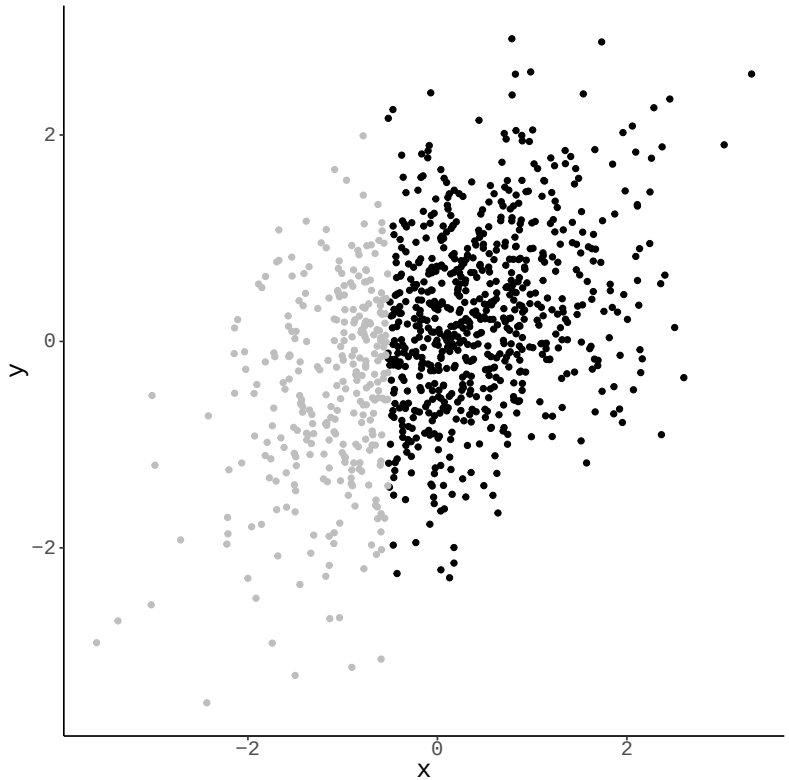
```
       y      x
1 -0.364 -1.564
2  0.408  1.630
3 -0.921  1.212
4 -0.906 -0.588
5 -0.182  0.870
```



Look at the Data

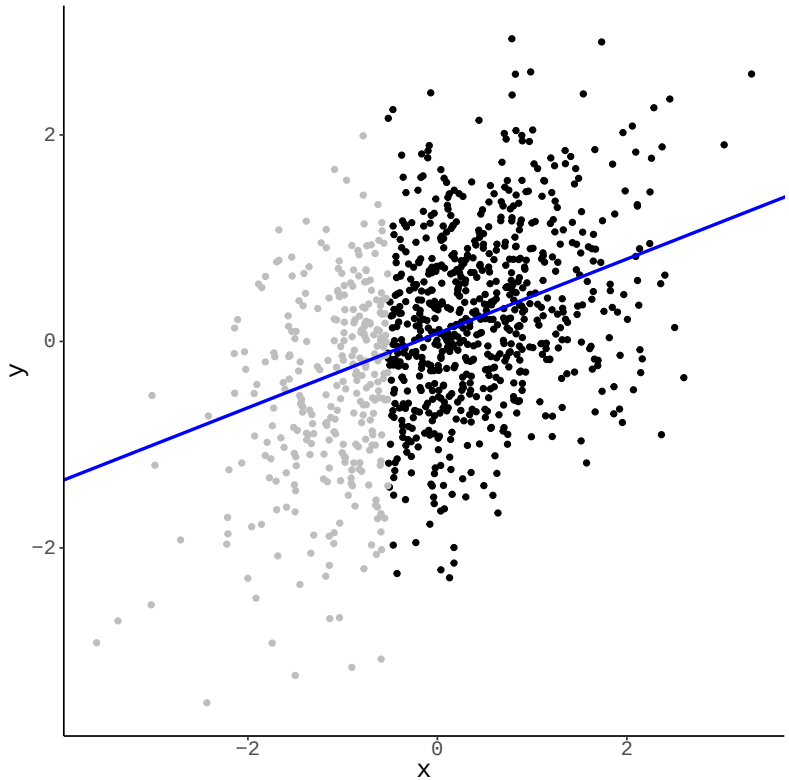
```
head(dat1, n = 5) %>% round(3)
```

```
       y      x
1     NA -1.564
2  0.408  1.630
3 -0.921  1.212
4     NA -0.588
5 -0.182  0.870
```



Expected Imputation Model Parameters

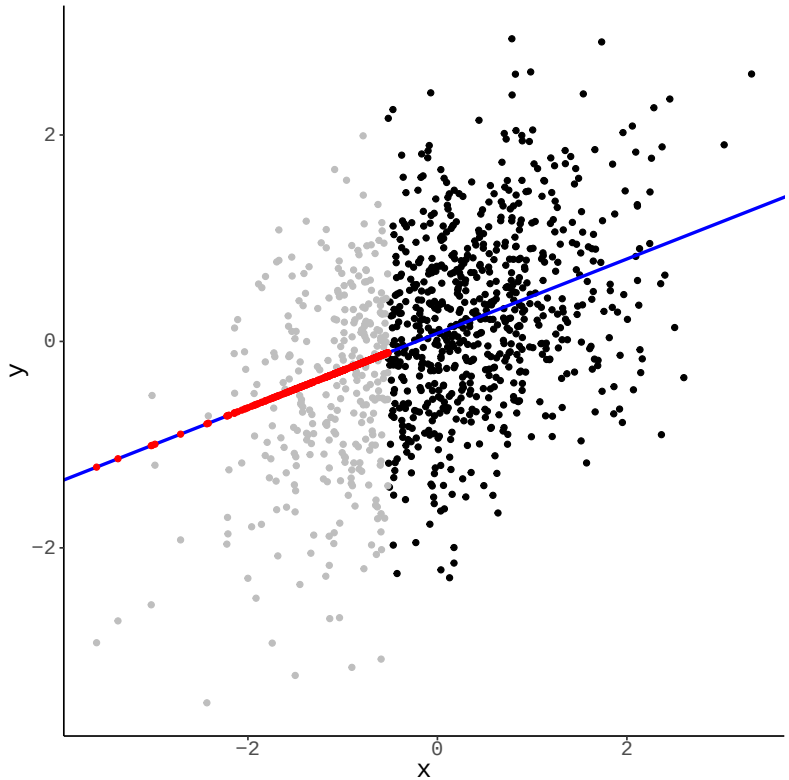
```
lsFit <- lm(y ~ x, data = yObs)

beta  <- coef(lsFit)
sigma <- summary(lsFit)$sigma

as.matrix(beta)

                   [,1]
(Intercept) 0.07810014
x           0.36026696

sigma

[1] 0.8348726
```



Conditional Mean Substitution

```
## Generate imputations:
imps <- beta[1] + beta[2] * yMis$x

## Fill missing cells in Y:
dat1[mVec, "y"] <- imps

head(dat1, n = 5) %>% round(3)
```

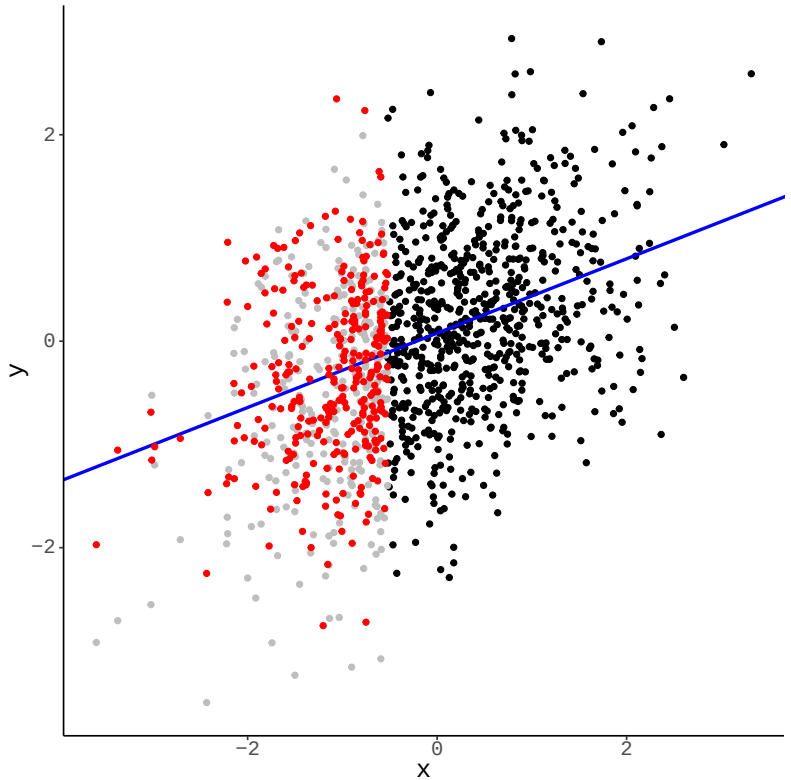
```
       y      x
1 -0.485 -1.564
2  0.408  1.630
3 -0.921  1.212
4 -0.134 -0.588
5 -0.182  0.870
```



Stochastic Regression Imputation

```
## Generate imputations:
imps <- imps +
    rnorm(nrow(yMis), 0, sigma)

## Fill missing cells in Y:
dat1[mVec, "y"] <- imps

head(dat1, n = 5) %>% round(3)
```

```
       y      x
1 -0.227 -1.564
2  0.408  1.630
3 -0.921  1.212
4  0.032 -0.588
5 -0.182  0.870
```



Multiple Imputation

Flavors of MI

MI simply repeats a single regression imputation M times.

- The specifics of the underlying regression imputation are important.

Flavors of MI

MI simply repeats a single regression imputation M times.

- The specifics of the underlying regression imputation are important.

Simply repeating the stochastic regression imputation procedure described above won't suffice.

- Still produces too many Type I errors

	Complete Data	PN-Type	PNE-Type
cor(X, Y)	0.499	0.499	0.498
Type I Error	0.040	0.066	0.046

Mean Correlation Coefficients and Type I Error Rates

- Type I error rates for PN-Type MI are much better than they were for single stochastic regression imputation, but they're still too high.

Proper MI

The problems on the previous slide arise from using the same regression coefficients to create each of the M imputations.

- Implies that you're using the "correct" coefficients.
- This assumption is plainly ridiculous.
  - If we don't know some values of our outcome variable, how can we know the "correct" coefficients to link the incomplete outcome to the observed predictors?

Proper MI

The problems on the previous slide arise from using the same regression coefficients to create each of the M imputations.

- Implies that you're using the "correct" coefficients.
- This assumption is plainly ridiculous.
  - If we don't know some values of our outcome variable, how can we know the "correct" coefficients to link the incomplete outcome to the observed predictors?
- Proper MI also models uncertainty in the regression coefficients used to create the imputations.
  - A different set of of coefficients is randomly sampled (using Bayesian simulation) to create each of the M imputations.
  - The tricky part about implemented MI is deriving the distributions from which to sample these coefficients.

Setting Up Proper MI

Our imputation model is simply a linear regression model:

$$Y = \mathbf{X}\beta + \varepsilon$$

To fully account for model uncertainty, we need to randomly sample both β and $\text{var}(\varepsilon) = \sigma^2$.

Setting Up Proper MI

Our imputation model is simply a linear regression model:

$$Y = \mathbf{X}\beta + \varepsilon$$

To fully account for model uncertainty, we need to randomly sample both β and $\text{var}(\varepsilon) = \sigma^2$.

For a simple imputation model with a normally distributed outcome and uninformative priors, we need to specify two distributions:

1. The marginal posterior distribution of σ^2
2. The conditional posterior distribution of β

Marginal Distribution of σ^2

We first specify the marginal posterior distribution for the noise variance, σ^2.

- This distribution does not depend on any other parameters.

$$\sigma^2 \sim \text{Inv-}\chi^2\left(N-P, MSE\right) \tag{1}$$

$$\text{with } MSE = \frac{1}{N-P}\left(Y-\mathbf{X}\hat{\beta}_{ls}\right)^T\left(Y-\mathbf{X}\hat{\beta}_{ls}\right)$$

- σ^2 follows a scaled inverse χ^2 distribution.

Conditional Distribution of β

We then specify the conditional posterior distribution for β.

- This distribution is conditioned on a specific value of σ^2.

$$\beta \sim \text{MVN}\left(\hat{\beta}_{ls},\ \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right) \tag{2}$$

- β (conditionally) follows a multivariate normal distribution.

PPD of the Missing Data

Once we've sampled our imputation model parameters, we can construct the posterior predictive distribution of the missing data.

- This is the distribution from which we sample our imputed values.
- In practice, we directly compute the imputations based on the simulated imputation model parameters.

$$Y_{imp} = \mathbf{X}_{mis}\tilde{\beta} + \tilde{\varepsilon} \tag{3}$$

$$\text{with } \varepsilon \sim \text{N}\left(0, \widetilde{\sigma^2}\right)$$

General Steps for Basic MI

With all of the elements in place, we can execute a basic MI by following these steps:

1. Find the least squares estimates of β, $\hat{\beta}_{ls}$, by regressing the observed portion of Y onto the the analogous rows of $\mathbf{X}$.

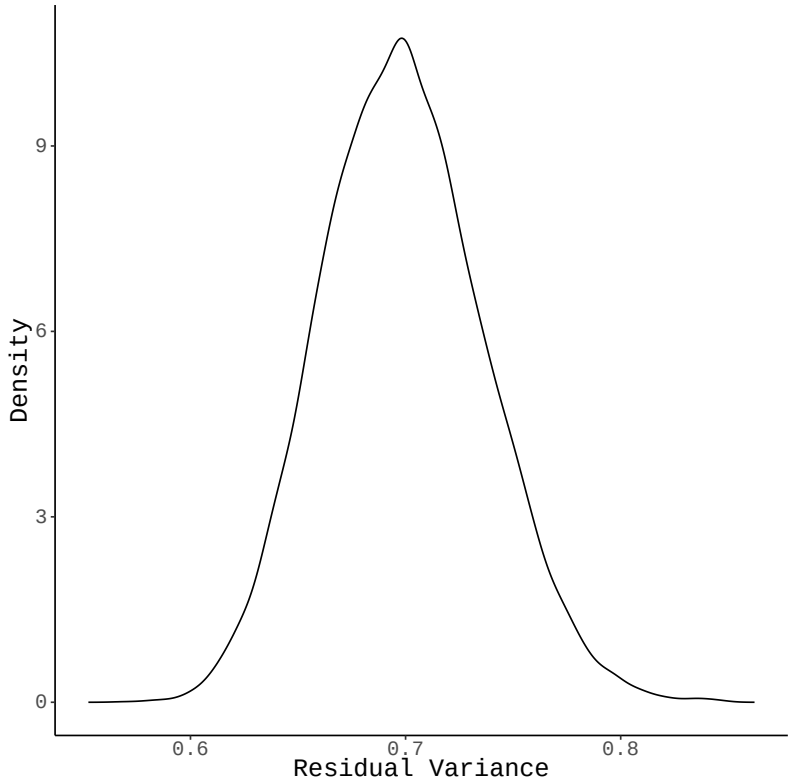
2. Use $\hat{\beta}_{ls}$ to parameterize the posterior distribution of σ^2, given by Equation 1, and draw M samples of σ^2 from this distribution.

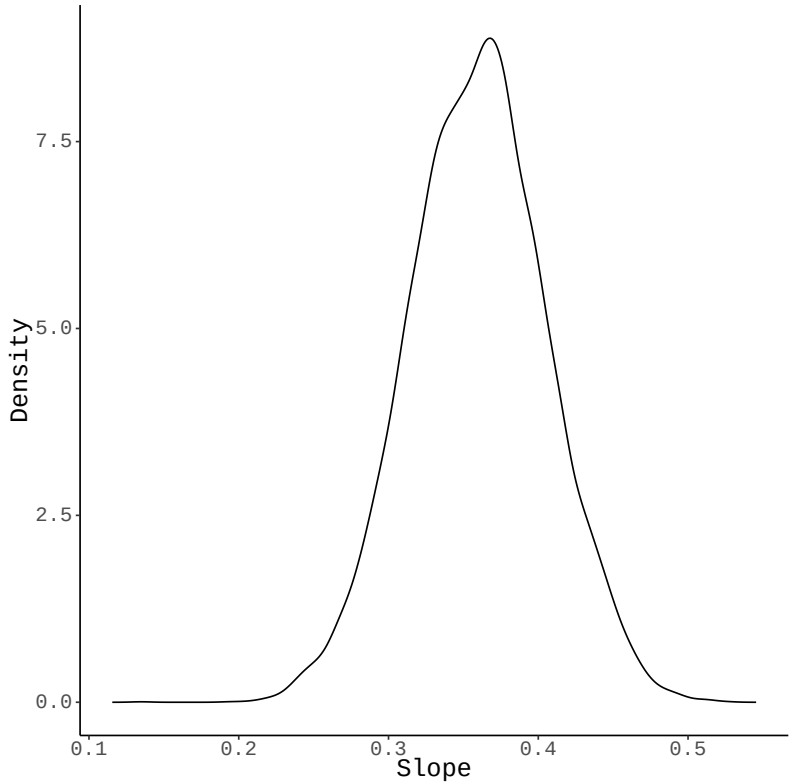
3. For each of the σ^2_m, sample a corresponding value of β from Equation 2.

4. Plug the M samples of β and σ^2 into Equation 3 to create the M imputations.

Visualizing MI

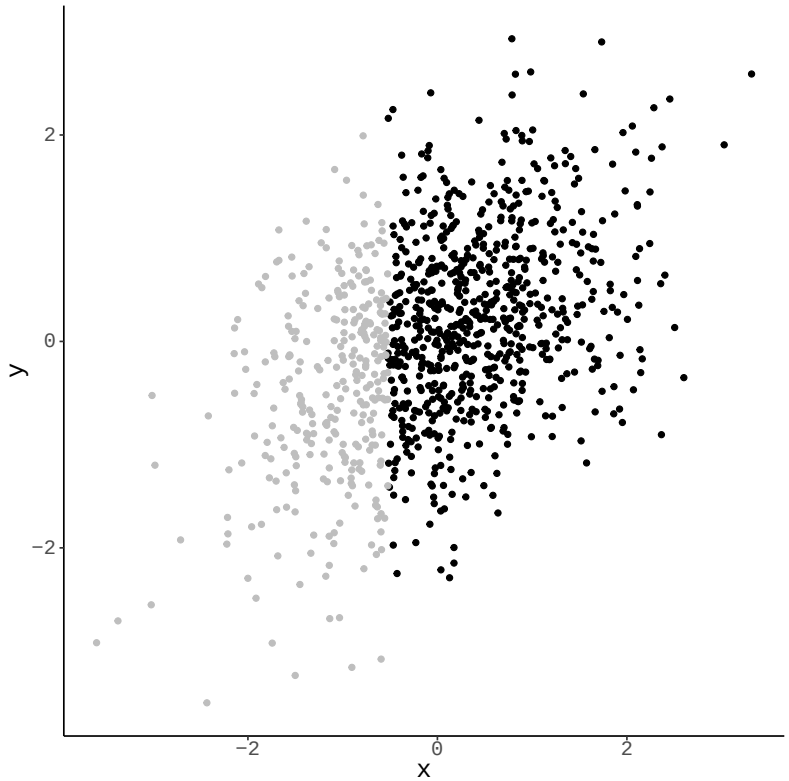
Use Bayesian simulation to estimate posterior distributions for the imputation model parameters:





Visualizing MI

Recall the incomplete data from the single imputation examples.

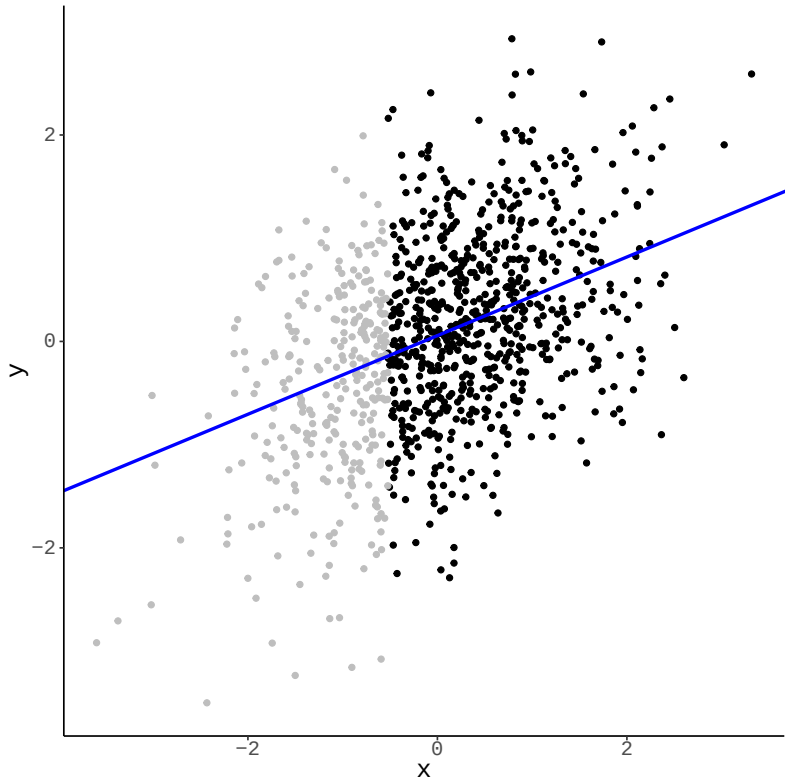


Visualizing MI

Sample values of β_0 and β_1:

- $\beta_0 = 0.056$
- $\beta_1 = 0.381$

Define the predicted best-fit line:
$\hat{Y}_{mis} = 0.056 + 0.381 X_{mis}$



Visualizing MI

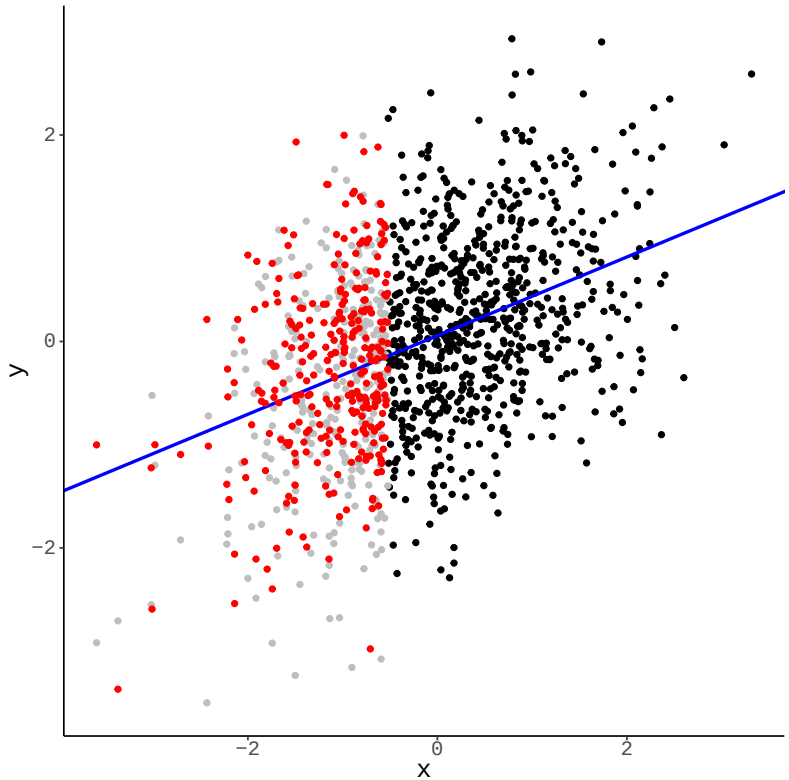
Sample a value of σ^2:

- $\sigma^2 = 0.634$

Generate imputations using the same procedure described in Single Stochastic Regression Imputation:

$$Y_{imp} = \hat{Y}_{mis} + \varepsilon$$

$$\varepsilon \sim \mathrm{N}(0, 0.634)$$

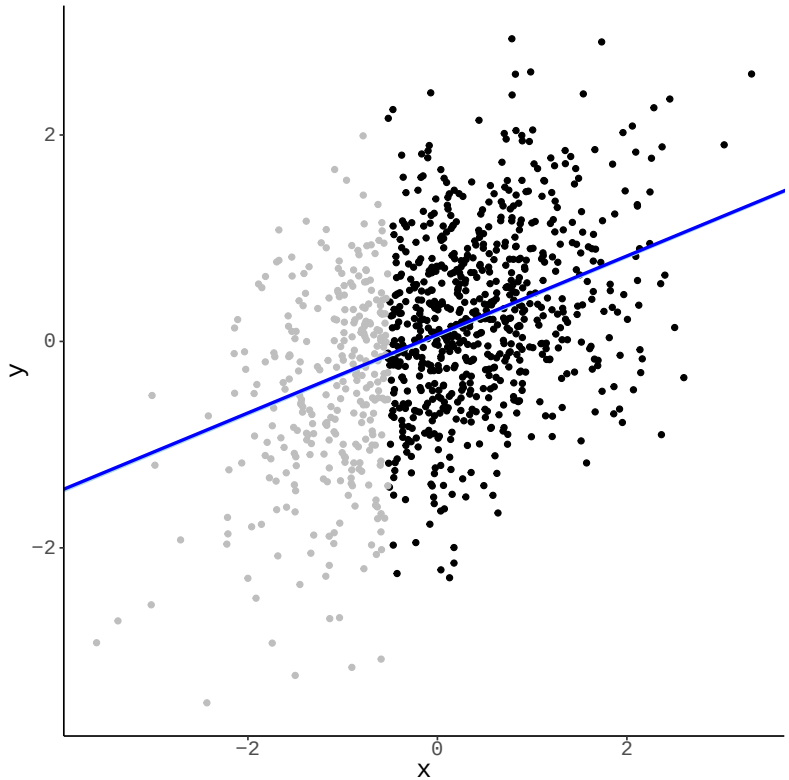


Visualizing MI

Sample values of β_0 and β_1:

- $\beta_0 = 0.068$
- $\beta_1 = 0.38$

Define the predicted best-fit line:
$\hat{Y}_{mis} = 0.068 + 0.38X_{mis}$



Visualizing MI

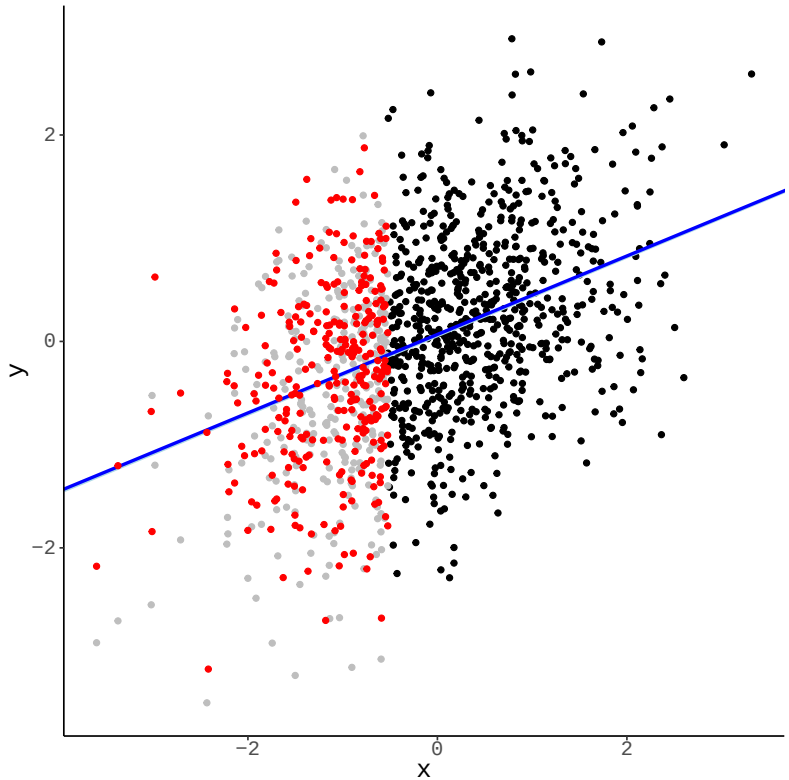
Sample a value of σ^2:

- $\sigma^2 = 0.714$

Generate imputations using the same procedure described in Single Stochastic Regression Imputation:

$$Y_{imp} = \hat{Y}_{mis} + \varepsilon$$

$$\varepsilon \sim \mathrm{N}(0, 0.714)$$

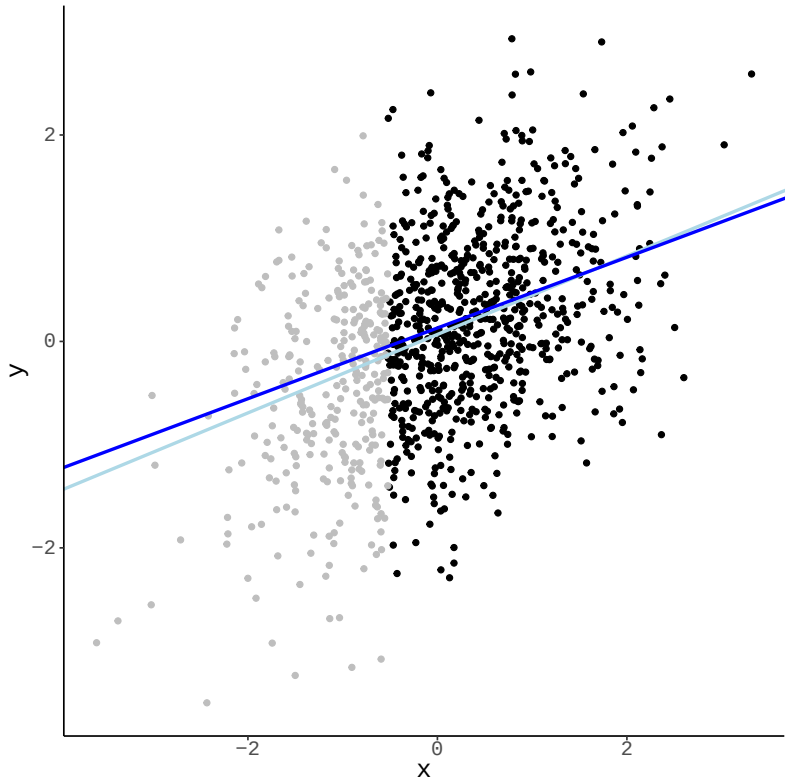


Visualizing MI

Sample values of β_0 and β_1:

- $\beta_0 = 0.13$
- $\beta_1 = 0.343$

Define the predicted best-fit line:
$\hat{Y}_{mis} = 0.13 + 0.343X_{mis}$



Visualizing MI

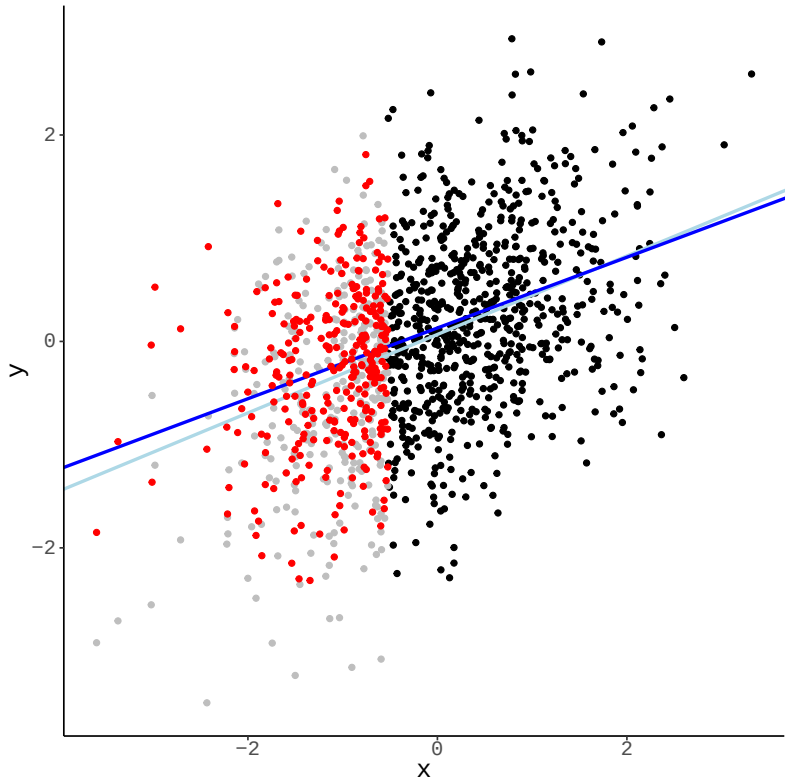
Sample a value of σ^2:

- $\sigma^2 = 0.688$

Generate imputations using the same procedure described in Single Stochastic Regression Imputation:

$$Y_{imp} = \hat{Y}_{mis} + \varepsilon$$

$$\varepsilon \sim \mathrm{N}(0, 0.688)$$



MI-Based Analysis

Doing MI-Based Analysis

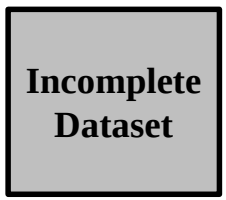
An MI-based data analysis consists of three phases:

1. The imputation phase
   - Replace missing values with M plausible estimates.
   - Produce M completed datasets.

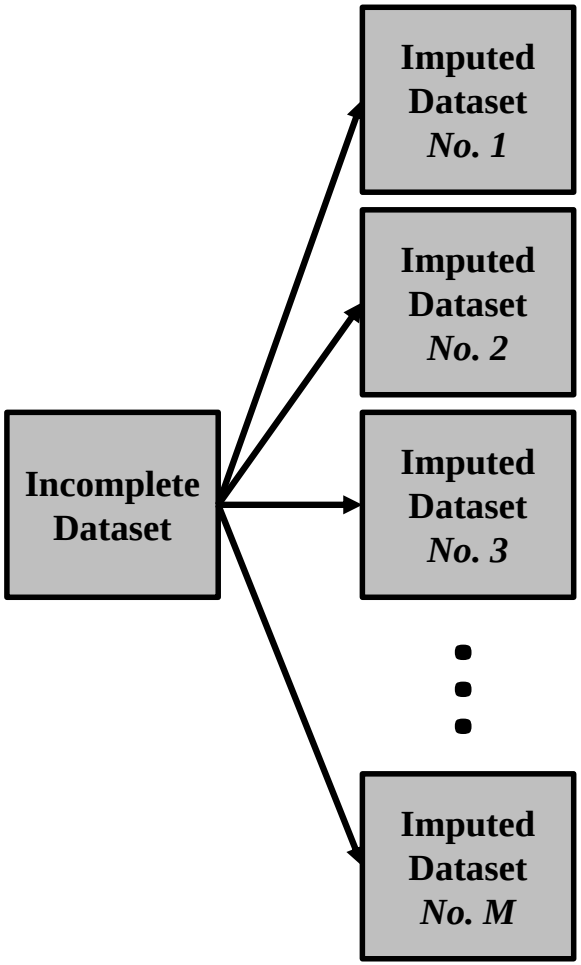
2. The analysis phase
   - Estimate M replicates of your analysis model.
   - Fit the same model to each of the M datasets from Step 1.

3. The pooling phase
   - Combine the M sets of parameter estimates and standard errors from Step 2 into a single set of MI estimates.
   - Use these pooled parameter estimates and standard errors for inference.

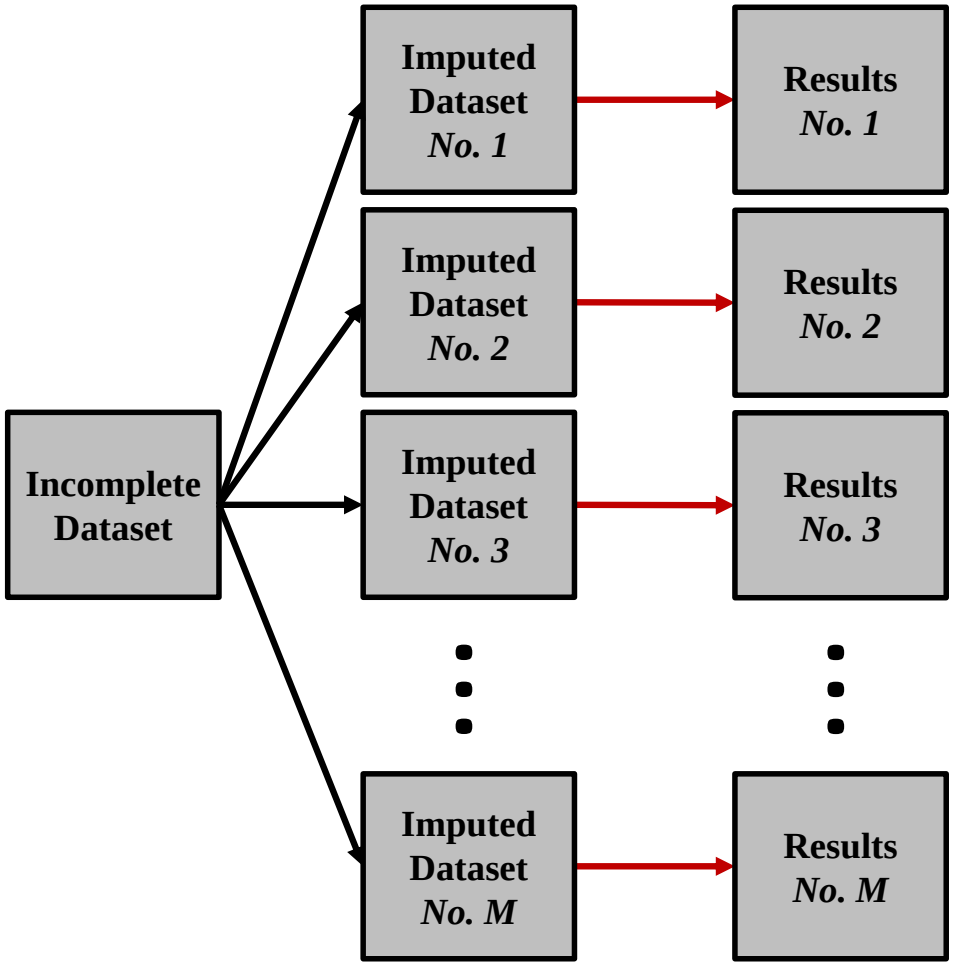
MI-Based Analysis



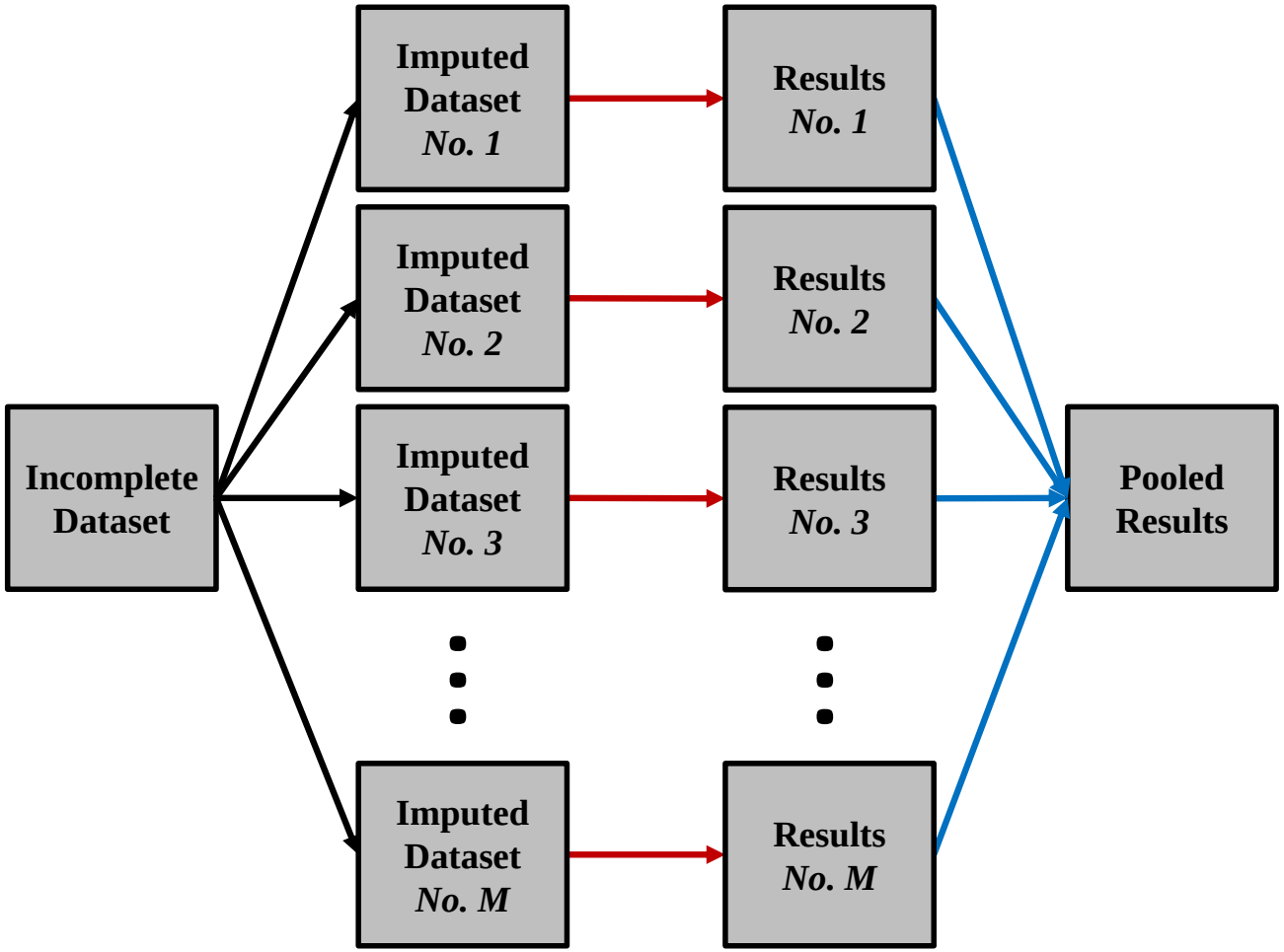
MI-Based Analysis



MI-Based Analysis



MI-Based Analysis



Pooling MI Estimates

Rubin (1987) formulated a simple set of pooling rules for MI estimates.

- The MI point estimate of some interesting quantity, Q^*, is simply the mean of the M estimates, $\{\hat{Q}_m\}$:

$$Q^* = \frac{1}{M} \sum_{m=1}^{M} \hat{Q}_m$$

Pooling MI Estimates

The MI variability estimate, T, is a slightly more complex entity.

- A weighted sum of the *within-imputation* variance, W, and the *between-imputation* variance, B.

$$W = \frac{1}{M}\sum_{m=1}^{M}\widehat{SE}_{Q,m}^{2}$$

$$B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Q}_m - Q^*\right)^2$$

$$T = W + \left(1 + M^{-1}\right)B$$

$$= W + B + \frac{B}{M}$$

Inference with MI Estimates

After computing Q^* and T, we combine them in the usual way to get test statistics and confidence intervals.

$$t = \frac{Q^* - Q_0}{\sqrt{T}}$$

$$CI = Q^* \pm t_{crit}\sqrt{T}$$

We must take care with our *df*, though.

$$df = (M-1)\left[1 + \frac{W}{(1+M^{-1})\,B}\right]^2$$

Fraction of Missing Information

Earlier today, we briefly discussed a very desirable measure of nonresponse: *fraction of missing information* (FMI).

$$FMI = \frac{r + \frac{2}{(df+3)}}{r+1} \approx \frac{(1+M^{-1})B}{(1+M^{-1})B+W} \rightarrow \frac{B}{B+W}$$

where

$$r = \frac{(1+M^{-1})B}{W}$$

The FMI gives us a sense of how much the missing data (and their treatment) have influence our parameter estimates.

- We should report the FMI for an estimated parameter along with other ancillary statistics (e.g., t-tests, p-values, effect sizes, etc.).

Special Pooling Considerations

The Rubin (1987) pooling rules only hold when the parameter of interest, Q, follows an approximately normal sampling distribution.

- For substantially non-normal parameters, we may want to transform before pooling and back-transform the pooled estimate.

The following table, reproduced from van Buuren (2018), shows some recommended transformations.

Statistic	**Transformation**	**Source**
Correlation	Fisher's z	Schafer (1997)
Odds ratio	Logarithm	Agresti (2013)
Relative risk	Logarithm	Agresti (2013)
Hazard ratio	Logarithm	Marshall et al. (2009)
R^2	Fisher's z on square root	Harel (2009)
Survival probabilities	Complementary log-log	Marshall et al. (2009)
Survival distribution	Logarithm	Marshall et al. (2009)

Pooling Predictions

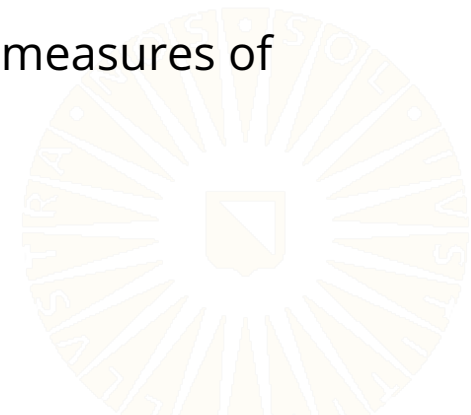
When doing an MI-based analysis, we generally want to pool results as late as possible in the analytic process.

- This pattern also holds when doing prediction with MI data (Wood, Royston, & White, 2015).
- When doing prediction, we pool the M sets of predictions.
  - We don't generate predictions using the pooled parameters.
  - *Caveat:* For GLMs, we pool predictions before applying the inverse link function.
- When pooling fit measures based on predictions (e.g., MSE), we pool the M estimates of fit.
  - We don't generate fit values using pooled predictions or parameters.
- Variability between the M predictions (or any estimates derived therefrom) quantifies uncertainty due to missing data.

Pooling Predictions

According to Wood et al. (2015), the most natural approach also tends to perform best:

1. Train the prediction model on each of the M imputed datasets separately.
2. Generate M sets of predictions by submitting the fully observed future data to the M models from above.
3. Average the M sets of predictions into a single vector of predicted values.
    - When estimating prediction error, calculate M separate measures of error, and pool these estimates.



Pooling Predictions

To cross-validate predictive models with MI data, we have a few options:

1. We can simply impute the entire sample before splitting and run the cross-validation procedure on each of the M imputed datasets.
2. We can split the sample first, train the imputation model on the training set, and also use this imputation model to generate imputations for the validation data.
3. We can train separate imputation models on the training and validation data.
    - When generating the validation-set predictions, we need to cross the training- and validation-set imputations.
    - I.e., for $M_1 = 10$ sets of training-set estimates and $M_2 = 10$ validation-set imputations, we'll have $M_1 \times M_2 = 100$ predictions.

Donor-Based Methods

Model-Based vs. Donor-Based Methods

They types of MI we've discussed above are all *model-based*.

- The imputations are randomly sampled from an estimated distribution of the missing values (i.e., a probability *model* of the missing data).

Model-based methods are theoretically ideal when the missing data truly follow the chosen distribution.

- If the missing data do not follow the model, performance suffers.

Sometimes, the solution is to employ a different probability model.

- We'll see this approach when we discuss MI for categorical variables.

Model-Based vs. Donor-Based Methods

If we're not able to choose a sensible distribution for the missing data, we can use *Donor-Based Methods*.

- Imputations are sampled from a pool of matched observed cases.
- The empirical distribution of the observed data is preserved.

One particularly useful donor-based method is *Predictive Mean Matching* (Little, 1988).

- The cases that make up the donor pool are matched based on their predicted outcome values.

Predictive Mean Matching: Procedure

Suppose we want to generate M imputations for an incomplete variable, Y, using some set of predictors, $\mathbf{X}$.

1. Regress Y_{obs} onto $\mathbf{X}_{obs}$ and compute the conditional mean of Y_{obs}:
   - $\hat{\mu} = \mathbf{X}_{obs}\hat{\beta}$
2. Do a Bayesian linear regression of Y_{obs} onto $\mathbf{X}_{obs}$ and sample M values of the posterior predicted mean of Y_{mis}:
   - $\tilde{\mu}_m = \mathbf{X}_{mis}\tilde{\beta}_m$.
3. Compute M sets of the matching distances:
   - $d(i,j)_m = (\tilde{\mu}_{mi} - \hat{\mu}_j)^2, \;\; i = 1, 2, \ldots N_{mis}, \;\; j = 1, 2, \ldots, N_{obs}$.

Predictive Mean Matching: Procedure

4. Use each $d(i,j)_m$ to construct N_{mis} donor pools.
   - Find the K (e.g., $K \in \{3, 5, 10\}$) cases with the smallest values of $d(1,j)_m$, $d(2,j)_m$, ..., $d(N_{mis},j)_m$.
5. For $m = 1, 2, \ldots, M$, select the final donor cases by randomly sampling a single observation from each of the N_{mis} donor pools defined in Step 4.
6. For each of the M imputations replace the missing values in Y with the donor data selected in Step 5.

Pros and Cons of Predictive Mean Matching

PMM tends to work well with continuous, non-normal variables.

- Relatively robust to misspecification of the imputation model
- Imputed values are always valid

PMM does have some important limitations.

- In small samples, the same donor cases can be re-used many times.
- PMM cannot extrapolate beyond the observed range of the data.
- PMM cannot be used with some variable types.
  - Nominal variables
- PMM may perform poorly when the number of predictor variables is small.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Harel, O. (2009). The estimation of r^2 and adjusted r^2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics*, *36*(10), 1109–1118. doi: 10.1080/02664760802553000

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296. doi: 10.1080/07350015.1988.10509663

Marshall, A., Altman, D. G., Holder, R. L., & Royston, P. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. *BMC Medical Research Methodology*, *9*(57). doi: 10.1186/1471-2288-9-57

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 519). New York, NY: John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data* (Vol. 72). Boca Raton, FL: Chapman & Hall/CRC.

References

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Boca Raton, FL: CRC Press.

Wood, A. M., Royston, P., & White, I. R. (2015). The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical Journal*, *57*(4), 614–632. doi: 10.1002/bimj.201400004