



Imputation



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 simulation

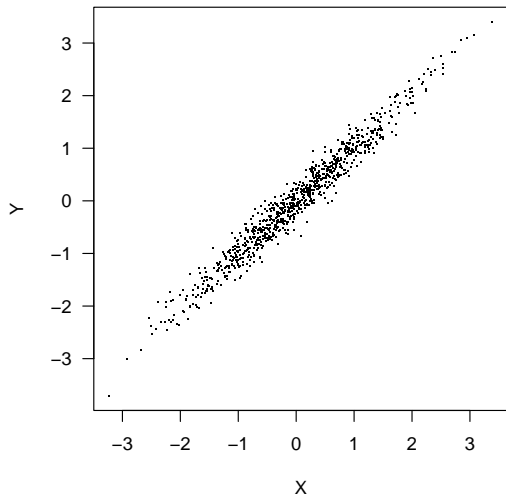


<model>	description
proxy	copy (transformation of) other variable(s)
median	(group-wise) median
rlm, lm, en	(robust) linear model, elasticnet regression
cart, rf	Classification And Regression Tree, RandomForest
em, mf	EM-algorithm (multivariate normal) missForest
knn	k nearest neighbours
shd, rhd	sequential,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

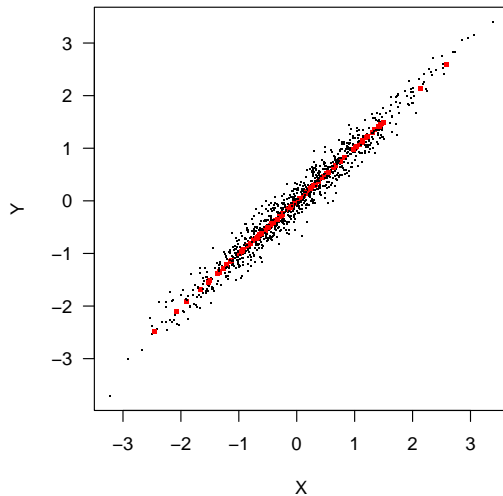
Imputation of the mean



10% missing in Y



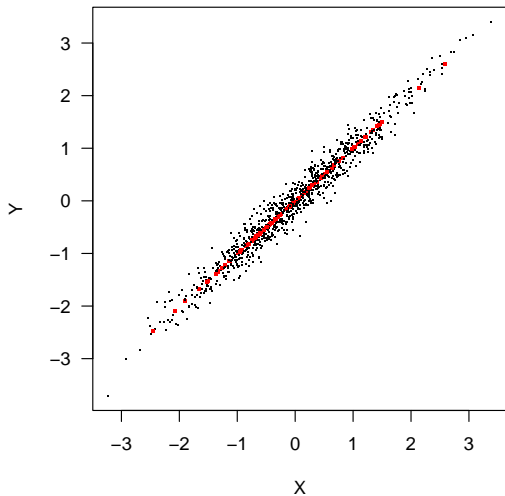
Imputation with model $Y = a + bX$



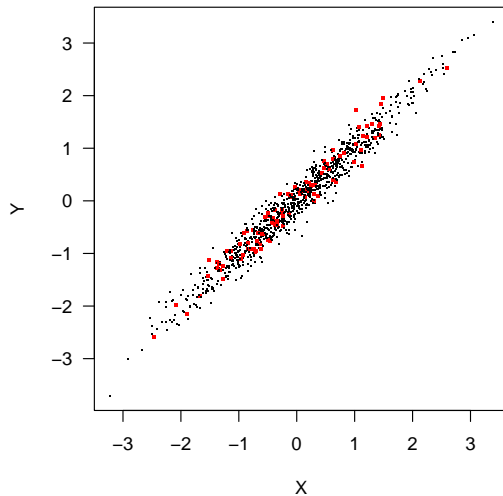
Adding a random residual



Imputation with model $Y = a + bX$



Imputation with $Y = a + bX + e$



Adding a random residual with simulation



Example

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

Options

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



Example

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



More on missing data and (s)imputation



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 completely 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 completely at Random (MCAR)

Model-based imputation

Missing at Random (MAR)

Model-based imputation

Not Missing at Random (NMAR)

No real solution.



Model based

Estimate a value based on observed variables.

Donor-imputation

Copy a value from a record that you did observe.

The simulation package



Provide

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

To facilitate

- experimentation
- configuration for production

The imputation package



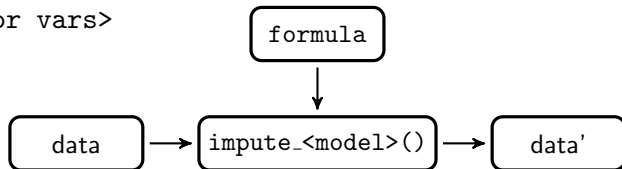
An imputation procedure is specified by

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

The simulation interface



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





```
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