Incomplete Data Analysis

V. Inácio de Carvalho & M. de Carvalho

University of Edinburgh

PRR

GOVERNO DE PORTUGAL

Financiado pela União Europeia NextGenerationEU

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Single imputation

Context

- ↪ Rather than discarding cases with missing data, another approach is to **fill in** or '**impute**' missing values.
- ↪ The term **single imputation** comes from the fact that these approaches generate a **single** replacement value for each missing observation. This is in contrast to multiple imputation, which creates several copies of the dataset and imputes each copy with different plausible estimates of the missing values.
- ↪ Single imputation by filling in the missing values leads to a completed dataset that can be subsequently analysed with any statistical method.

Single imputation

Context

- The methods that we will cover are
  - Unconditional mean imputation.
  - Conditional mean imputation/regression imputation.
  - Stochastic regression imputation.
  - Hot deck imputation.
  - Last observation carried forward.

Single imputation

Context

- Whenever a single imputation strategy is used, the standard errors of the estimates tend to be too low. The intuition behind this is that we obviously have considerable uncertainty about the values that are missing, but by choosing a single imputation we are kind of pretending that we know the true value with certainty. That is why multiple imputation is recommended.

- Nevertheless, single imputation is still paid attention in the literature.

- Further, some of the single imputation methods are a step in the right direction and form the basis of some multiple imputation techniques.

Single imputation mechanisms

- We shall use a small bivariate dataset to illustrate ideas.
- 20 chronic patients enrolled in a pain management programme.
- Patients with mid pain are more likely to refuse the depression measure.

Pain severity	Depression
4	NA
6	NA
7	14
7	11
8	NA
9	NA
9	11
10	NA
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

Single imputation

Mean imputation

- In **mean imputation** (also known as **unconditional/marginal mean imputation**) each missing value is filled by the overall mean of the observed values for that variable.
- We may use the mode for categorical data.
- Thus, for our illustrative dataset we would replace each of the 5 missing values by the mean of the depression score of the observed 15 cases. This would be 15.
- Arithmetically, the estimate of the mean after mean imputation must always be the same as the mean for the observed data.

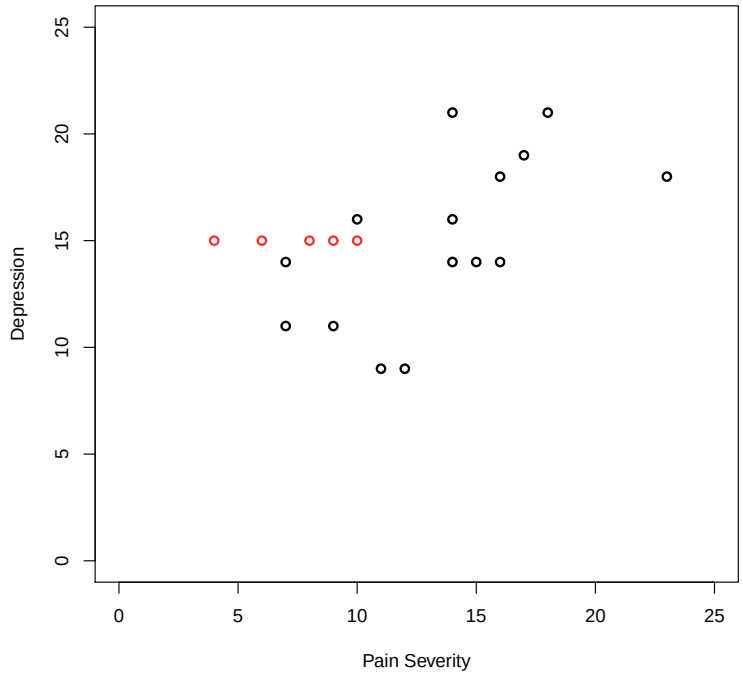
Single imputation

Mean imputation

- ↪ Intuitively, imputing values at the center of the distribution reduces the variability of the data.
- ↪ Thus, it makes sense that mean imputation will attenuate the standard deviation of the variable with missing values.
- ↪ For instance, in our toy example, the standard deviation of the completed variable (after mean imputation) is 3.37, which is smaller than that from the observed data in that variable (3.93).
- ↪ Imputing the mean of a variable also attenuates the magnitudes of covariances and correlations between variables.
- ↪ Remember the covariance formula. Cases with missing values on either one of the variables contribute with a value of zero to the numerator formula.

Single imputation

Mean imputation



- ↪ Notice that the imputed values (red dots) fall directly on a horizontal line, which implies that the correlation is zero for the subset of cases with imputed depression scores.
- ↪ The correlation between the depression and pain measures drops from 0.616 in the observed data to 0.483 in the completed dataset.

Single imputation

Mean imputation

- Mean imputation is fast and a simple fix for the missing data problem.
- However, it will underestimate the variance, attenuate the correlations between variables, bias any almost estimate other than the mean (and all these even under a MCAR mechanism) and bias the estimate of the mean when data are not MCAR.
- Several studies suggest that mean imputation is possibly the worst missing data handling method available.
- Quoting Julie Josse (from her website):

  "In case of MCAR values, it is really preferable to suppress observations that have missing data (indeed, we end up with just a small sample size but the estimators calculated on this sample will be unbiased but necessarily more variable) rather than imputing by the mean which can lead to biased estimators."

Single imputation

Conditional mean imputation

- **Conditional mean imputation** (also known as **regression imputation**) is an improvement on the mean imputation approach since it replaces each missing value with a predicted conditional mean from a regression equation.
- For now, we will focus in the case where only one variable has missing values. The extension to the case where several variables may be subject to missingness will be detailed next week.
- The idea behind this approach is appealing: use information from the complete variables to fill in the incomplete variable.
- Variables tend to be correlated so it makes sense to generate imputed values that borrow information from the observed data.
- The first step is to fit the regression model to the complete cases, where the variable that has missing values is regressed on the fully observed variables.
- The second step plugs in the observed values of the complete variables into the estimated regression equation, thus obtaining predicted values for the missing values.

Single imputation

Conditional mean imputation

- Let us reconsider the bivariate data on pain/depression scores.
- The 15 complete cases were used to estimate the regression of depression score on pain score

$$\text{DEP}_i = \beta_0 + \beta_1 \text{PS}_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2), \quad i = 1, \ldots, 15. \tag{1}$$

- The resulting equation for the predicted values is

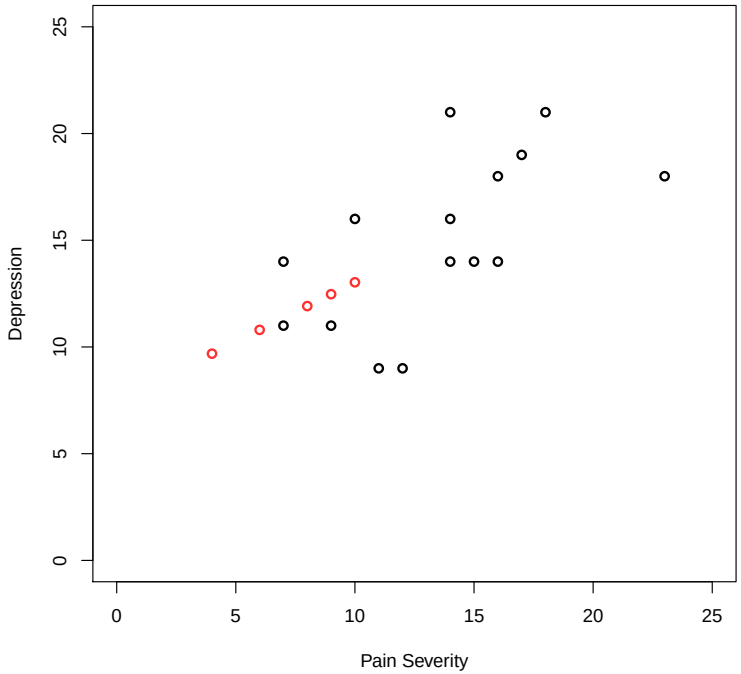
$$\begin{aligned} \widehat{\text{DEP}}_i &= \widehat{\beta_0} + \widehat{\beta_1} \text{PS}_i \\ &= 7.4568 + 0.5574 \text{PS}_i. \end{aligned}$$

- Substituting the appropriate pain scores into the regression equation yields

Pain score	Depression score	Predicted depression score
4	NA	9.686
6	NA	10.801
8	NA	11.916
9	NA	12.473
10	NA	13.031

Single imputation

Conditional mean imputation



- The figure above shows a scatterplot of the imputed data. The ensemble of imputed values vary less than the observed values.
- We can notice immediately that the imputed values fall directly on a regression line with nonzero slope. This implies that the correlation between pain and depression scores is 1 in the subset of cases with imputed values.

Single imputation

Conditional mean imputation

- ↪ It thus turns out that regression imputation suffers from the exact opposite problem as mean imputation because it imputes data with perfectly correlated scores.
- ↪ Consequently, regression imputation overestimates correlations even when the data are MCAR.
- ↪ The correlation for our running example increases from 0.616 in the observed data to 0.706 in the completed dataset.
- ↪ Also, since the imputed values fall on a straight line, the imputed data lack variability that would have been observed had the data been complete.

Single imputation

Stochastic regression imputation

- **Stochastic regression imputation** is a refinement of conditional mean imputation.
- It also uses a regression model to predict the incomplete variables from the complete variables, but it takes an extra step by adding noise to the predictions.
- Adding noise to the imputed values restores lost variability to the data.
- It has been shown that stochastic regression imputation gives unbiased parameter estimates (correlations, regression coefficients, etc) under a MAR missing data mechanism.

Single imputation

Stochastic regression imputation

- In the context of our example, we would have

$$\widehat{\text{DEP}}_i = \widehat{\beta_0} + \widehat{\beta_1}\text{PS}_i + z_i, \qquad z_i \sim \text{N}(0, \widehat{\sigma}^2).$$

- Stochastic regression uses the same basic procedure as standard conditional imputation, so the regression coefficients for our example will be identical to those in regression imputation (as in both cases the basic underlying model is the one in (1)), i.e., $\widehat{\beta}_0 = 7.4568$ and $\widehat{\beta}_1 = 0.5574$.

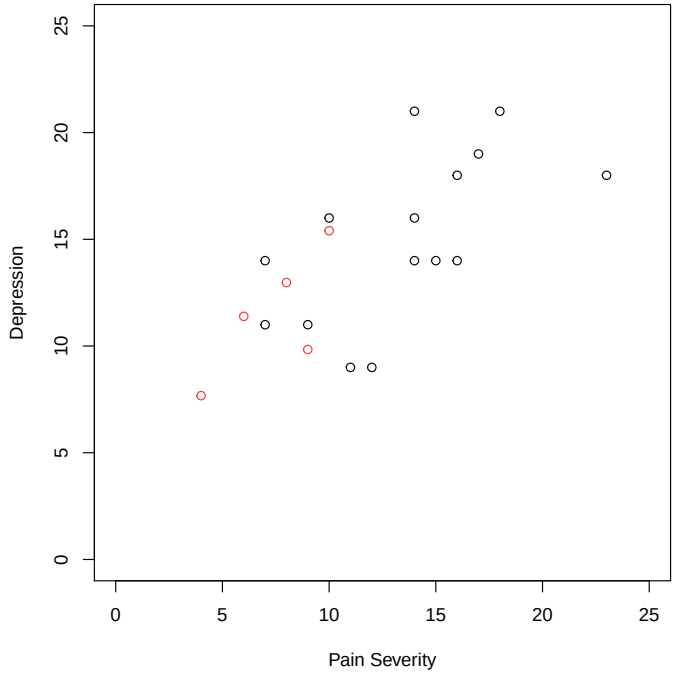
- However, the prediction equation has now a residual term, which is a random value from a normal distribution with mean zero and variance equal to the estimated variance of the residuals.

- The complete cases regression analysis produce a residual variance of $\widehat{\sigma}^2 = 3.211^2$.

- I then generated 5 values from a normal distribution with such variance and add to the predicted scores already obtained.

Single imputation

Stochastic regression imputation



- ↪ Above, it is shown a stochastic regression scatterplot of the pain and depression scores.
- ↪ Without any doubt, from the three methods described, it is the only one that produces a reasonable scatterplot.
- ↪ Note that in some cases/examples, stochastic regression imputation, by adding a random noise term, can lead to implausible predictions. For instance, in this specific example, it does not make sense to impute a negative value for the depression score.

Single imputation

Stochastic regression imputation

- ↪ At a first look, stochastic regression imputation may be a viable alternative to more sophisticated techniques.
- ↪ However, as any single imputation technique, stochastic regression attenuate standard errors of the estimates.
- ↪ Our intuition would dictate that a missing data analysis should produce larger standard errors than a hypothetical complete data analysis.
- ↪ However, standard analyses techniques treat the imputed values values as real data and the additional sampling error from the missing data is completely ignored.
- ↪ However, stochastic regression imputation has good features, it is an important step forward, and it actually form the basis of multiple imputation, a more sophisticated technique that we will learn later in this course.

Single imputation

Hot deck imputation

- **Hot deck imputation** is a procedure that has a long history in survey applications.
- Hot deck imputation is a collection of techniques that impute the missing values with values from 'similar' individuals. This is in contrast to 'cold deck' methods, where the imputations come from a previously collected data source.
- Several modifications have been proposed through the years.
- The basic idea is to impute missing values from other individuals.
- In its simplest version, a random draw from the observed data of the variable containing missing values replaces each missing value.

Single imputation

Hot deck imputation

- The more typical application replaces each missing value with a random draw from a subsample of individuals that have similar values on a set of matching variables (age, gender, race, marital status, etc).
- Note that the matching variables need not be categorical, since there are hot deck procedures that match individuals on continuous variables (e.g., nearest neighbour hot deck).
- Hot deck imputation generally does not attenuate the variability of the imputed data to the same extent as other imputation methods.
- However, hot-deck approaches can produce substantially biased estimates of correlations and regression coefficients (Schafer and Graham, 2002).
- Like any other single imputation procedure, hot deck type of procedures underestimate standard errors, although corrections have been proposed.

Single imputation

Last observation carried forward

- For longitudinal medical studies, the **last observation carried forward** approach imputes the missing value at one point in time with its value at the previous time point.
- Implicitly, this technique assumes that values do not change after the last observed measurement or during the intermittent period where scores are missing.
- This technique may produce biased estimates even when data are MCAR.
- Although widely used in medical studies and clinical trials, a growing number of empirical studies suggest that this method is a poor strategy to handle missing data.

Single imputation

↪ To finish, I quote Dempster and Rubin (1983):

> "*The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.*"