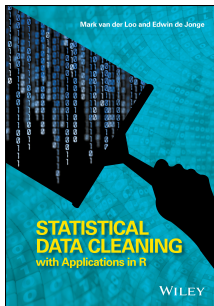# Imputation

Mark van der Loo and Edwin de Jonge

Statistics Netherlands Research & Development
Twitter: @markvdloo @edwindjonge

ENBES|EESW 2019

O3valid/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
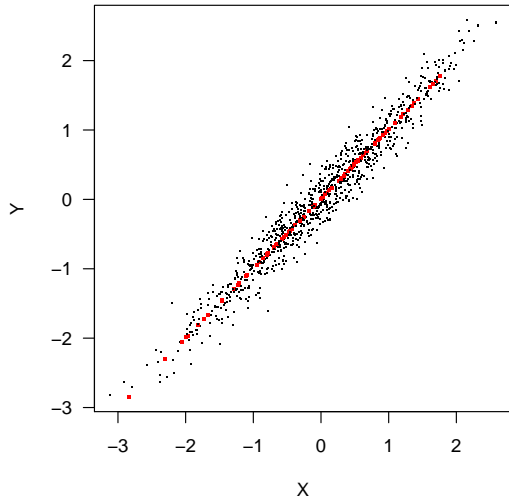

10% missing in Y
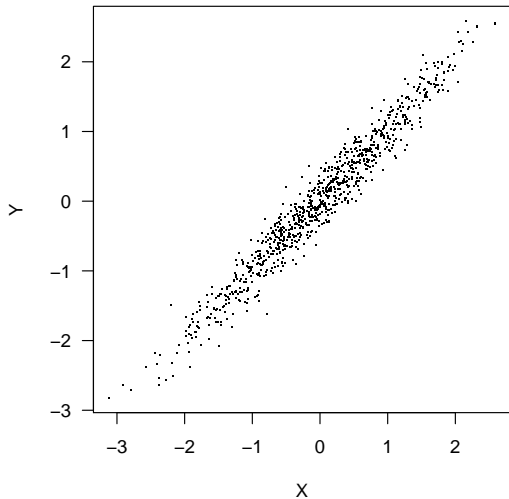
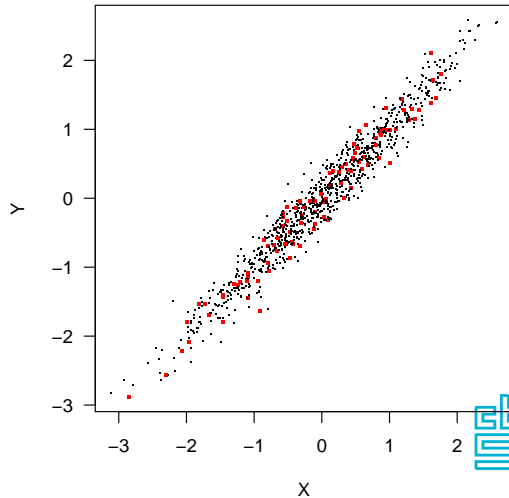**Imputation with model Y = a + bX**
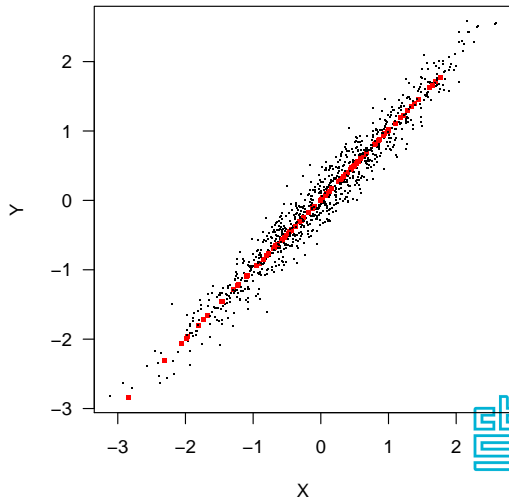
Adding a random residual

Imputation with model Y = a + bX

Imputation with Y = a + bX + e

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 11–14 (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: . ~ .)