WHAT (NOT) TO DO WITH MISSING DATA?

WHY SHOULD I CARE?!

Ignoring the missingness can result in not being able to calculate statistics, a loss of power, or even worse; wrong conclusions! How to deal with the missingness depends on the missingness mechanism.

WHAT ARE MISSINGNESS MECHANISMS?

Every datapoint has some likelihood to be missing. The process that governs these probabilities is called the "missing data mechanism". We distinguish 3 types: MCAR, MAR & MNAR.

MCAR

Missing Completely At Random, which means that the missingness pattern is totally random.

Congratulations, your data is not biased! You can choose to either remove the missing cases or do multiple imputation.

MAR

Missing At Random, which means that the missingness pattern is not random, but known. The missingness depends on data that is observed.

Next step: correct for the biased-ness in your data

MNAR

Missing Not At Random, which means that the missingness pattern is not random and unknown. The missingness depends on the missing data

REMOVE MISSING CASES

Only analyze the observed data; your conclusions will not be biased! However, less data means less power...

IMPUTATION

If the observed information holds the essence of the missing information, you can predict the unobserved values based on the observed information.

The next four are some common imputation methods. Choose wisely!

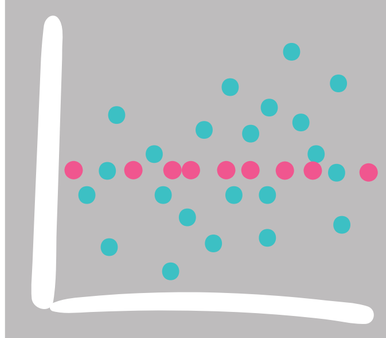
NOPE.

Your data is biased and you cannot correct for the biasedness by using imputation...

Better luck next time!

MEAN IMPUTATION

The missing values are replaced with the mean value

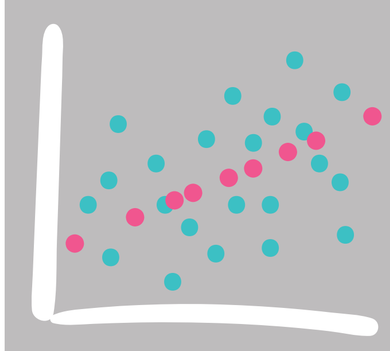


+ unbiased for the mean under MCAR
- biased under MAR
- underestimates the variance
- biases correlations to zero

DON'T USE THIS

REGRESSION IMPUTATION

Impute values on the regression line

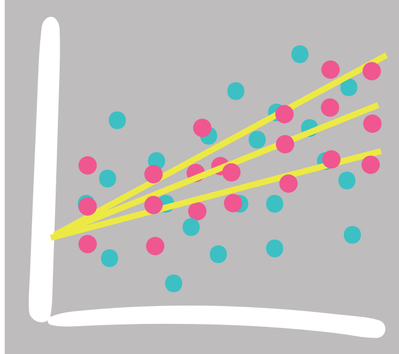


+ good approximation if the expplained variance is high
- artificially increases correlations, resulting in too optimistic p-values & too narrow CI's!

DON'T USE THIS

BAYESIAN REGRESSION IMP.

Imputes values multiple times while adding noise to both the regression coefficient & imputed values

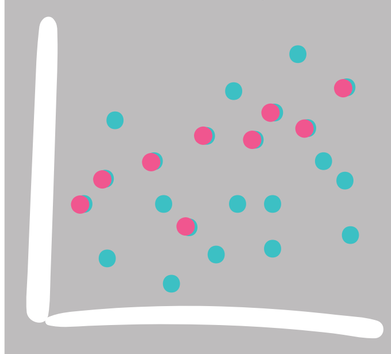


+ takes the uncertainty of the predictions into account (correct point & variance estimation)
- might lead to implausible values

GOOD OPTION

PREDICTIVE MEAN MATCHING

Missing values are replaced by an observed value from another case that is the most similar in multivariate space



+ only imputes values within the range of the observed values
- cannot impute values that were not already observed

GOOD OPTION