Data Cleaning Tutorial: imputation

Mark van der Loo

https://github.com/markvanderloo/UFPEL2019

Try the code

`03valid/impute.R`

Imputing data

Need to specify

- Imputation method
- Variable(s) to impute
- Variables used as predictor

Simputation's goal

Easy to experiment, robust enough for production.

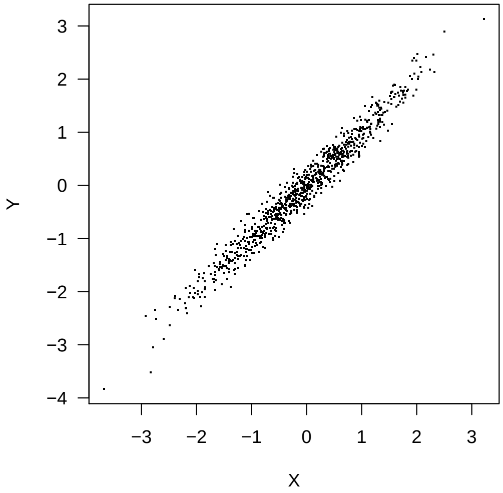
Simputation interface

```
impute_<model>(data, imputed_variables ~ predictors, ...)
```

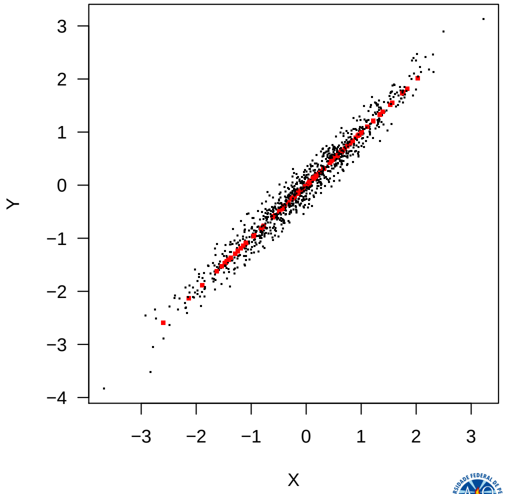
Imputing data with simputation

`<model>`	description
`proxy`	copy (transformation of) other variable(s)
`median`	(group-wise) median
`rlm`, `lm`, `en`	(robust) linear model, elasticnet regression
`cart`, `rf`	Classification And Regression Tree, RandomForest
`em`, `mf`	EM-alogithm (multivariate normal) `missForest`
`knn`	*k* nearest neighbours
`shd`, `rhd`	sequential, random, hot-deck
`pmm`	predictive mean matching
`impute_model`	use pre-trained model

Imputation of the mean



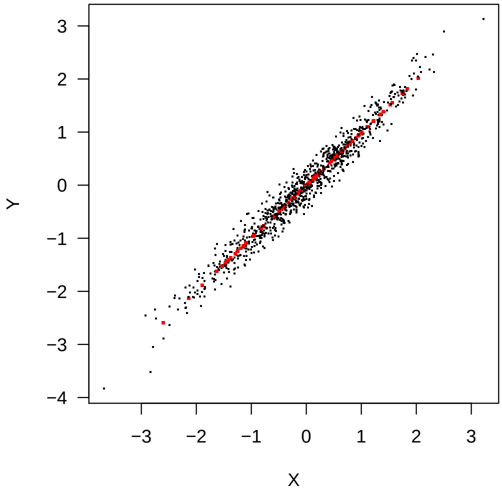
10% missing in Y



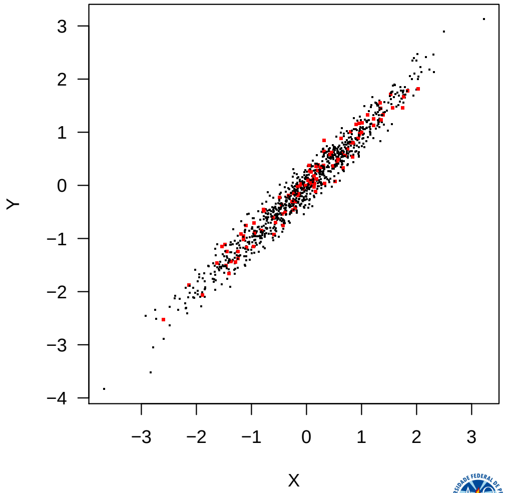
Imputation with model Y = a + bX

https://github.com/markvanderloo/UFPEL2019

Adding a random residual



Imputation with model Y = a + bX



Imputation with Y = a + bX + e

https://github.com/markvanderloo/UFPEL2019

Adding a random residual with `simputation`

Example

```
impute_rlm(companies, other.rev ~ turnover
         , add_residual = "normal")
```

Options

- “none”‘: (default)
- “normal”‘: from $N(0, \hat{\sigma})$
- “observed”‘: from observed residuals

Chaining methods

Example

```
companies %>%
  impute_lm(turnover ~ staff + profit) %>%
  impute_lm(turnover ~ staff)
```

Assignment

1. Read `errors_located.csv` (`stringsAsFactors=FALSE`)
2. Make a separete data frame, selecting columns 7–14 (`staff–vat`)
3. Implement the following imputation sequence:
   - Impute `turnover` by copying the `vat` variable (`impute_proxy`)
   - Impute `staff` with a robust linear model based on `staff.costs`
   - Impute `staff` with a robust linear model based on `total.costs`
   - Impute `profit` as `total.rev - total.costs` (`impute_proxy`)
   - Impute everything else using `missForest` (formula: `. ~ .`)

More on missing data and (s)imputation

Missing data

WHOLE MILK
Milk
MISSING

Name : Value Six
Missing Since : 30-06-2014
Last seen at : input.csv

If you have seen Value Six or know someone who has, please contact your local statistician at +31 415 926 53

Missing data

Reasons

- nonresponse, data loss
- Value is observed but deemed wrong and erased

Solutions

- Measure/observe again
- Ignore
- Take into account when estimating
- **Impute**

Missing data mechanisms

Missing comletely at Random (MCAR)

Missingness is totally random.

Missing at Random (MAR)

Missingness probability can be modeled by other variables

Not Missing at Random (NMAR)

Missingness probability depends on missing value.

You can't tell the mechanism from the data

NMAR can look like MCAR

Given Y, X independent. Remove all $y \geq y^*$. Observer 'sees' no correlation between missingness and values of X: MAR.

NMAR can look like MAR

Given Y, X with $\text{Cov}(Y, X) > 0$. Remove all $y \geq y^*$. Observer 'sees' that higher X correlates with more missings in Y: MCAR.

Dealing with missing data mechanisms

Missing comletely at Random (MCAR)

Model-based imputation

Missing at Random (MAR)

Model-based imputation

Not Missing at Random (NMAR)

No real solution.

Imputation methodology

Model based

Estimate a value based on observed variables.

Donor-imputation

Copy a value from a record that you did observe.

The simputation package

Provide

- a *uniform interface*,
- with *consistent behaviour*,
- across *commonly used methodologies*

To facilitate

- experimentation
- configuration for production

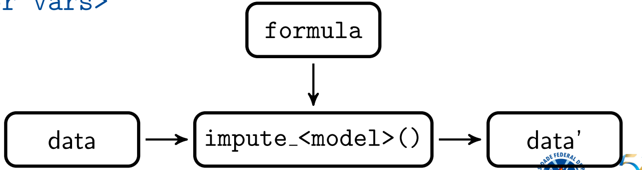
The simputation package

An imputation prodedure is specified by

1. The variable to impute
2. An imputation model
3. Predictor variables

The simputation interface

```
impute_<model>(data
  , <imputed vars> ~ <predictor vars>
  , [options])
```



Chaining methods

```r
ret %>%
  impute_rlm(other.rev ~ turnover) %>%
  impute_rlm(other.rev ~ staff) %>% head(3)
```

```
##   staff turnover other.rev total.rev staff.costs total.costs profit vat
## 1    75       NA  64.88174      1130          NA       18915  20045  NA
## 2     9     1607  17.25247      1607         131        1544     63  NA
## 3    NA     6886 -33.00000      6919         324        6493    426  NA
```

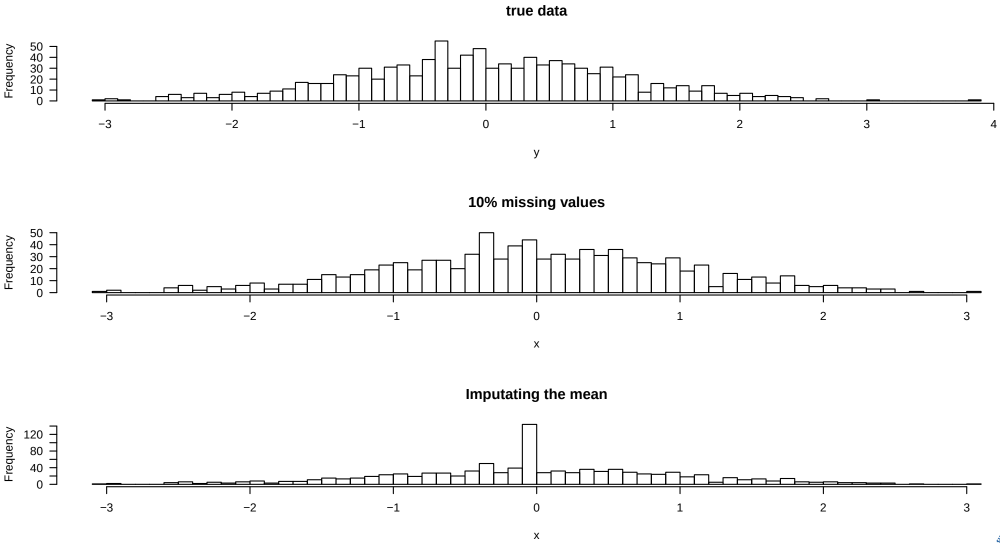
Example: Multiple variables, same predictors

```
ret %>%
  impute_rlm(other.rev + total.rev ~ turnover)

ret %>%
  impute_rlm( . - turnover ~ turnover)
```

Example: grouping

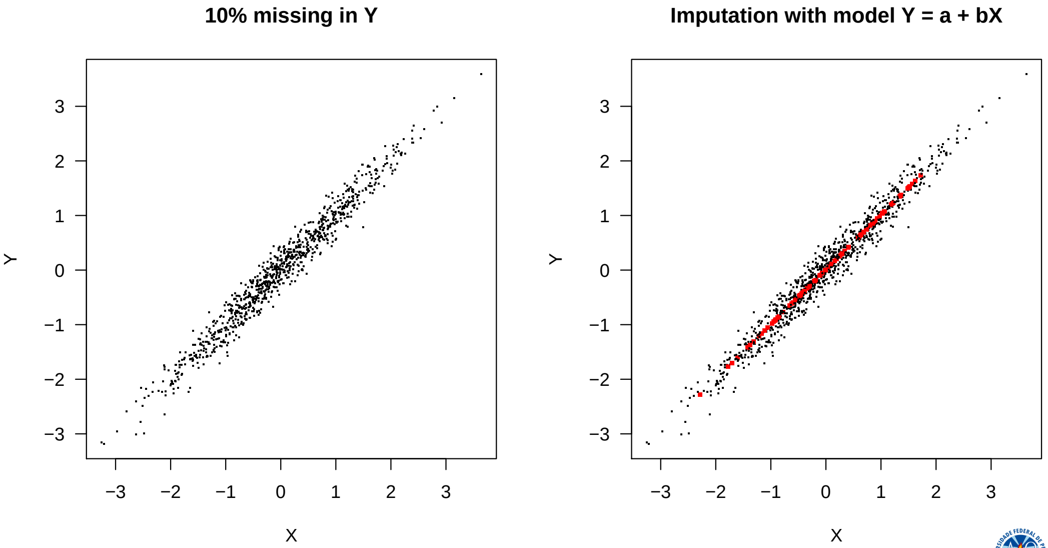
```r
retailers %>% impute_rlm(total.rev ~ turnover | size)

# or, using dplyr::group_by
retailers %>%
  group_by(size) %>%
  impute_rlm(total.rev ~ turnover)
```

Imputation and univariate distribution



https://github.com/markvanderloo/UFPEL2019

Imputation and bivariate distribution



10% missing in Y

Imputation with model Y = a + bX

Adding a random residual

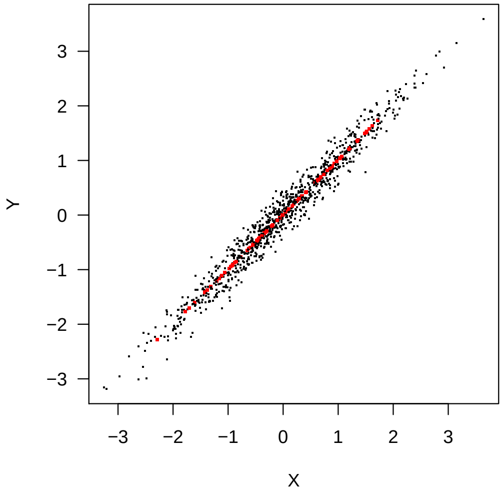
$$\hat{y}_i = \hat{f}(X_i) + \varepsilon_i$$

- $\hat{y}_i$ estimated value for record i
- $\hat{f}(X_i)$ model value
- ε_i random perturbation
  - Either a residual from the model training
  - OR sampled from $N(0, \hat{\sigma})$

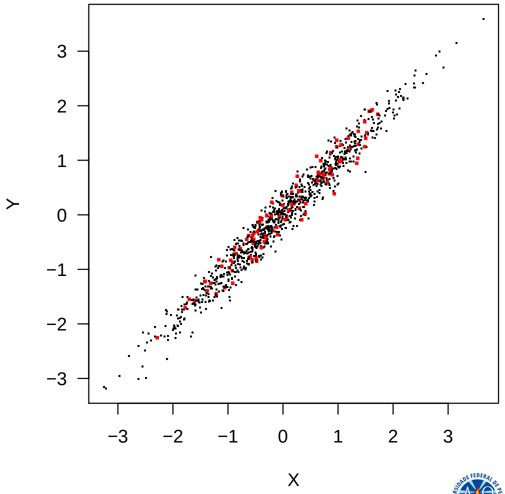
+ Better (multivariate) distribution

– Less reproducible

Adding a random residual



Imputation with model Y = a + bX



Imputation with Y = a + bX + e

Adding a residual with `simputation`

Code

```r
ret %>%
  impute_rlm(other.rev ~ turnover
    , add_residual = "normal") %>% head(3)
```

Options

- `add_residual = "none"`: (default)
- `add_residual = "normal"`: from $N(0, \hat{\sigma})$
- `add_residual = "observed"`: from observed residuals

Compute the variance of `other.rev` after each option.

Ten models.

1. Impute a proxy

$$\hat{\boldsymbol{y}} = \boldsymbol{x} \text{ or } \boldsymbol{y} = f(\boldsymbol{x}),$$

where $\boldsymbol{x}$ is another (proxy) variable (e.g. VAT value for turnover), and f a user-defined (optional) transformation.

```
# simputation
impute_proxy()
```

2. Linear model

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta},$$

where

$$\hat{\beta} = \arg\min_{\beta} \sum_{i} \epsilon_i^2$$

```
# simputation:
impute_lm()
```

3. Regularized linear model (elasticnet)

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta},$$

where

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2}\sum_i \epsilon_i^2 + \lambda\left[\frac{1-\alpha}{2}\|\beta^*\|^2 + \alpha\|\beta^*\|_1\right]$$

- $\alpha = 0$ (Lasso) $\cdots$ $\alpha = 1$ (Ridge)
- β^*: β w/o intercept.

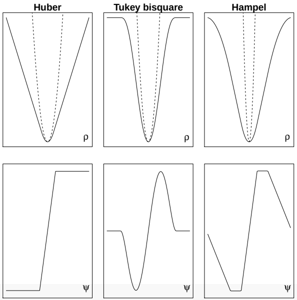
```
# simputation:
impute_en()
```

4. *M*-estimator

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta},$$

where

$$\hat{\beta} = \arg\min_{\beta} \sum_i \rho(\epsilon_i)$$



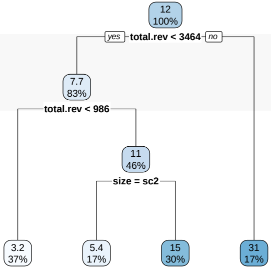
```
# simputation:
impute_rlm()
```

5. Classification and regression tree (CART)

$$\hat{\boldsymbol{y}} = T(\boldsymbol{X}),$$

where T represents a set of binary questions on variables in $\boldsymbol{X}$. There are spare questions for when one of the predictors is missing.

```
# simputation:
impute_cart()
```



6. Random forest

$$\hat{\boldsymbol{y}} = \frac{1}{|\text{Forest}|}\sum_{i\in\text{Forest}} T_i(\boldsymbol{X}),$$

where each T_i is a simple decision tree without spare questions. For categorical $\boldsymbol{y}$, the majority vote is chosen.

```
# simputation
impute_rf()
```

7. Expectation-Maximization

Dataset $\boldsymbol{X} = \boldsymbol{X}_{obs} \cup \boldsymbol{X}_{mis}$. Assume $\boldsymbol{X} \sim P(\boldsymbol{\theta})$.

1. Choose a $\hat{\boldsymbol{\theta}}$.
2. Repeat until convergence:
   2.1 $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = \ell(\boldsymbol{\theta}|\boldsymbol{X}_{obs}) + E_{mis}[\ell(\boldsymbol{X}_{mis}|\boldsymbol{\theta}, \boldsymbol{X}_{obs})|\hat{\boldsymbol{\theta}}]$
   2.2 $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$
3. $\hat{\boldsymbol{X}}_{mis} = \arg\max_{\boldsymbol{X}_{mis}} P(\boldsymbol{X}_{mis}|\hat{\boldsymbol{\theta}})$

```
# simputation (multivariate normal):
impute_em()
```

8. missForest

Dataset $\boldsymbol{X} = \boldsymbol{X}_{obs} \cup \boldsymbol{X}_{mis}$.

1. Trivial imputation of $\boldsymbol{X}_{mis}$ (median for numeric variables, mode for categorical variables)
2. Repeat until convergence:
    2.1 Train random forest models on the completed data
    2.2 Re-impute based on these models.

```
# simputation:
impute_mf()
```

9.a Random hot deck

1. Split the data records into groups (optional)
2. Impute missing values by copying a value from a random record in the same group

```
# simputation
impute_rhd(data, imputed_variables ~ grouping_variables)
```

9.b Sequential hot-deck

1. Sort the dataset
2. For each row in the sorted dataset, impute missing values from the last observed.

```
# simputation
impute_shd(data, imputed_variables ~ sorting_variables)
```

9.c *k*-nearest neighbours

For each record with one or more missings:

1. Find the *k* nearest neighbours (Gower's distance) with observed values
2. Sample value(s) from the *k* records.

```
# simputation
impute_knn(data, imputed_variables ~ distance_variables)
```

10. Predictive mean matching

1. For each variable X_i with missing values, estimate a model $\hat{f}_i$.
2. Estimate all values, observed or not.
3. For each missing value, impute the observed value, of which the prediction is closest to the prediction of the missing value.

```
# simputation: (currently buggy!)
impute_pmm()
```