Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2018/2019

Single imputation mechanisms

Exercise

↪ The following table shows a small artificial dataset with 4 missing values:

Subject number	Sex	Age(years)	Systolic blood pressure (mmHg)
1	Male	50	163.5
2	Male	41	126.4
3	Male	52	150.7
4	Male	58	190.4
5	Male	56	172.2
6	Male	45	NA
7	Male	42	136.3
8	Male	48	146.8
9	Male	57	162.5
10	Male	56	161.0
11	Male	55	148.7
12	Male	58	163.6
13	Female	57	NA
14	Female	44	140.6
15	Female	56	NA
16	Female	45	118.7
17	Female	48	NA
18	Female	50	104.6
19	Female	59	131.5
20	Female	55	126.9

Single imputation mechanisms

Exercise

(i) Carry out a complete case analysis to find the mean value of systolic blood pressure overall, and by sex. Also compute the associated standard error of the mean.

(ii) Impute the missing values of systolic blood pressure by mean imputation. Use these filled in values to estimate the mean systolic blood pressure with corresponding standard error.

(iii) Impute the missing values for systolic blood pressure by regression imputation, regressing on sex and age. Write down the regression equation used and the four imputed values obtained. Use these to estimate the mean systolic blood pressure and associated standard error.

(iv) The same as in (iii) but now using stochastic regression imputation.

(v) Suppose that hot deck imputation is to be used with strata defined by sex and age (≤ 50 years and > 50 years). Estimate the mean systolic blood pressure with associated standard error.

Single imputation mechanisms

Exercise – solution

- ↪ Firstly, recall that the standard error of the mean is a measure of the precision of the mean.

- ↪ It depends on both the standard deviation and the sample size and it is defined as $\frac{\sigma}{\sqrt{n}}$.

- ↪ Since the standard deviation is seldom known in practice, we estimate it by the sample standard deviation, and thus the estimated standard error of the mean is given by $\frac{s}{\sqrt{n}}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

- ↪ Conducting a complete case analysis we obtain an overall estimated mean of 146.53, with associated standard error of 5.53.

- ↪ The mean SBP for men and for women are respectively 156.55 (5.27) and 124.46 (6.10). The values in brackets are the standard errors.

Single imputation mechanisms

Exercise – solution

- Using mean imputation, the estimated SBP mean is 146.53 (we already knew this from the complete case analysis), while the associated standard error is 4.39.
- Of course, the standard error of the mean based on mean imputation is smaller than that of the complete case analysis. Why? Have a look at the sample variance formula.
- We will now use regression imputation, conditioning on sex and age, to impute the missing SBP values and consequently compute its mean and associated standard error.
- Let us use the following regression model

$$\text{SBP} = \beta_0 + \beta_1\text{Age} + \beta_2\text{Sex} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2).$$

- Here, gender is a dummy (binary) variable taking the value 0 if the subject is female and 1 if the subject is male.

Single imputation mechanisms

Exercise – solution

$\hookrightarrow$ The expected values for SBP are then

$$E(SBP) = \begin{cases} \beta_0 + \beta_1 \text{Age} & \text{if the subject is female,} \\ (\beta_0 + \beta_2) + \beta_1 \text{Age} & \text{if the subject is male.} \end{cases}$$

$\hookrightarrow$ We fit the regression model to the complete cases, obtain estimated regression coefficients and then impute the SBP values based on the regression equation.

$\hookrightarrow$ The estimated regression coefficients are $\widehat{\beta}_0 = 40.8784$, $\widehat{\beta}_1 = 1.6518$, and $\widehat{\beta}_2 = 29.6318$.

$\hookrightarrow$ The imputed values are then

Subject number	Sex	Age	Imputed SBP
6	Male	45	144.84
13	female	57	135.03
15	female	56	133.38
17	Female	48	120.17

Single imputation mechanisms

Exercise – solution

- For the above regression imputation, the estimated mean SBP is 143.89 with a standard error of 4.65.
- We will impute the SBP values based on stochastic regression imputation.
- The steps are essentially the same as those in regression imputation, but we add a random term, normally distributed with mean zero and variance equal to the estimated variance of the residuals, which in this case is 13.34^2.
- The imputed values in this case are

Subject number	Sex	Age	Imputed SBP
6	Male	45	133.90
13	female	57	126.75
15	female	56	148.38
17	Female	48	119.95

- Of course, since we are adding a random term, different execution of the code will lead to different imputed values. The values obtained above were obtained using `set.seed(1)`.

Single imputation mechanisms

Exercise – solution

- $\hookrightarrow$ The estimated mean is thus 143.67 (4.71).
- $\hookrightarrow$ We now turn our attention to hot deck imputation. We define strata based on gender and age (≤ 50 and > 50).
- $\hookrightarrow$ For instance, we see that subject number 6 is male and is 45 years old. Thus, possible candidates to give him his SBP value are subjects number 1, 2, 7, and 8. We now just need to pick one of these subjects in a random fashion. Using the seed equal to 1, we have obtained subject number 8 to be the donor.
- $\hookrightarrow$ We proceed in the same fashion for the other subjects.

Single imputation mechanisms

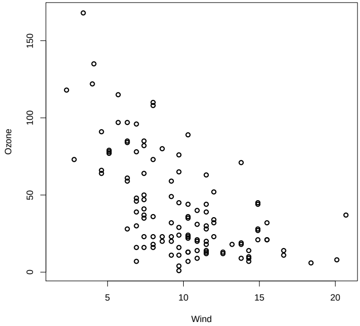
Cautionary note

- The imputed values from regression imputation and stochastic regression imputation are only as good as the model, and so assumptions must be checked.
- So far, we have used normal linear regression models, which assume, normality of the error terms, homoscedasticity (i.e., the variance of the error terms is constant, or equivalently, does not depend on the regressors), independence of the error terms, and linearity of the predictor.
- All these assumptions can be checked informally using diagnostic plots (c.f. your favourite regression book).
- We give a specific example concerning the linearity of the predictor. Let us use the `airquality` dataset again.

Single imputation mechanisms

Cautionary note

- Below we have a scatterplot of the variables wind and ozone (which has missing values, but this is not the point here).

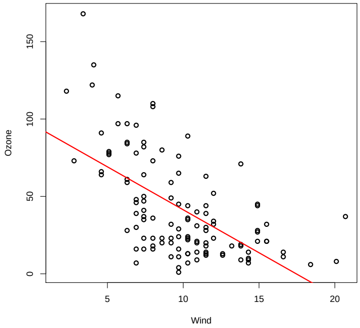


- The relationship does not seem to be linear, as the plot in the next slide indicates.

Single imputation mechanisms

Cautionary note

↪ Scatterplot with the fitted regression line superimposed.



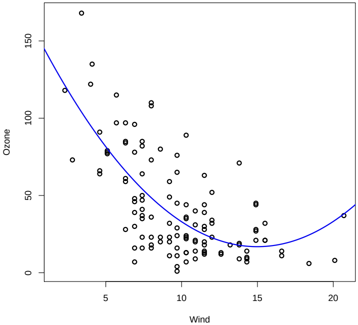
↪ The nonlinear curvature might suggest the inclusion of a quadratic term.

Single imputation mechanisms
Cautionary note

↪ We have then fitted the model

$$\text{Ozone} = \beta_0 + \beta_1 \text{Wind} + \beta_2 \text{Wind}^2 + \epsilon, \qquad \epsilon \sim \text{N}(0, \sigma^2).$$

↪ Below, we show the scatterplot along with the quadratic fit, which seems much better.



Single imputation mechanisms

Illustrative simulation study

- ↪ To better illustrate the performance of some of the traditional methods to handle missing data I have conducted a Monte Carlo simulation study.
- ↪ A Monte Carlo simulation study generates a large number of samples (e.g., 1000) from a population with a specified set of parameter values.
- ↪ Estimating a statistical model on each sample and saving the resulting parameter estimates creates an empirical sampling distribution for each model parameter.
- ↪ The difference between the average parameter estimate and the true population parameter is of particular importance because it quantifies the bias.
- ↪ I've uploaded on Learn three interesting papers (in my opinion!) about simulation studies.

Single imputation mechanisms

Illustrative simulation study

- I have generated 1000 datasets, each consisting of 200 observations, from a bivariate normal distribution with the following structure

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \qquad \rho = 0.5.$$

- We then write

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- We will impose missingness on Y_2 only, in a similar fashion as we did in Exercise 3 of Workshop 1.

Single imputation mechanisms

Illustrative simulation study

$\hookrightarrow$ The first simulation created MCAR data by imposing

$$\Pr(R_2 = 1 \mid Y_1, Y_2) = 0.5.$$

$\hookrightarrow$ The second simulation created MAR data by imposing

$$\Pr(R_2 = 1 \mid Y_1, Y_2) = \frac{e^{\beta_0 + \beta_1 Y_1}}{1 + e^{\beta_0 + \beta_1 Y_1}}, \quad \beta_0 = 1.5, \quad \beta_1 = 3.$$

$\hookrightarrow$ Finally, the third simulation study created MNAR data by imposing

$$\Pr(R_2 = 1 \mid Y_1, Y_2) = \frac{e^{\beta_0 + \beta_1 Y_1 + \beta_2 Y_2}}{1 + e^{\beta_0 + \beta_1 Y_1 + \beta_2 Y_2}}, \quad \beta_0 = 1.5, \quad \beta_1 = 3, \quad \beta_2 = 5.$$

Single imputation mechanisms

Illustrative simulation study – MCAR results

Estimate	Population parameter	CC	MI	RI	SRI
μ_1	0	-0.0056	-0.0029	-0.0029	-0.0029
μ_2	0	0.0007	0.0007	0.0021	-0.0005
σ_1^2	1	0.9992	1.0006	1.0006	1.0006
σ_2^2	1	1.0016	0.499	0.6270	1.0036
ρ	0.5	0.4962	0.3499	0.6279	0.4971

Single imputation mechanisms

Illustrative simulation study – MAR results

Estimate	Population parameter	CC	MI	RI	SRI
μ_1	0	0.4671	-0.0029	-0.0029	-0.0029
μ_2	0	0.2352	0.2352	0.0007	0.0024
σ_1^2	1	0.6008	1.0006	1.0006	1.0006
σ_2^2	1	0.9000	0.5984	0.7539	1.0042
ρ	0.5	0.4060	0.2569	0.5715	0.4935

Single imputation mechanisms

Illustrative simulation study – MNAR results

Estimate	Population parameter	CC	MI	RI	SRI
μ_1	0	0.5064	-0.0029	-0.0029	-0.0029
μ_2	0	0.6030	0.6030	0.5286	0.5289
σ_1^2	1	0.6592	1.0006	1.0006	1.0006
σ_2^2	1	0.5212	0.3025	0.3204	0.5337
ρ	0.5	0.1626	0.1006	0.2531	0.1963

Single imputation mechanisms

Imputation of several missing variables – routine multivariate imputation

- When presenting the conditional mean imputation and stochastic regression imputation methods, we assumed that only one variable was subject to missingness.
- We now put ourselves in the situation where several variables have missing values.
- One possibility, directly extending what we have seen before, is to specify a multivariate regression model. However, the main drawback of this approach is that it is not trivial to set up a reasonable multivariate regression model.
- Usually, a multivariate normal or t distribution is used for continuous variables, and a multinomial distribution for discrete outcomes.
- Positive point is that software exist to fit such models.

Single imputation mechanisms

Imputation of several missing variables – iterative regression imputation

- A different way to extend the univariate methods seen before is to apply them iteratively to the variables with missingness.
- Supposing the variables with missingness are $Y_{(1)}, Y_{(2)}, \ldots, Y_{(k)}$. Suppose further that Y_{obs} denotes the fully observed variables (i.e., no missingness).
- Iterative regression imputation consists of the following steps:
  1. Imputing all missing values in $Y_{(1)}, Y_{(2)}, \ldots, Y_{(k)}$ using some crude imputation method (e.g., mean imputation).
  2. Impute $Y_{(1)}$ given $Y_{(2)}, \ldots, Y_{(k)}$ and Y_{obs} (using for instance stochastic regression imputation).
  3. Impute $Y_{(2)}$ given $Y_{(1)}, Y_{(3)} \ldots, Y_{(k)}$ and Y_{obs} (using the newly imputed values for $Y_{(1)}$).
  4. ...
  5. Impute $Y_{(k)}$ given $Y_{(1)}, Y_{(2)} \ldots, Y_{(k-1)}$ and Y_{obs} (using the newly imputed values for $Y_{(1)}, Y_{(2)} \ldots, Y_{(k-1)}$).
- Loop until approximate convergence.

Single imputation mechanisms

Imputation of several missing variables – iterative regression imputation

- $\hookrightarrow$ Iterative regression imputation has the advantage that, when compared to the full multivariate model, the set of regression models (one for each variable with missingness) is easier to understand, thus allowing the user to fit a reasonable model at each step.
- $\hookrightarrow$ The disadvantage is that the user needs to be more careful in this setting in order to ensure that the separate regression models are consistent with each other.
- $\hookrightarrow$ For instance, on a survey, it would not make sense to impute age based on income but then later ignoring age when imputing income.

Single imputation mechanisms

Summary

- $\hookrightarrow$ Best approach is possibly stochastic regression imputation.
- $\hookrightarrow$ Can be reasonable with only a mild percentage of missing values (e.g., $< 5\%$).
- $\hookrightarrow$ We should be aware that all methods, although to different extents, results in overly precise estimates (i.e., standard errors too small):
  - $\hookrightarrow$ Analyses after single imputation do not know that some of the values have been imputed.
  - $\hookrightarrow$ Simply treats imputed values as if they were observed.
  - $\hookrightarrow$ Does not take into account the uncertainty in the imputed values.