Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2018/2019

Naive methods

- ↪ Here we will review some simple methods for handle missing data. It is worth keeping in mind that such methods only work under very specific and (somehow) restrictive conditions.
- ↪ Nevertheless, these methods have been widely used in practice and so they deserve our attention.
- ↪ These methods deal with missing data by removing the cases/observations with incomplete data or by filling in the missing values (one single value is used–single imputation).
- ↪ Some of these methods are:
  1. Complete case analysis.
  2. Available case analysis.
  3. Unconditional mean imputation.
  4. Conditional or regression imputation.
  5. Stochastic regression imputation.
  6. Hot deck imputation.
- ↪ These methods have the potential of inducing bias as well as understating variability and are generally not recommended (unless in very specific situations).

Naive methods

Bias of an estimator – definition

- $\hookrightarrow$ We have just mentioned that the procedures to be described have the potential to induce bias. It is then worth defining bias before proceeding.
- $\hookrightarrow$ Suppose that $X_1, \ldots, X_n$ are iid random variables, each with pdf/pmf $f_X(x \mid \theta)$, θ unknown.
- $\hookrightarrow$ If $\widehat{\theta} = T(\mathbf{X})$ is an estimator of θ, then the bias of $\widehat{\theta}$ is the difference between its expectation and the 'true' value, i.e.,

$$\text{bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

- $\hookrightarrow$ An estimator $T(\mathbf{X})$ is unbiased if

$$E\{T(\mathbf{X})\} = \theta \quad \text{for all } \theta,$$

otherwise it is biased.

Naive methods

Bias of an estimator – example

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, 1)$.
- A possible estimator for μ is

$$T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- We have

$$\begin{aligned} E\{T(\mathbf{X})\} &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \mu. \end{aligned}$$

- Thus, T is unbiased for μ.

Naive methods

Methods that discard data

- ↪ Many missing data approaches simplify the problem by discarding data.
- ↪ The primary advantage of these approaches is that they are convenient to implement and the default option is statistical software packages.
- ↪ These approaches assume MCAR and can lead to biased estimates when this assumption does not hold.
- ↪ In addition, even if MCAR assumption is plausible, eliminating data is wasteful and can lead to estimates with larger standard errors due to reduced sample size.

Naive methods

Methods that discard data

- ↪ Consequently, there is little to recommend these techniques unless the proportion of missing data is trivially small.
- ↪ If, say, only 1% of cases have missing values, then deletion would certainly be the quickest way to deal with missing data.
- ↪ Graham's (2009) recommendation is that if 5% or more of cases are missing, one should then do something else than deletion.

Naive methods

Methods that discard data – complete case analysis

- ↪ Complete case analysis (also known as listwise deletion) excludes the data for any case that has one or more missing values.

- ↪ The data are treated as if the cases with missing values were not there.

- ↪ Restricting the analyses to complete cases eliminates the need for specialised software and for advanced missing data handling procedures.

- ↪ This method is, by far, the most commonly used method in the applied (medical, social, etc) sciences.

Naive methods

Methods that discard data – complete case analysis

- ↪ But what are the costs associated to such simplicity?
- ↪ If the data are MCAR, then the observed data are a random subsample of the whole sample.
- ↪ Hence, any estimates derived from the restricted subsample will be unbiased estimates of corresponding quantities in the sample.
- ↪ Assuming that the sample is, itself, a random sample of the overall population of interest, the results will be unbiased in the general sense and complete case analysis is a valid approach.
- ↪ However, the sampling variability/standard error will be larger than in the case of no missing data, so that confidence intervals will be wider and power reduced, compared with the no missing situation.

Naive methods

Methods that discard data – complete case analysis

- ↪ Even if the data are only MAR, inferences from a complete case analysis can be biased.

- ↪ If many variables are included in a model, there may be very few complete cases, so that most of the data would be discarded for the sake of a simple analysis.

- ↪ As an interesting aside note, it is worth mentioning the result of Little (1992): complete case analysis can produce unbiased estimates of regression slopes under any missing data mechanism, provided that missingness is a function of a predictor variable and not of the outcome variables.

- ↪ We should keep in mind that this aforementioned very particular scenario is possibly the only situation in which complete case analysis is likely to outperform more advanced techniques (maximum likelihood and multiple imputation) for MNAR data.

Naive methods

Methods that discard data – available case analysis

- ↪ Available case analysis, also known as pairwise deletion, attempts to mitigate the data loss problem of complete case analysis.

- ↪ In available case analysis different aspects of a problem are studied with different subsets of the data.

- ↪ The typical application of available case analysis is the computation of covariance/correlation matrices, where different subsets of cases are used to compute each element in the covariance/correlation matrix.

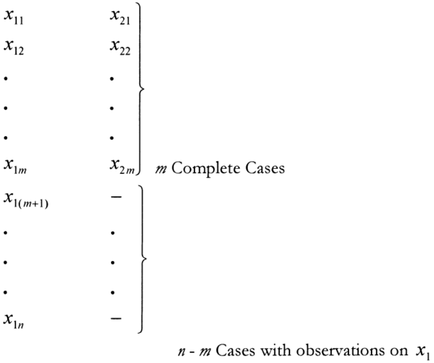
- ↪ Consider as an example, a simple two variable (X_1, X_2) data matrix with only one variable, X_2, subject to missingness.

- ↪ In available case analysis, all cases would be used to estimate the mean and variance of X_1, but only the complete cases would contribute to an estimate of X_2 and the covariance between X_1 and X_2.

Naive methods

Methods that discard data – available case analysis

- ↪ Suppose that data are reordered and so the first m cases have complete data on both variables and the remaining $n - m$ cases have missing values on X_2.



$$
\begin{array}{ll}
x_{11} & x_{21} \\
x_{12} & x_{22} \\
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot \\
x_{1m} & x_{2m}
\end{array}
\left.\vphantom{\begin{array}{l} x \\ x \\ \cdot \\ \cdot \\ \cdot \\ x \end{array}}\right\} m \text{ Complete Cases}
$$

$$
\begin{array}{ll}
x_{1(m+1)} & - \\
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot \\
x_{1n} & -
\end{array}
\left.\vphantom{\begin{array}{l} x \\ \cdot \\ \cdot \\ \cdot \\ x \end{array}}\right\}
$$

n - m Cases with observations on x_1

Figure from Pigott (2001).

Naive methods

Methods that discard data – available case analysis

↪ The available case estimates are

$$\begin{aligned}
\bar{x}_1 &= \frac{1}{n}\sum_{i=1}^{n} x_{1i},\\
\bar{x}_2 &= \frac{1}{m}\sum_{i=1}^{m} x_{2i},\\
s_1^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2,\\
s_2^2 &= \frac{1}{m-1}\sum_{i=1}^{m}(x_{2i}-\bar{x}_2)^2,\\
s_{1,2} &= \frac{1}{m-1}\sum_{i=1}^{m}(x_{1i}-\bar{x}_{1(m)})(x_{2i}-\bar{x}_2),
\end{aligned}$$

where $\bar{x}_{1(m)}$ is the mean calculated from the m cases.

Naive methods

Methods that discard data – available case analysis

↪ The covariance matrix is then given by

$$\begin{bmatrix} s_1^2 & s_{1,2} \\ s_{1,2} & s_2^2 \end{bmatrix}$$

↪ Similarly, the correlation between X_1 and X_2, under available case analysis, would be

$$\rho_{1,2} = \frac{s_{1,2}}{s_{1(m)} s_2}.$$

↪ There are software packages that compute the variances s_1^2 and s_2^2 in the correlation from separate subsamples (e.g., compute it using $s^2_{1(n)}$ instead of using $s^2_{1(m)}$). This approach is problematic because it may produce implausible correlations, i.e., correlations that exceed plus or minus one.

Naive methods

Methods that discard data – available case analysis

- ↪ Another problem is that it is not clear which sample size should be used to compute standard errors (since no single value of n is applicable to the entire covariance matrix).
- ↪ Some software packages use the average sample size per variable but this approach is likely to underestimate the standard error for some variables and overestimate the standard errors for other variables (Little, 1991).
- ↪ This method is simple, uses all available information and produces consistent estimates of mean, correlations and covariances under MCAR (Little and Rubin, 2002, p.55).
- ↪ However, the estimates can be biased if the data are not MCAR.

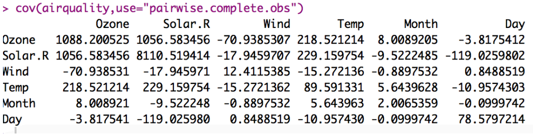
Naive methods

Methods that discard data – available case analysis

- In R, available cases analysis for covariances and correlations matrices can be easily computed by using the extra argument use=”pairwise.complete.obs”.

- For illustration purposes, consider the airquality dataset. We already know that the variables Ozone and Solar.R have missing values.

- To obtain the covariance matrix we simply do



```
> cov(airquality,use="pairwise.complete.obs")
              Ozone     Solar.R        Wind       Temp      Month          Day
Ozone   1088.200525 1056.583456 -70.9385307 218.521214  8.0089205   -3.8175412
Solar.R 1056.583456 8110.519414 -17.9459707 229.159754 -9.5222485 -119.0259802
Wind     -70.938531  -17.945971  12.4115385 -15.272136 -0.8897532    0.8488519
Temp     218.521214  229.159754 -15.2721362  89.591331  5.6439628  -10.9574303
Month      8.008921   -9.522248  -0.8897532   5.643963  2.0065359   -0.0999742
Day       -3.817541 -119.025980   0.8488519 -10.957430 -0.0999742   78.5797214
```

- If we want the correlation matrix, simply type cor instead of cov.

Single imputation mechanisms

- ↪ Rather than removing variables or observations 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.

- ↪ 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.

- ↪ The methods will be presented briefly and given their drawbacks, single imputation techniques are, in general, not recommended.

Single imputation mechanisms

- We shall use a small bivariate data to illustrate ideas.
- 20 chronic patients enrolled in a pain management program.
- 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 mechanisms

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.

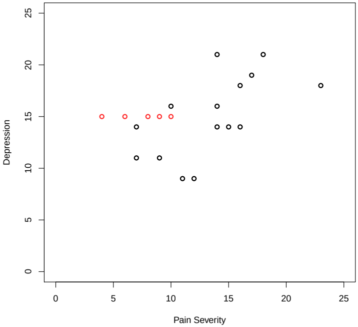
Single imputation mechanisms

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.
- ↪ Imputing the mean of a variable also attenuates the magnitudes of covariances and correlations.
- ↪ Remember the covariance formula. Cases with missing values on either one of the variables contribute with a value of zero to the numerator formula. Obviously, the same is then true for the correlation.

Single imputation mechanisms

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.
- ↪ Several studies suggest that mean imputation is possibly the worst missing data handling method available.

Single imputation mechanisms

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.

- 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 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 mechanisms

Conditional mean imputation

- ↪ Reconsider the bivariate data on pain/depression scores.
- ↪ The 15 complete cases were used to estimate the regression of depression score on pain score.
- ↪ The resulting regression equation is

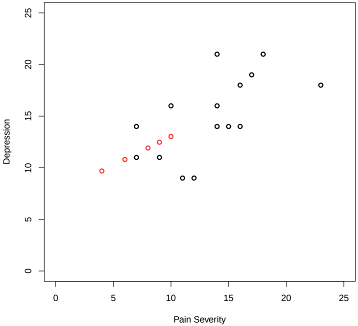
$$\begin{aligned}\widehat{\text{DEP}}_i &= \widehat{\beta_0} + \widehat{\beta_1}\text{PS}_i \\ &= 7.4568 + 0.5574\text{PS}_i.\end{aligned}$$

- ↪ 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 mechanisms

Conditional mean imputation



- ↪ The figure above shows a scatterplot of the imputed data. 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 mechanisms

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.

- ↪ 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.

- ↪ As a result, variances and covariances are attenuated, although not to the same extent as in mean imputation.

Single imputation mechanisms

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 under a MAR missing data mechanism.

Single imputation mechanisms

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, i.e., $\widehat{\beta}_0 = 7.4568$ and $\widehat{\beta}_1 = 0.5574$.

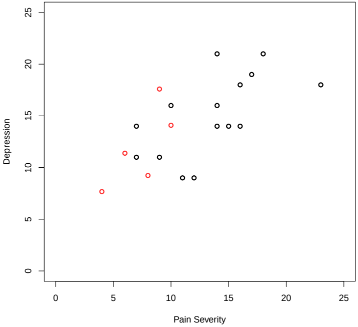
- However, the prediction equation has now a residual term, which is a random value from a normal distribution with mean zero and variance equal to the estimated variance of the residuals.

- 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 mechanisms

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.

Single imputation mechanisms

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.
- ↪ 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 and it actually form the basis of multiple imputation, a more sophisticated technique that we will learn later in this course.

Single imputation mechanisms

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 replaces each missing value.

Single imputation mechanisms

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 mechanisms

Last observation carried forward

- ↪ For longitudinal medical studies, this approach imputes the missing value at one point in time with its value at the previous time point, or used the last observation carried forward.

- ↪ 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 mechanisms

Question from last’s year exam

↪ Consider a two variable (Y_1, Y_2) data matrix with only Y_2 subject to missingness. Conducting a complete case analysis, one has obtained that the estimated mean of Y_2 is 35. For this same dataset, if we were to conduct a mean imputation approach instead of a complete case analysis, what would be the estimated mean of Y_2 after imputation? Justify.

Single imputation mechanisms

Another question from last's year exam

↪ Consider the following bivariate dataset composed of five observations on each variable. While Y_1 is completely observed, Y_2 has two missing values.

Y_1	Y_2
-0.1	-1.0
-0.1	0.4
-1.1	**NA**
1.1	1.7
0.4	**NA**

When answering the next two questions, use the `R` output provided at the end of this question.

(i) Explicitly write down the regression imputation equation (also known as conditional mean imputation equation) and use it to impute values for the two observations missing.

(ii) If you were to conduct a stochastic regression imputation, what would be the exact regression equation you would use to impute the missing values.

Single imputation mechanisms

Another question from last’s year exam (continued)

```
Call:
lm(formula = y2[r2 == 1] ~ y1[r2 == 1])

Residuals:
       1         2         3
-7.00e-01  7.00e-01 -1.11e-16

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1333     0.6469  -0.206    0.871
y1[r2 == 1]   1.6667     1.0104   1.650    0.347

Residual standard error: 0.9899 on 1 degrees of freedom
Multiple R-squared:  0.7313,Adjusted R-squared:  0.4625
F-statistic: 2.721 on 1 and 1 DF,  p-value: 0.3469
```