DS 7333 – Quantifying the World

LIVE SESSION - 9

Handling Missing Data

World Changers Shaped Here

SMU

Imputation

- Missing Data Types
- Methods of Filling In
- Questions

Types of Missing Data

- **MCAR (Missing Completely at Random)**
  - No reason why! Totally random
- **MAR (Missing at Random)**
  - Difference between MCAR and MAR is MAR is typically missing only in a subset of data – the subset of data is determined by another value, and randomly missing; often requires domain knowledge
- **MNAR (Missing NOT at Random) – missing data can contain information**
  - There is definitely a pattern as to why the data is missing

Listwise Deletion

- Use only complete data
- If you have missing data drop that “row” or observation
  - Example:
    - In some study, data collected on height, weight, and age. Each “person” is a row of data or an observation. If any of the data is not collected (i.e., we forget to weigh someone) we discard ALL of the data from that person (their height and age for example).
    - We have a list of people with data in rows.
    - We delete people from the list with incomplete data
- Great for “clean” data sets
- (I’ve never used it)

Pairwise Deletion

- Estimate statistics based on actual data
  - Many methods use results of the individual data distribution (means, standard deviation) to build a model
  - Pairwise distribution is a good choice for these models
- Be aware when doing column comparison if the method is pairwise vs. listwise

Single Imputation

- Used quite frequently in my experience
- Many advanced methods do not handle missing data (R, Python)
  - Make incorrect assumptions
    - Missing = “0”
    - Program cannot handle NA/Blanks

Example

Id	Timestamp	Full_sq	Life_sq	Floor	Max_floor
8952	7/4/2013	123	123	3	10
8953	7/4/2013	25	10	NA	NA
8954	7/5/2019	32	NA	4	NA
8955	7/5/2019	51	30	4	NA
8956	7/5/2019	40	20	5	9
8957	7/5/2013	142	78	NA	NA
8958	7/5/2013	64	NA	5	NA

Look at your data

Full_sq	Life_sq
123	123
25	10
32	NA
51	30
40	20
142	78
64	NA

These columns seem closely related!
We could compute Life_sq by:

- A regression of Full_sq
- Substitute Full_sq when Life_sq is missing
- Drop Life_sq altogether
- Perform an ensemble approach using all of the above

Hot Deck

- Use similar values to estimate missing data
  - Guess the price of size of a home based on neighborhood
  - Guess the weight of a child based on age
  - Find correlations by k-means
- Another method – can be combined with means/regression (take the mean of the group and fit the group with a linear regression based on some additional variable).

Resources

- https://www.displayr.com/different-types-of-missing-data/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130338/
- https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4
- https://www.kaggle.com/rdslater/data-cleanup-munging