

Outliers and Missing data (Part 2, Missing Data)

Daniel Lawson University of Bristol

Lecture 05.2.2 (v1.0.1)

Signposting

- ▶ Missing Data is an essential topic in **Data Cleaning**.
 - ▶ It is about reasoning about what your data are, rather than what you assume them to be.
 - ▶ It is how you might detect problems.
- ▶ It relates especially to Block 1's EDA lecture, but will be practically essential for every project.
- ▶ This is part 1 of Lecture 4.2:
 - ▶ Part 1 is about outliers,
 - ▶ Part 2 is about missing data.

Intended Learning Outcomes

- ▶ ILO1 Be able to **access and process cyber security data** into a format suitable for mathematical reasoning
- ▶ ILO2 Be able to **use and apply basic machine learning** tools
- ▶ ILO3 Be able to make and report appropriate inferences from the results of applying basic tools to data

Types of missing data

1. **Missing completely at random.**

- ▶ The missing data are completely independent of everything, and can be modelled independently.
- ▶ This sort of missingness is often called **ignorable**.

2. **Missing at random.**

- ▶ The missing data are dependent on observed variables, and can therefore missing status be **modelled independently of the values**.
- ▶ For example, data might be more missing if it is UDP than TCP.

3. **Missing dependent on unobserved parameters.**

- ▶ The missing data are also dependent of latent properties of the model, and therefore must be modelled **at the same time as the values**.
- ▶ For example, data might be more missing if it is from a particular cluster.

4. **Missing dependent on the missing value.**

- ▶ Whether something is missing depends of the value it takes.
- ▶ This is called **censoring** and is a modelling category of its own.

Missingness

- ▶ When inferring missingness, it is possible to prove that **it is impossible to detect the type of missingness**.
- ▶ This is because more complex forms of missingness can always be constructed. . .
 - ▶ That appear, given the data available, to be from a simpler class of missingness.
- ▶ It is therefore always an assumption that you have handled missingness “correctly”.

Methods that discard data

- ▶ Discarding data that contain missingness is a common strategy.
 - ▶ It can **cause biased inference**, when data are not missing completely at random.
 - ▶ It also reduces power.
- ▶ We make two main distinctions for how to remove records:
 - ▶ **Complete case analysis**: keep all cases that contain no missing data.
 - ▶ **Available case analysis**: keep all cases that are complete for each question independently. Therefore different analyses may be differently biased.
- ▶ Discarding **variables** (features) due to missingness rate has a similar flavour to available case analysis.
 - ▶ Philosophically it can be concerning. “What if I had never measured this feature?” leads to “What if there is some feature I need that I have not measured?”
 - ▶ Similar questions arise in sampling. “What if there is an important population that I did not sample?”

Example of available case analysis

```
> summary(lm1)
...
(188178 observations deleted due to missingness)
```

Methods that impute missing data

- ▶ There is only one statistically defensible way to do imputation:
 1. Design a **model that you believe** could be true
 2. Treat **missing data as parameters** of that model
 3. **Test the assumptions** behind the model
 4. **Repeat** until the model assumptions cannot be falsified
- ▶ In practice this is rarely possible.

Imputation procedures

- ▶ In order of decreasing difficulty of implementation:
- ▶ **Bayesian model-based inference.**
 - ▶ You may not believe your model, but it is still your best model. Use it for your inference goal.
- ▶ **Monte-carlo model-based imputation.**
 - ▶ Use your best model, and make multiple random datasets which you then insert into your inference framework.
- ▶ **Model-based imputation.**
 - ▶ Use your best model, insert the best-guess for all the missing data.
- ▶ **Regular imputation.**
 - ▶ Use a fast model to impute rapidly.
- ▶ Throughout, many biases can be reduced by **retaining and using indicators of missingness status.**

Imputation approaches for large datasets

- ▶ Assuming that you can't just run a plausible model, approaches include:
- ▶ **Mean imputation.**
 - ▶ Replace missing values by the mean.
 - ▶ This tends to create many distortions but is often OK when detecting outliers though an appropriate method, e.g. PCA.
- ▶ **Regression** or other predictive models.
 - ▶ Try to mean-predict the missing values based on what else is present.
- ▶ **Nearest neighbour prediction.**
 - ▶ Using the mean value of the nearest k-neighbours can work surprisingly well for some problems, though may propagate measurement error.
- ▶ **Conservative replacement.**
 - ▶ If directions of effects are known apriori, it is sometimes possible to construct a conservative estimate.
 - ▶ This requires care and understanding what the variables mean.

Example: Mean imputation

```
## Mean imputation
```

```
conndataMmean=conndataM
```

```
conndataMmean[is.na(conndataMmean[, "logduration"]),  
               "logduration"] =
```

```
  mean(na.omit(conndataMmean[, "logduration"]))
```

```
tdata=conndataMmean[conndataMmean[, "logduration"] < log(threshold)
```

```
lm2=lm(logduration~proto+service +ts +id.resp_p+day,  
        data=tdata)
```

Activity 4 of the workshop.

Regression based imputation

- ▶ This is a direct extension of mean-imputation.
- ▶ We build a model for the covariate,
 - ▶ Regression is popular,
 - ▶ (though ideally, the model would be robust to missing data itself. . .)
- ▶ And replace the values with the predictions.
- ▶ This is conceptually still mean imputation, but where covariates matter.

Nearest neighbour imputation

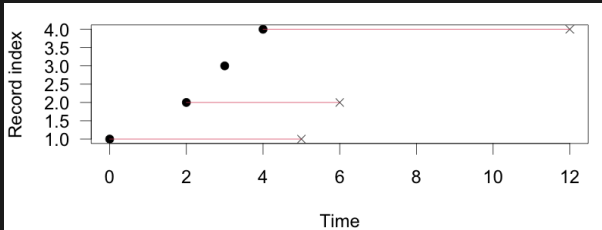
- ▶ Define the set of neighbours for each **record** according to a **distance measure**.
- ▶ Form a graph with records as nodes in a graph.
- ▶ Missing data on a node is **imputed** as the mean/median/etc of its **neighbours**.
- ▶ Local graph computations are efficient.

Activity 6 of the Workshop.

Conservative Imputation

- ▶ Imputation that is conservative **relative** to some task.
- ▶ Usually involves a statistical test. . .
- ▶ In which you can **guarantee** that the test statistic is going to **monotonically decrease** under application of the imputation (assuming that big values are evidence against the null).
- ▶ If you can do conservative imputation, and false positives are your target whilst false negatives are not of concern, then conservative imputation is to be recommended.

Example: Conservative testing with censoring



- ▶ Is a record B “nested” inside record A?
 - ▶ Make “segments” out of each record, i.e. a start and end time.
 - ▶ For missing B events, we can impute conservative end times by setting duration to 0.
 - ▶ For missing A events, this is not possible.
- ▶ **Activity 5** of the Workshop.

Many missing covariates

- ▶ When multiple covariates are missing, there is no “trivial” imputation.
- ▶ The previous methods can be used with a **iterative scheme**, where an imputation method is used for each in turn.
- ▶ **Model-based methods** such as Bayesian models handle missing values as parameters.
 - ▶ This can be efficient if missingness is sparse.
- ▶ In general, if missingness is dense, there may be multiple possible solution modes.
 - ▶ Finding these, and expressing uncertainty, is often a challenge.

Testing imputation procedures

- ▶ You should always **test everything**.
- ▶ In missing data problems, this means:
 - ▶ Taking data that is not missing,
 - ▶ Making it missing according to your **best beliefs** (NOT your model!)
 - ▶ Applying your missingness model,
 - ▶ Seeing how your inference goal is affected by that missingness,
 - ▶ Only proceeding if it is not!
- ▶ **Activity 7**: checking the imputation models.

Note on special values

- ▶ Imputation procedures can only handle special values appropriately if they know about them.
- ▶ **Cyber data are full of special values:**
 - ▶ 0 is often special: 0 bytes in a packet mean that a data transfer failed; 0 counts of an event may mean that a detector had failed, etc.
 - ▶ Often a **zero-inflated** model is needed to handle this: the data are either zero with some probability, or taken from their usual distribution.
- ▶ Other values are special.
 - ▶ Ports are all special and should often be considered as categorical. There are magic numbers in packet size that give away some protocols.
- ▶ **Categorical variables** in general are particularly hard to impute.
 - ▶ If you use “best guess” you may change the mean as the most frequent option is artificially even more frequent. Other guesses are worse on average.

Missing data Roundup

- ▶ Cyber data are often missing at the data collection stage: the collection procedure is so hopelessly **biased** that additional bias from the treatment of missing data is negligible.
- ▶ In this case, ask questions that you believe are **robust** to the data that were available, or are specific to them.
- ▶ For example, if you are lucky you may get a good dataset of what your company's network traffic looks like, at a given time, at the perimeter.
 - ▶ So ask questions about **changes** to the perimeter over time, not questions about what is going on over the network as a whole.

Reflection

- ▶ How do you know what type of missingness are in your data?
- ▶ What are the approaches to handling this? What are the challenges?
- ▶
- ▶ By the end of the course, you should:
 - ▶ Be able to QC your data for missingness,
 - ▶ Be able to appraise others' QC attempts,
 - ▶ Be able to perform basic imputation.

Signposting

- ▶ Next session: Workshop on Missing Data.
- ▶ Next block: Moving closer to advanced **machine learning** with Latent Dirichlet Allocation, and the high-level view of the Bayesian methodology that underpins it.
- ▶ Further reading:
 - ▶ Chapter 9.6 of The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Friedman, Hastie and Tibshirani)
 - ▶ Andrew Gelman's Missing Data Notes