Univariate Multiple Imputation

Utrecht University Winter School: Missing Data in R

Utrecht University

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

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.

*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).

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")
```

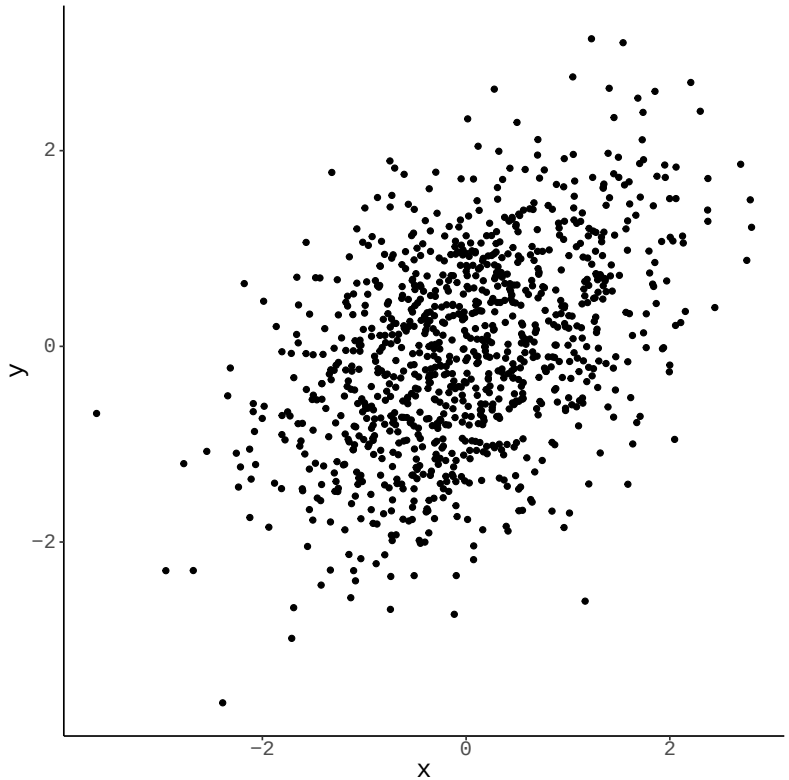
Simulate Some Toy Data

```
## 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)
```

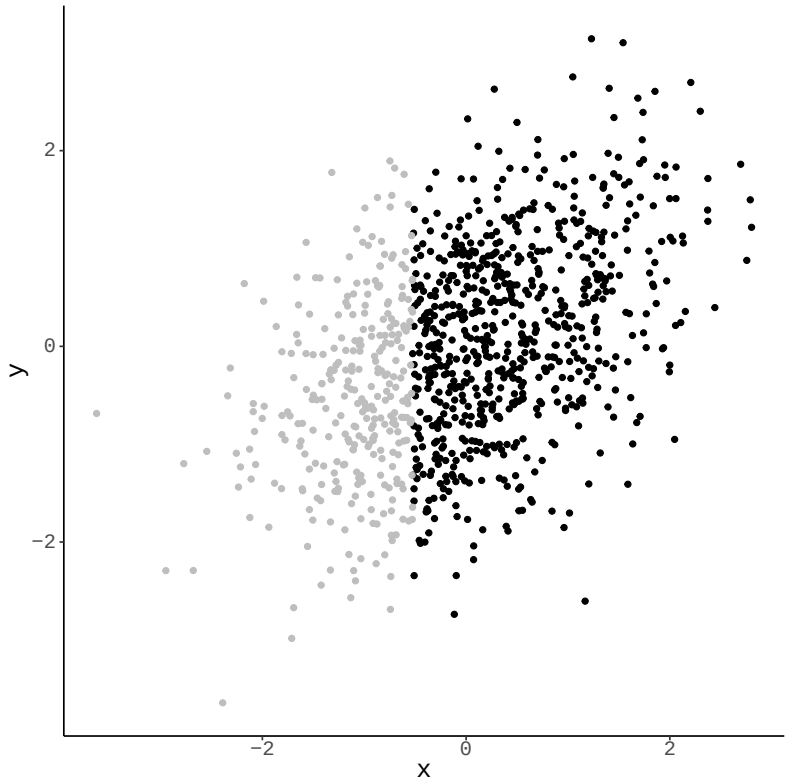
```
       y      x
1 -0.961 -0.912
2  1.467  0.667
3 -0.361 -0.017
4  0.928 -0.447
5 -2.292 -2.678
```



Look at the Data

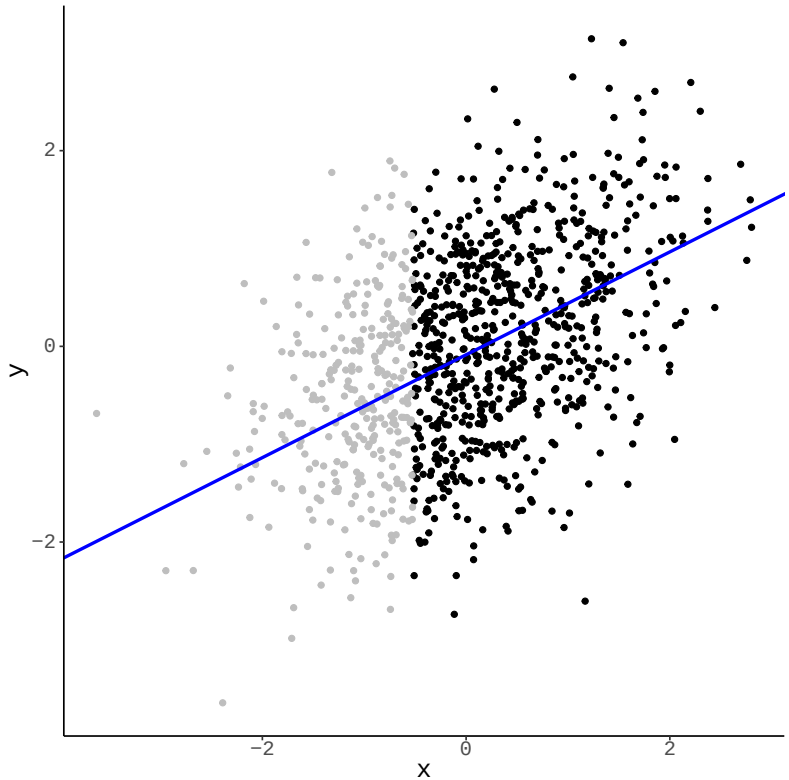
```
head(dat1, n = 5) %>% round(3)
```

```
      y      x
1    NA -0.912
2 1.467  0.667
3 -0.361 -0.017
4 0.928 -0.447
5    NA -2.678
```



Expected Imputation Model Parameters

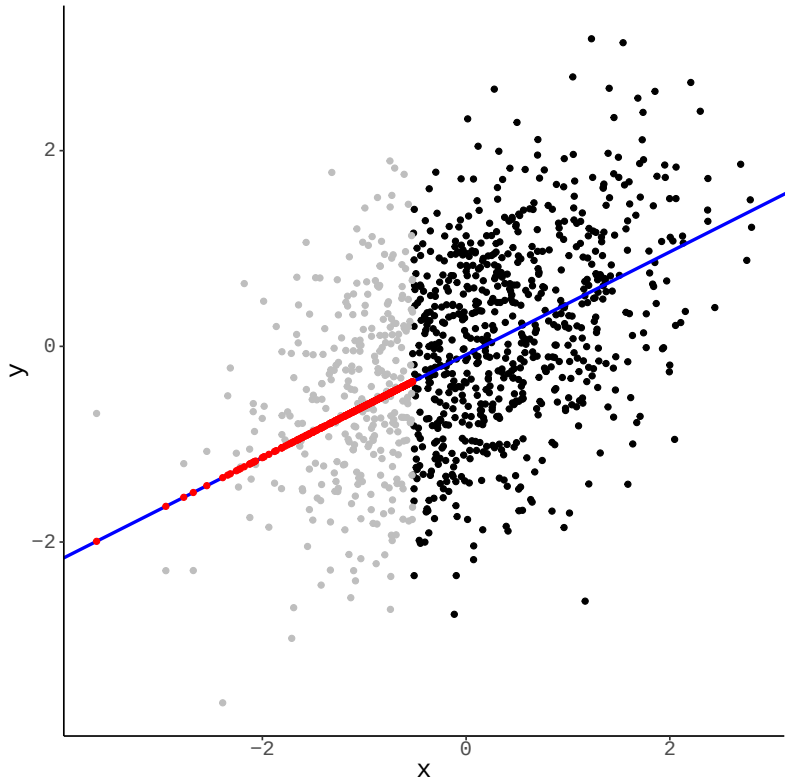
```
lsFit <- lm(y ~ x, data = yObs)

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

as.matrix(beta)

                    [,1]
(Intercept) -0.08610404
x            0.52564595

sigma

[1] 0.9080502
```



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)
```

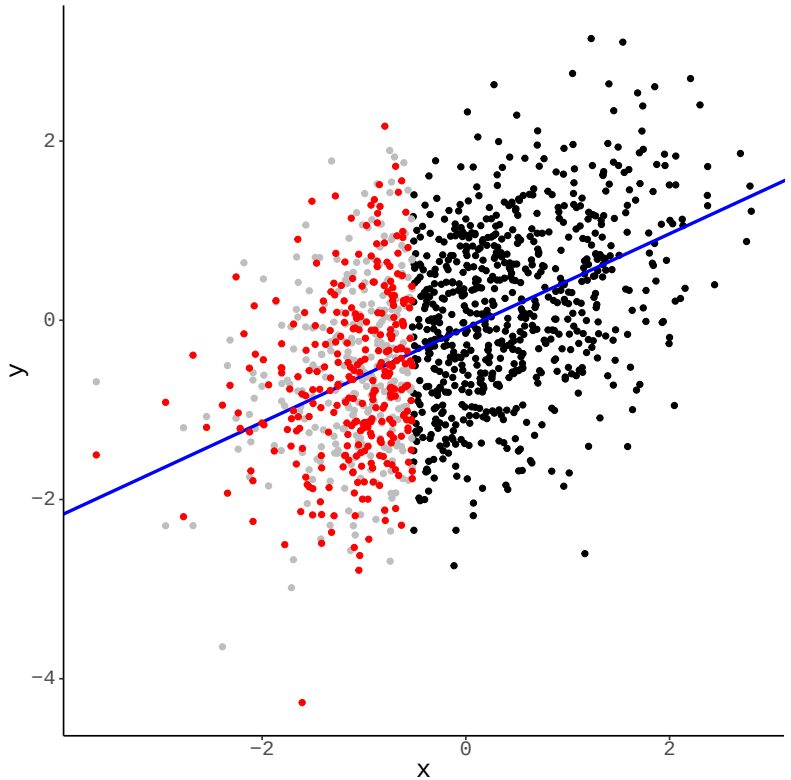
```
       y      x
1 -0.566 -0.912
2  1.467  0.667
3 -0.361 -0.017
4  0.928 -0.447
5 -1.494 -2.678
```



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 -0.885 -0.912
2  1.467  0.667
3 -0.361 -0.017
4  0.928 -0.447
5 -0.390 -2.678
```



Setting Up Proper MI

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.

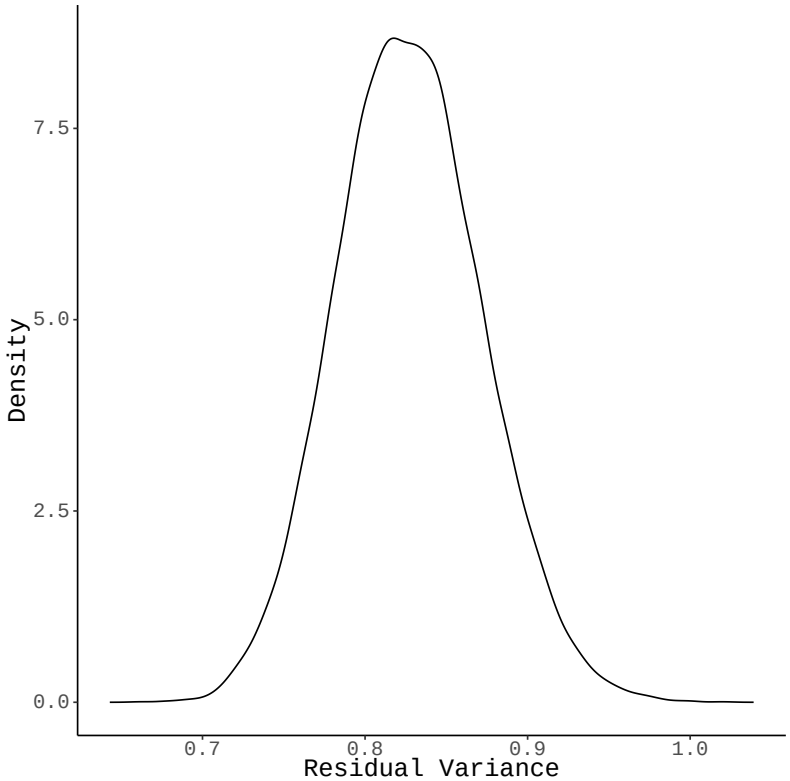
Our imputation model is simply a linear regression model:

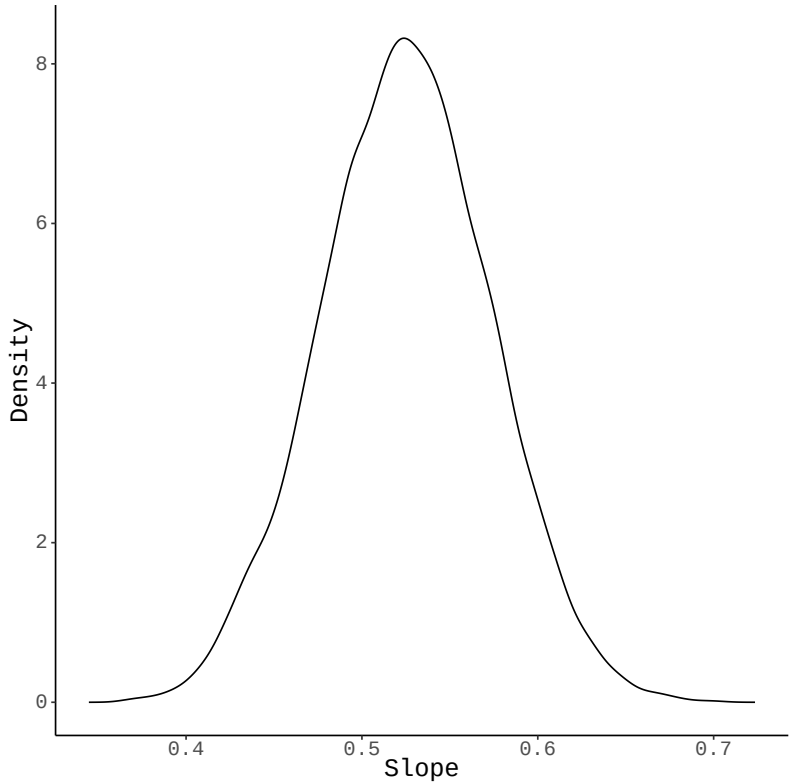
$$Y = \mathbf{X}\beta + \varepsilon$$

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

Visualizing MI

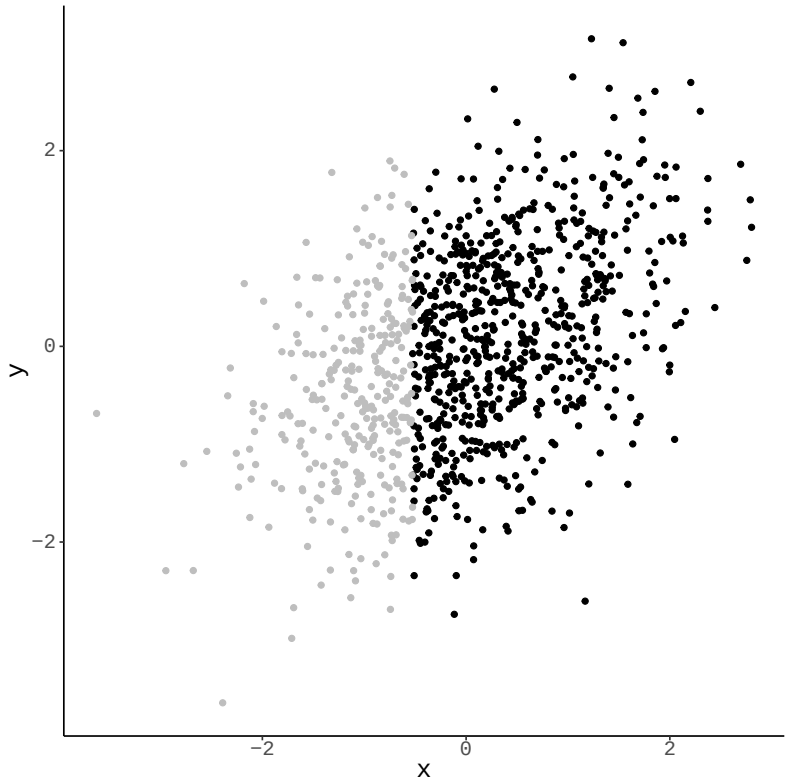
Use Bayesian simulation to estimate posterior distributions for the imputation model parameters:





Visualizing MI

Recall the incomplete data from the single imputation examples.



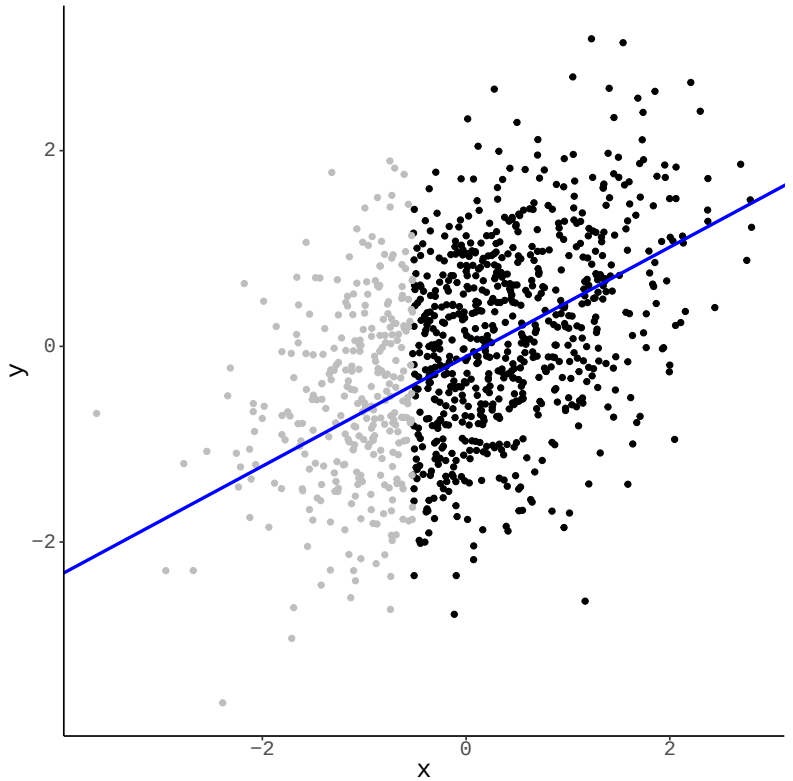
Visualizing MI

Sample values of β_0 and β_1:

- $\beta_0 = -0.105$
- $\beta_1 = 0.56$

Define the predicted best-fit line:

$\hat{Y}_{mis} = -0.105 + 0.56X_{mis}$



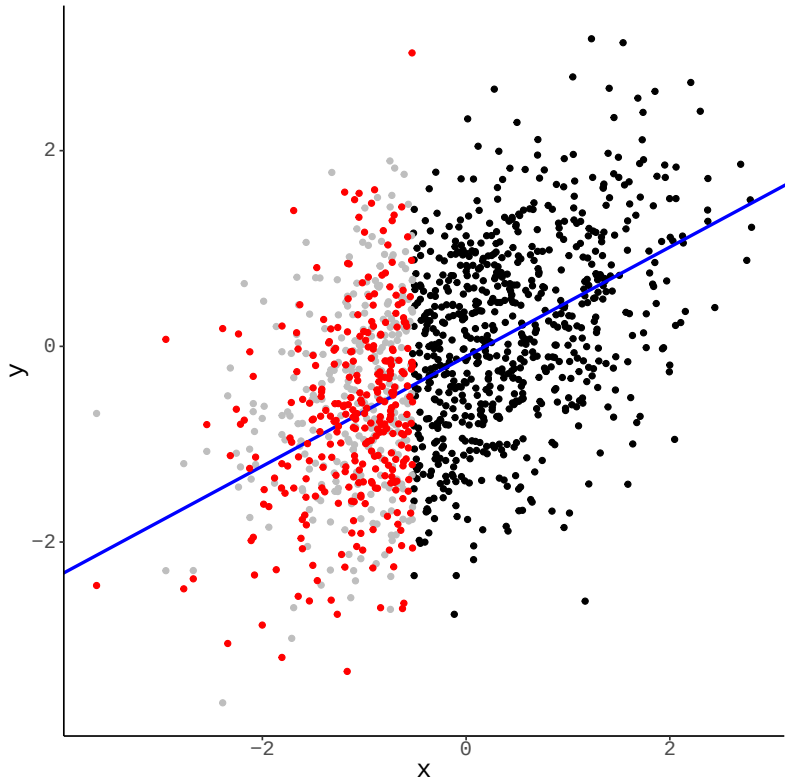
Visualizing MI

Sample a value of σ^2:

- $\sigma^2 = 0.849$

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

$$Y_{imp} = \hat{Y}_{mis} + \varepsilon$$
$$\varepsilon \sim \mathrm{N}(0, 0.849)$$



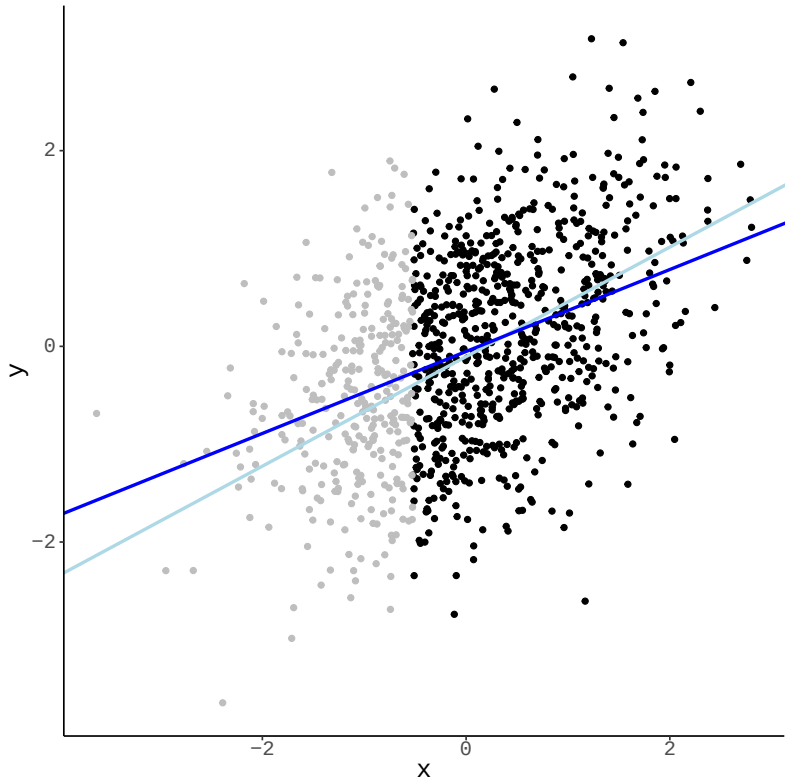
Visualizing MI

Sample values of β_0 and β_1:

- $\beta_0 = -0.053$
- $\beta_1 = 0.419$

Define the predicted best-fit line:

$\hat{Y}_{mis} = -0.053 + 0.419X_{mis}$



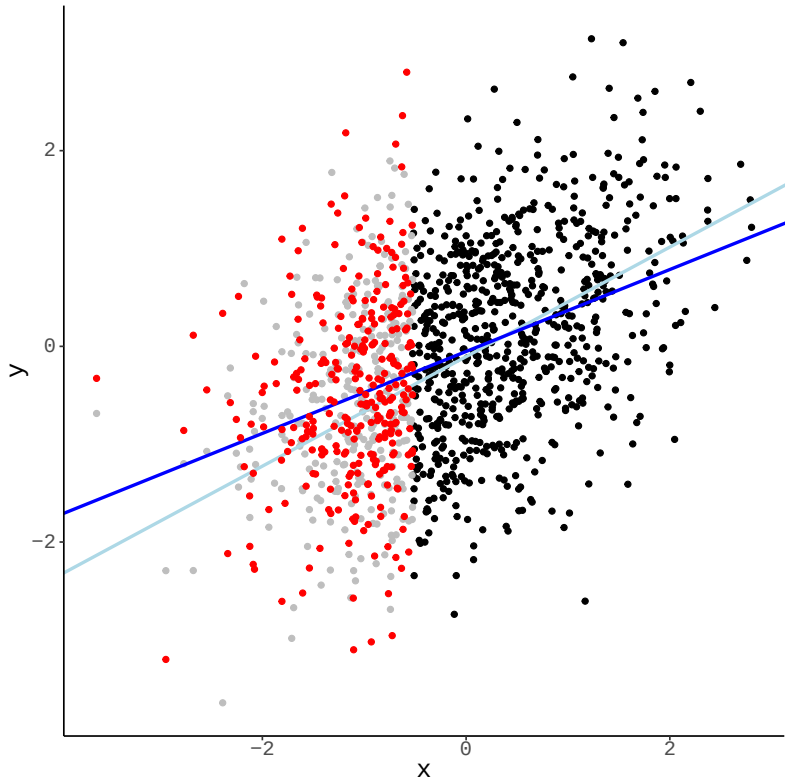
Visualizing MI

Sample a value of σ^2:

- $\sigma^2 = 0.888$

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

$$Y_{imp} = \hat{Y}_{mis} + \varepsilon$$
$$\varepsilon \sim \mathrm{N}(0, 0.888)$$



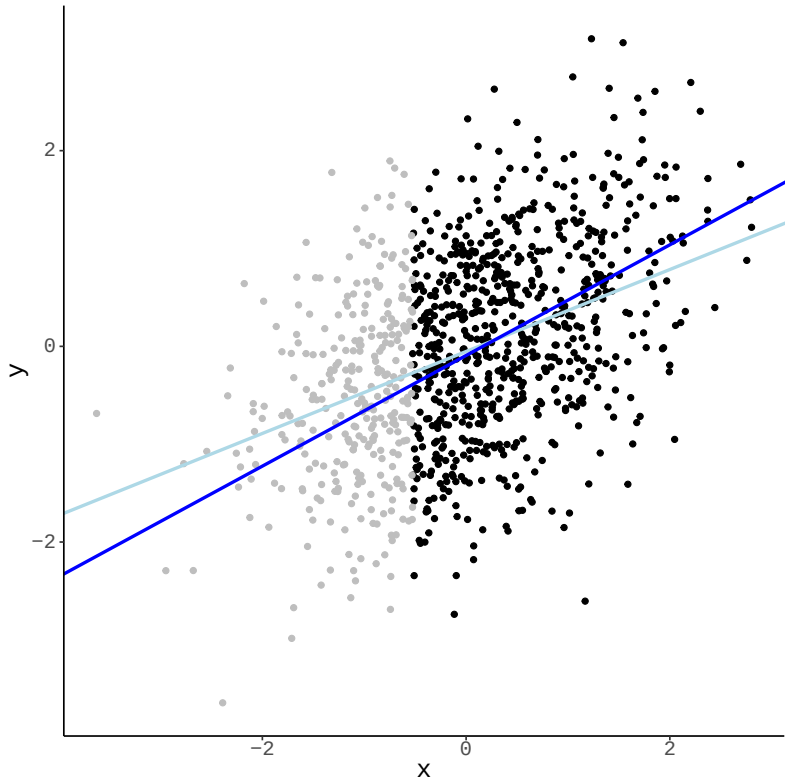
Visualizing MI

Sample values of β_0 and β_1:

- $\beta_0 = -0.093$
- $\beta_1 = 0.565$

Define the predicted best-fit line:

$\hat{Y}_{mis} = -0.093 + 0.565X_{mis}$



Visualizing MI

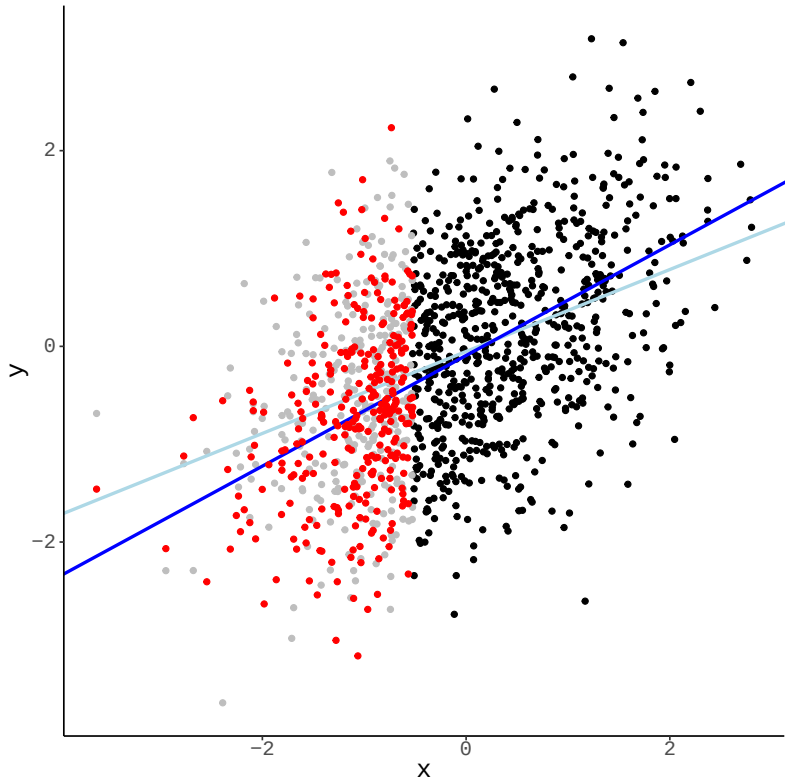
Sample a value of σ^2:

- $\sigma^2 = 0.819$

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

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

$$\varepsilon \sim \mathrm{N}(0, 0.819)$$



MI-Based Analysis

Doing MI-Based Analysis

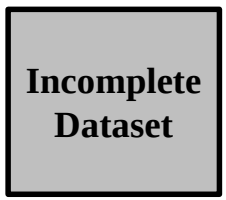
An MI-based data analysis consists of three phases:

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

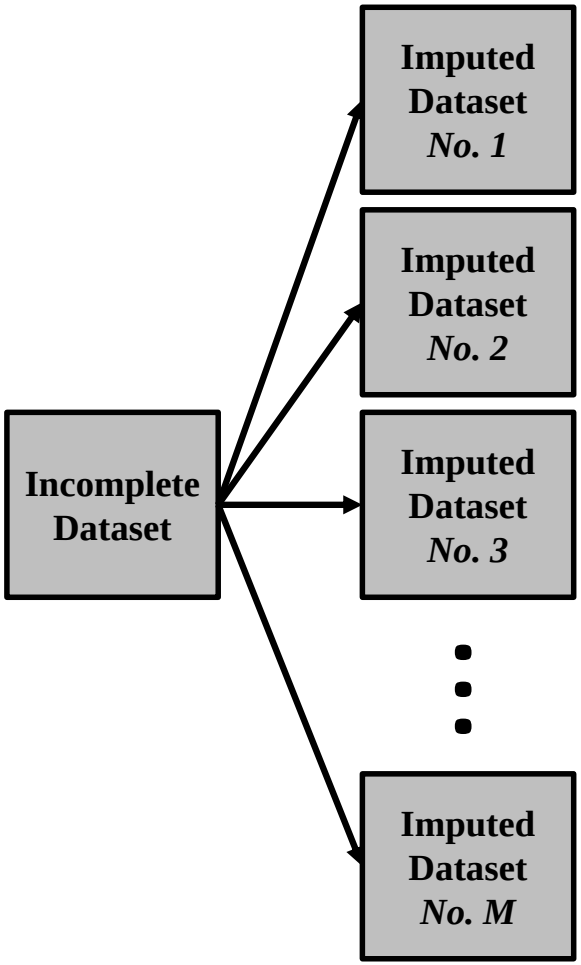
2. The analysis phase
   - Estimate M replicates of your analysis model.
   - Fit the same model to each of the M datasets from Step 1.

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.

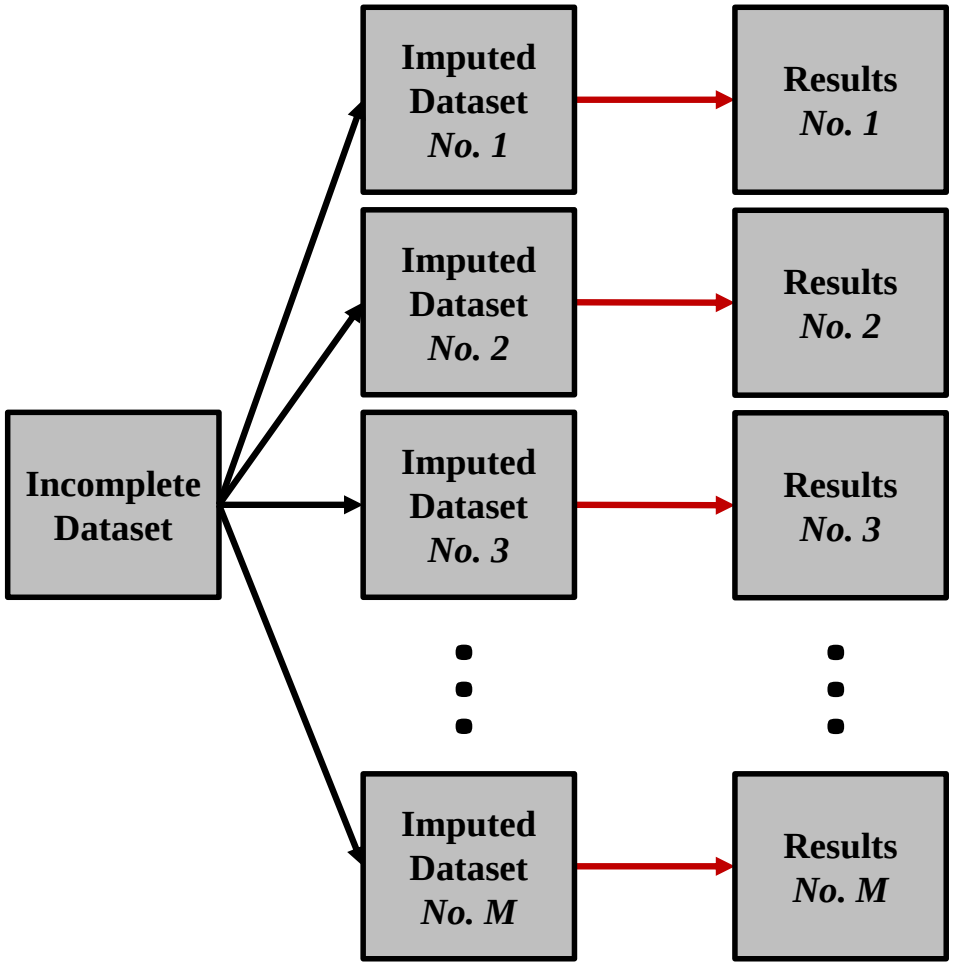
MI-Based Analysis



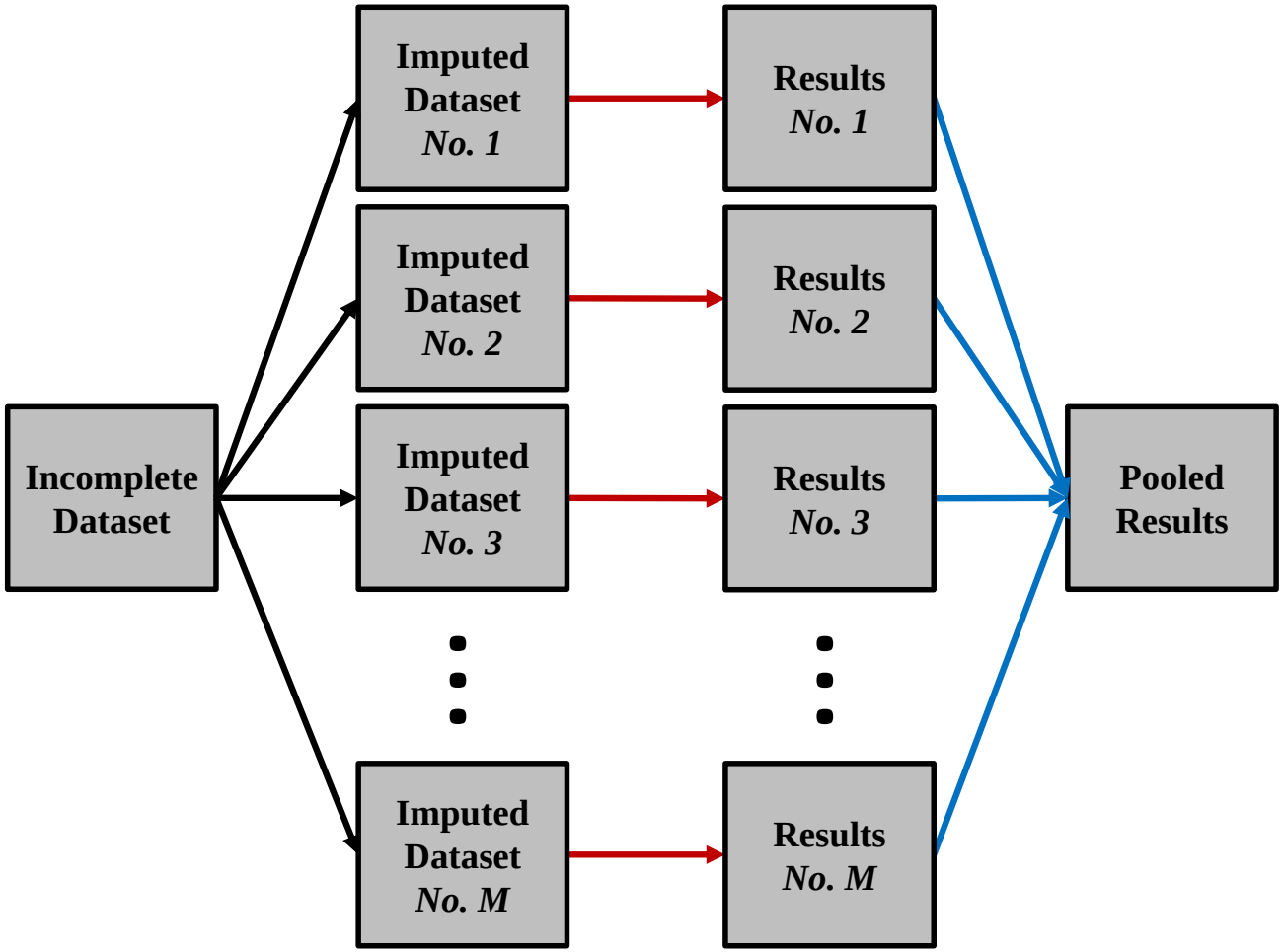
MI-Based Analysis



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)

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

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.

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