Data Quality

Lecture 3

Xiaofei Shi

Learning objectives

- Definition and measures for data quality
- Implications of data quality
- How to detect issues in data quality?

The pain is real

Daniel Ting likes this

Felix Naumann • 3rd+
Professor at Hasso-Plattner-Institute (Potsdam, Germany)
1w • Edited •

My data won't load ...
...because nobody bothered to use escape symbols.
...because ` is not a proper quotation symbol.
...because the maximum line length is exceeded.
...because there is a header row.
...because there is no header row.
...because the first line is the table-name.
...because some lines are empty.
...because it is too large.
...because it is encoded in CP-1252.
...because columns are shifted every ten rows.
...because a numeric column contains a string in line 590450.
...because that column contains another string in line 844026.
...because some lines are two fields shorter.
...because umlauts are not supported.
...because someone added footnotes.
...because who uses § as a delimiter?
...because the file contains multiple tables.
...because tab and space are not the same thing.

How to define data quality

- Is it just the absence of selection bias?

How to define data quality

- Is it just the absence of selection bias?
- How do we define mental illnesses?

Conceptual definition:

Practical view:

Data can be wrong:

- Sampling
- Measurement
- Processing

More theoretical: reliability

- Reliability - consistency
  - E.g. sample proportions vs population proportions
  - Similar to variance

More theoretical: reliability vs validity

- Reliability - consistency
  - E.g. sample proportions vs population proportions
  - Similar to variance
- Validity - correctness
  - E.g. GPA vs true understanding
  - Can something be valid if it is not reliable?
  - Similar to ... ?

How to measure?

Reliability:

- Test re-test
- 2 sets of questions on the same topic

Validity:

- Ground truth or expert comparison

Implication: big data vs quality data

- Crowdsourcing / Mechanical Turk

- Expertly labeled data

How to detect there are data quality issues?

Usual advice:

- Outliers (or impossible values)
- Prevalence of missing data
- Inconsistency with prior knowledge or expectations
    - Inconsistency across datasets

Why do these issues above matter?

Consequences of mishandling data quality:

Outliers

- Sometimes the most sensitive are the outliers (e.g. tech whales)

Consequences of mishandling data quality:

Outliers

- Sometimes the most sensitive are the outliers (e.g. tech whales)

Prevalence of missing data

- Missing data can be informative

Consequences of mishandling data quality:

Outliers

- Sometimes the most sensitive are the outliers (e.g. tech whales)

Prevalence of missing data

- Missing data can be informative

Inconsistency with prior knowledge or expectations

- Can sometimes reveal new hypotheses about the system, e.g. when marketing efforts decrease the average pageviews for a tech company

Data quality depends on the problem at hand

You cannot predict all the problems the data will possibly support.

How do you know what matters?

Example: what problems can the linear regression solve?

How do you know what matters?

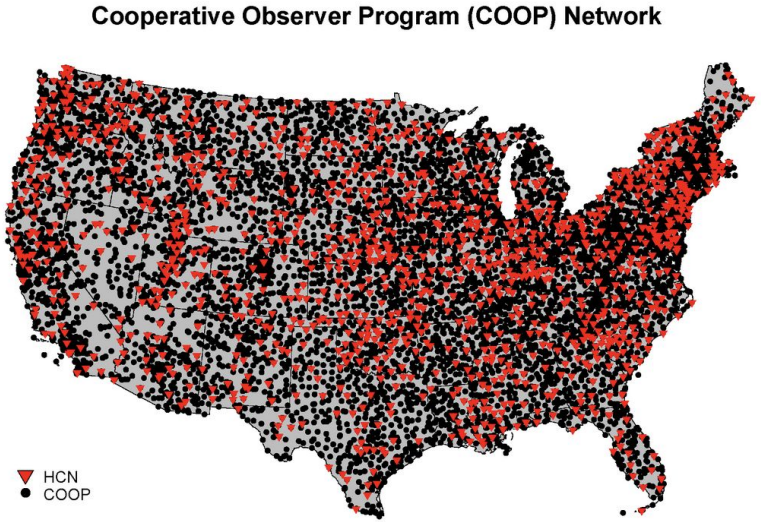
Example: what problems can the linear regression solve?

- Prediction

- Inference

Example: US Historical Climatology Network

US Historical Climatology Network (USHCN) measures historical measurements on daily min/max temperature, and precipitation.



Example: US Historical Climatology Network

US Historical Climatology Network (USHCN) measures historical measurements on daily min/max temperature, and precipitation.

What can go wrong with these measurements?

Flagging the data, let the researcher decide

```
QCFLAG: quality control flag, seven possibilities within
        quality controlled unadjusted (qcu) dataset, and 2
        possibilities within the quality controlled adjusted (qca)
        dataset.

        Quality Controlled Unadjusted (QCU) QC Flags:

        BLANK = no failure of quality control check or could not be
                evaluated.

        D = monthly value is part of an annual series of values that
            are exactly the same (e.g. duplicated) within another
            year in the station's record.

        I = checks for internal consistency between TMAX and TMIN.
            Flag is set when TMIN > TMAX for a given month.
```

Example: missing values

Missing values can be treated like a data quality flag.

Big data

- What does big data mean?
- What is its goal?

Can data collection be an ethical problem?

- Should we collect the candidate’s ethnicity in job interviews?

Summary

- Data quality depends on the problem at hand
- Trust in your data is *very* important, actively provide reasons to trust your data
- Statistical theory can inform what matters vs not in data quality
  - Necessary and sufficient conditions
- Big data has many new ethical concerns