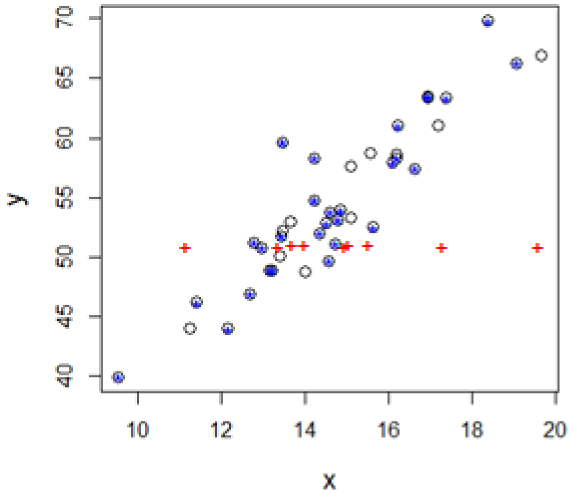
In-class exercise on topic “imputation of missing values”
Solution

Complete case analysis vs mean imputation for a special missing-process

Here, we investigate the association between a continuous predictor x and a continuous outcome y.

a) We look at simulated data where a random process was used to pick which y-values are replaced by NA (missing). Since the data are simulated we can also look at the relationship between x and y before some of the y-values were deleted (see all points - filled and empty circles - in the scatterplot): We can compare that with the scatterplot of the complete data (filled circles).



(i) What kind of missing process do you assume based on the above plot?

MCAR, since the missingness seems not to depend on x or y.

(ii) Do you expect to get different results if you would perform a linear regression fit once with all data (all circles) and once with the data after some y-values were deleted (only filled circles). If yes, describe what will change and why.

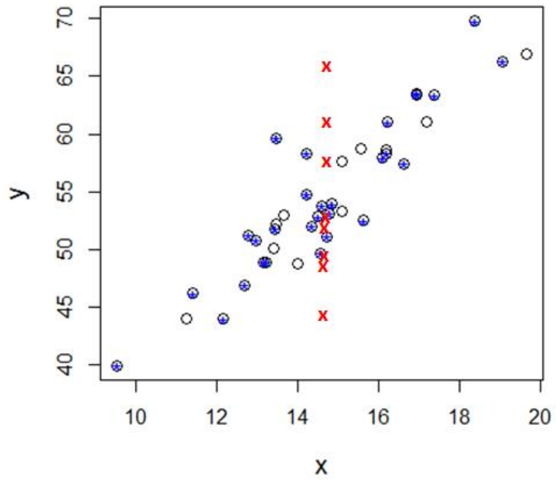
We will get unbiased estimates (more or less same values for the estimated coefficients), however the CI will be wider, since we have less data. That means the power to detect an effect gets lower and we have a smaller chance to detect a slope which is significantly different from zero.

(iii) Now perform a mean-imputation for the missing y-values. This means we replace each missing y-value with the mean of all observed y-values which is here 54. Indicate the position of the imputed data points in the above plot by "+"-Symbols.

Imagine that you perform a linear regression fit with the imputed data set.
Do you expect to get different results if you would perform a linear Regression fit once with all data (all circles) and once with the imputed data set (only filled circles and imputed pluses). If yes, describe what will change.

We will get biased estimates – the slope will be too flat. Also the CI will be changed (because we have wider spread than in the unobserved full data set).

(iv) Imagine now that the missing values (empty circles) are due to a missing x-value. Now perform a mean-imputation for the missing x-values (mean of all observed x-values is 15) and indicate the imputed data points in the plot below with "x"-Symbols. Do you get unbiased estimates for the regression slope?



The estimate for the slope will be biased – we get a too large slope and also the CI is not valid. We are prone to find a significant effect where there might be no effect (or a much smaller) in the unobserved full data set – this corresponds to a type-I error.

Is it always better to work with the complete cases instead of an imputed data set if we want to avoid biases in the estimates?

No! In many situation we get biased estimates if we perform an analysis using only the completely observed data instead of working with an imputed data set (see example in the lectures with simulated data for a linear regression when missingness of x depends on y). However, we should use a good imputation approach and mean imputation is often a bad approach.