Data Imputation

Complete-Case Analysis: Where all cases where information is missing are excluded

Can think of imputation as supplying distributions for the input variables/predictors

Why data is missing

1. *Missingness Completely At Random* (MCAR): A piece of data is missing completely at random if the probability of missingness is the same for all units, for example if every respondent decided to answer a question by rolling a dice than the data is missing completely at random.
2. *Missingness at Random* (MAR): The probability that a variable is missing depends only on the available information. It is missing completely at random for each category of another variable. i.e in the social survey data carried out by the new york city council, earnings is not completely missing at random as can be seen when looking at the different response rates of the race categories 'white' and 'black'. For objects where the variable race is on the class 'black' than the data for earnings can be said to be missing completely at random.
3. *Missingness that depends on unobserved predictors*: The missingness depends on information that has not been recorded and also predicts the missing values
4. *Missingness that depends on the missing value itself*: The probability of missingness depends on the (potentially missing) variable itself. for example higher earners may be more unwilling to reveal their earnings and so the variable itself is the cause of the missingness.

Discarding Data

Discarding objects with missing data can lead to:

1. biased estimates
2. Estimates with large standard errors

The methods of discarding data are:

1. Complete Case Analysis
2. Available-case analysis
3. Complete-Variable analysis
4. Nonresponse weighting

Complete Case Analysis

A direct approach to missing data is to exclude them

- In the regression context, this is usually complete-case analysis, that means excluding all units for which the outcome or any of the inputs are missing

Issues:

- What if the units that are missing differ systematically from the observed cases? This could bias the complete-case analysis
- If many variables are included in a model, there may be very few complete cases and a large portion of the data is discarded for the sake of simple analysis

Available-case analysis

To analyse each variable based on the available data for that particular variable.

Has the issue that the analysis for different variables will be based off of different subsets of the data.

Same bias problems as complete case analysis

Complete Variable Analysis

Excluding a variable or set of variables from the analysis die to their missing data.

This can lead to the omission of a variable that is necessary to satisfy the assumptions required for prediction.

Non-response weighting

Easy to use when only one - mainly categorical - variable has missing data

- Build a model using all other variables to predict the non-response in that variable
- The Inverse of predicted probabilities of response from this model could then be used as survey weights to make the complete-case sample representative (along the dimensions measured by the other predictors) of the full sample

Give a higher weight to those cases that are less likely to respond so as to balance out the bias in the sample. The logic here is that those who are more liekely to respond will be over represented and those who are less likely to respond will be under-represented. By weighting the less likely respondants higher than the likely respondants you balance the sample to more accurately represent the population.

Imputation

1. Mean Imputation
2. Last Value Carried Forward
3. Using information from related studies

4. Indicator Variables for missingness of categorical predictors
5. Indicator Variables for missingness of continuous predictors
6. Imputation based on logical rules
7. Simple random imputation
8. Using regression predictions to perform deterministic imputation
9. Using regression predictions to perform random imputation
10. Two-stage modeling to impute a variable that can be positive or zero
11. Matching and hot-deck imputation
12. Routine multivariate imputation
13. Iterative regression imputation

Mean Imputation

Easiest method of imputation.

- can severely distort the distribution of the variable, leading to complication with summary measures
- underestimates of the standard deviation. Pulling the distribution towards the mean.
- distorts relationships between variables by pulling estimates of the correlation towards zero. The mean has nothing to do with the variables that would have influenced the true value of a missing piece of data and so this is pulling the correlation towards zero i.e no correlation

Last Value Carried Forward

In the case of a linear time series of data, the approach of carrying the last observed value forward to impute missing values can be used.

Or in another case if the data had some prior estimate of the value, called a pre-treatment value. You could carry that value forward to impute the missing data.

Issues with this approach is that it can be conservative as it would underestimate the true treatment effect.

It can also be anti-conservative in some cases, for example control surveys where there probably would have been a marked decrease in some form of risky behavior but instead it remains the same, making things look worse than they are.

Using information from related studies

Used in cases where data of a related variable could be a plausable substitute for the missing data. For example replacing missing data about a fathers income with the income of the mother.

Can propagate measurement error.

Is there incentive for the reporting individual to missrepresent the measurement

Indicator variables for missingness of categorical predictors

A simple means of imputation is to add an extra category for the variable indicating missingness

Indicator variables for missingness of continuous predictors

- Add an extra variable indicating which observations have missing data
- Then replace the missing values in the partially observed predictor by zeroes of the mean
- Prone to yield biased coefficient estimates for the other predictors included in the model.

Imputation based on logical rules

- Making logical assumptions of the value based off of related variables. For example is the income is not reported but the hours worked are zero then the income could be assumed as zero.

Random Imputation of a single variable

When more than a trivial fraction of the data is missing then we prefer to perform more formal imputation

Simple random imputation

Replacing missing values of a variable based on the observed values of the variable itself. Basically randomly choose from the observed values to replace the missing values.

Zero-Coding and Top-Coding

- Regression model will be fit to those respondents whose earnings were observed and positive.
- Top-Code all observations above a certain value. i.e replace the value above a maximum value with the maximum permitted value. Reducing the sensitivity to the highest values.
- Lose information about higher values but has no effect on categorization.

This can improve the predictive power of the regression models.

Using regression predictions to perform deterministic imputation

Use simple linear regression to impute the missing values (the *lm()* function in R). The dependent variable is the variable to impute and the independent variables are the other columns in the dataset that could in some way influence the dependent variable.

Can be poor at actually matching the variance and deviation in the data. Leading to false implications

Using regression predictions to perform random imputation

Putting uncertainty back into the imputations by adding prediction error into the regression.

Could create random predicted values for the missing cases using the normal distribution.

Two-stage modeling to impute a variable that can be positive or zero

Can be required when some information about a variable is not available to determine weather the variable is positive or zero

1. Imputing an indicator of if the variable is positive
2. then imputing the variable itself based off of the positive indication.

For the first part perform a logistic regression for I^y the indicator that y is positive and is 1 if $y > 0$

For the second part perform a linear regression for the square root of y^pos, which is y if $y > 0$.

Matching and hot-deck imputation

- for each unit with a missing value, y, find a unit with a similar value of X in the observed data and take its y value.

Routine multivariate imputation

Setting up a multivariate model to all the variables that have missingness

The issue of this approach is that it requires a lot of effort to set up a reasonable multivariate regression model

Iterative regression imputation

Iteratively imputing the missing values in variables. So if the variables with missingness are Y and the complete variables are X then you impute Y(1) given y(2)... Y(k) and X then Y(2) given Y(1) and Y(3)... Y(k) and X and so on (using the newly imputed values of Y(1)).

Acronyms