Introduction to Missing Data

Almost All Datasets Will Be Incomplete

- Incomplete can take several forms.
  - Missing data
  - Incomplete data
  - Incorrect data
- One of the biggest challenges is dealing with incomplete data.

The Bulk of Incomplete Data Is Missing Data

- How do we know it is missing?
- What do we do if it is missing?
- What kind of impact does missing data have on our analysis?

Imputation: Dealing with Missing Data

- The way we deal with missing data is to make a guess.
  - The better our guess, the better we can deal with missing data.
- Why can't we just ignore the missing data?
  - Our algorithms many times make assumptions when data is missing.
    - Those assumptions could be good or bad.
  - Some algorithms simply error when data is missing.
    - We have to supply **something.**
      - Usually the default is 0.
      - Is 0 the best guess?

DataScience@SMU

Missing Data Patterns

Patterns of Missing Data

- There are three patterns of missing data.
  1. Missing completely at random
  2. Missing at random
  3. Not a random
- The pattern of missing data can give clues for imputation.
- Identifying the pattern is the first step to identifying a method of imputation.

Missing Completely at Random

- No pattern!
- There is no indication why the data is missing.

Age	Value
15	62
32	
25	15
4	
71	
55	165
14	22
44	161
31	
51	25
26	

Missing at Random

When controlled for a third variable, the pattern is random.

Age	Class	Value
15	1	62
32	2	
25	1	15
4	1	51
71	2	
55	1	165
14	2	22
44	2	161
31	1	52
51	2	25
26	2	

Not Random

Pattern to the data

Age	Class	Value
15	1	62
32	2	
25	1	15
4	1	51
71	2	
55	3	165
14	3	22
44	3	161
31	1	52
3	2	
26	2	

DataScience@SMU

Imputation of Missing Data

Imputation

- Ignore missing data.
- Use a method based on summary statistics.
- Replace missing data with our best guess.
  - Guess is subjective
    - Mean
    - Median
    - Mode
    - Fit
    - Sample

Pairwise Deletion

- Delete any rows with missing data (even if more than two columns)
- Can substantially reduce dataset size
- Really only useful with small amounts of missing data

Age	Value
15	62
~~32~~	~~NAN~~
25	15
~~4~~	~~NAN~~
~~71~~	~~NAN~~
55	165
14	22
44	161

NAN is how we represent missing data, especially in Python. It is short for “not a number.”

Listwise Deletion

Uses summary statistics to model data.

- Limits your model choices!
- Different columns have different statistical power.

Age	Value
15	62
32	NAN
25	15
4	NAN
71	NAN
55	165
14	22
44	161

Column	Mean	Std. Dev	N
Age	32.50	21.32	8
Value	85.00	65.69	5

Fill In Missing Values

- There are a number of ways to attack this:
  - If the data is normally distributed, try the mean (as in replace all missing values with the mean).
  - If the data is categorical, try the mode.
  - If the data is non-normal, try the median.
  - If multiple variables are correlated, either drop the column with missing data or perform a fit to predict missing data.

Age	Value
15	62
32	NAN
25	15
4	NAN
71	NAN
55	165
14	22
44	161

- There is also the possibility of replacing missing values with a value to “flag” the value as missing.
- In this case the “Value” column is always positive, so replacing missing values with “0,” “−1,” or even “−100” could allow the model to learn how to deal with missing values.

Fill In Missing Values (cont.)

- This is less science and more art form
- Leverage domain experience
- Use multiple methods and combine your results
  - AKA: create an ensemble
- Search for relations between variables

DataScience@SMU

Domain Knowledge

Domain Knowledge Is a Key Part of Imputation

- Understanding the problem can be a key help in figuring out missing data.
- Specialized expertise can give you insight into the data that may not be obvious.

Example: Housing Data

- Most of you have some experience with houses.
- How can we fill in Column 2? We could use any number of the methods listed in the imputation section, most likely the mean or median.
- But what if you knew something about Col 1 and Col 2?

Col 1	Col 2
59	NA
74	40
33	18
45	30
38	23
73	43
123	NA
71	40
80	44
40	20
56	29
32	18
45	30
43	27
38	22
32	21
87	50
75	NA
30	17
42	28

Domain Knowledge

- If you know what the columns represent, now we can make a much more accurate guess. The total area of a house is very closely correlated to the area of living space. So we can use the first column to predict the second.
- We do a simple linear fit, replace the missing value with the value from the first column, or do both and ensemble the result.

Total Square Footage	Living Square Footage
59	NA
74	40
33	18
45	30
38	23
73	43
123	NA
71	40
80	44
40	20
56	29
32	18
45	30
43	27
38	22
32	21
87	50
75	NA
30	17
42	28

Other Examples of Domain Knowledge

- If the material of a house is missing, why not look at what nearby houses are made of?
- If the square footage total is missing, why not look at nearby houses?
- What about the tax bill or heating bill—look at the neighbor's data. (The heating bill in Florida is probably a lot different than the heating bill in Minnesota).
- In short, use your knowledge of how the data is related to improve your guess in any way possible.

DataScience@SMU