Precept 6: Imputation Methods and Bootstrapping

COS424/524/SML302 Spring 2021

Xiaoyan Li

Topics:

- Assumptions on Missing Data
  - MCAR, MAR, and MNAR
- Imputation Methods
- Bootstrapping
  - Bootstrap methods for estimating prediction error

Missing Data

- Missing Completely at Random(MCAR)
- Missing at Random(MAR)
- Missing Not at Random(MNAR)

Missing Data

- **Missing Completely at Random(MCAR)**
  - when the probability of missing data on a variable is
    - unrelated to any other measured variable
    - and is unrelated to the variable with missing values itself.
  - Example: Some survey responses are randomly lost/damaged.

Missing Data

- **Missing at Random(MAR)**
  - when the probability of missing data on a variable is
    - related to some other measured variable in the model,
    - but not to the value of the variable with missing values itself.
  - Examples:
    - Older people are likely to have missing values for IQ; (The probability of missing data on IQ is related to Age)
    - Females are likely to have missing values for Age. (The probability of missing data on Age is related to Gender)

Missing Data

- **Missing Not at Random(MNAR)**
  - when the missing values on a variable are related to the values of that variable itself
  - Examples:
    - People with high incomes tend to have missing values for Income
    - Students with low GPA tend to have missing values for GPA

Missing Data in Fragile Family Studies

- **Item non-response:**
  - Respondents simply refuse to answer a survey question.
- **Survey non-response:**
  - Respondents cannot be located or refuse to answer any questions in an entire wave of the survey.

https://www.fragilefamilieschallenge.org/missing-data/

Missing Data in Fragile Family Studies

- **-9 Not in wave** – Did not participate in survey/data collection component
- **-8 Out of range** – Response not possible; rarely used
- **-7 Not applicable** (also -10/-14) – Rarely used for survey questions
- **-6 Valid skip** – Intentionally not asked question; question does not apply to respondent or response known based on prior information.
- **-5 Not asked "Invalid skip"** – Respondent not asked question in the version of the survey they received.
- **-3 Missing** – Data is missing due to some other reason; rarely used
- **-2 Don't know** – Respondent asked question; Responded "Don't Know".
- **-1 Refuse** – Respondent asked question; Refused to answer question

Topics:

- Assumptions on Missing Data
  - MCAR, MAR, and MNAR
- Imputation Methods
- Bootstrap methods
  - Bootstrap methods for estimating prediction error

Imputation Methods: single imputation

- (1) Mean imputation
- (2) Regression imputation
- (3) Matching methods
- (4) Last observation carried forward
- Come up with your own method…

Imputation Methods: single imputation

- (1) Mean imputation
  - Replace the missing value with the mean of the observed values for that variable
    - Similarly, can have median imputation, mode (most frequent) imputation
- (2) Regression imputation
  - Build a regression model for the variable X(with missing values) with the other observed variables as predictors/covariates,
  - train the model on the observed values for that variable.
  - Use the regression model to predict the missing value.

Imputation Methods: single imputation

- (3) Matching methods
  - Find similar samples (could use KNN) for a sample with missing value
  - Impute the missing value with the values of those similar samples
- (4) Last observation carried forward
  - Assume an individual is repeatedly measured over a period of time
  - Then the missing value can be imputed with the last observation of the individual(assume it has not changed since last observation).
- Come up with your own method…

Imputation Methods: multiple imputation

- Imputation process is repeated multiple times and generate multiple imputed datasets to deal with the uncertainty of missing values.
- Three steps:
  - Imputation
    - Missing values are imputed multiple times resulting in multiple data sets
  - Analysis
    - Each data set is analyzed
  - Pooling
    - Summarize the multiple results, calculate mean, and variance, etc.
- Many multiple imputation methods.

Topics:

- Assumptions on Missing Data
  - MCAR, MAR, and MNAR
- Imputation Methods
- Bootstrap methods
  - Bootstrap methods for estimating prediction error

The bootstrap: examples

The bootstrap answers the question: if I had another data set drawn from the exact same population, how different might my result be?

A statistic is a scalar numerical value that is quantified by applying a specific function to a collection of samples.

Examples of statistics to bootstrap

In next years' COS 424 survey, presumably drawn from a similar population as you, how different might our estimates be of:

- mean male and female height?
- variance of male and female height?
- mean number of siblings?
- proportion of females?

What can the bootstrap do?

The bootstrap was developed to quantify:

- statistic bias
- statistic variance
- confidence interval on statistic estimate
- prediction error
- ... and others.

A resampling method to recover these quantities is useful because often analytic calculations of these quantities are not available.

The bootstrap

To perform a bootstrap analysis for a data set with n samples:

The "case resampling" bootstrap method

- Repeat K times:
  - resample *with replacement* n samples
  - estimate the statistic of interest in this resampled data
- Estimate the probability of the true estimate of the statistic with respect to the K resampled estimates of the statistic
- Estimate the variance of the K resampled estimates of the statistic

Bootstrap: practical examples

First, resample the data $K = 1000$ times.

In every resampled data set, there will (probably) be multiple copies of some samples, and other samples will be omitted.

Next, compute the statistic of interest for each of the resampled data.

Then, look at the empirical distribution of the statistic across the K resampled data sets.

Bootstrap: practical example

Bootstrap the mean male height

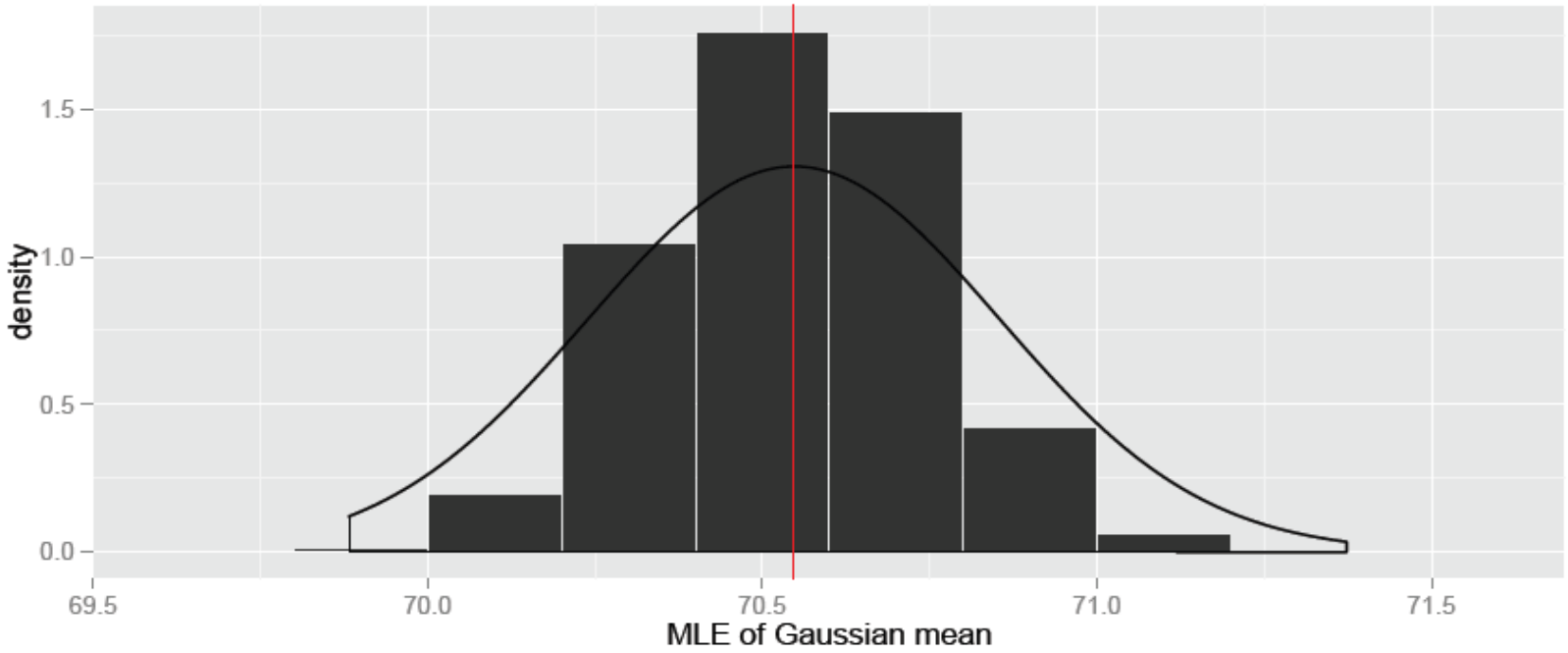
First, resample the data $K = 1000$ times: draw n_{males} samples with replacement from the set of males in our data.

For each set of resampled data set, y^k_{males}, compute the statistic of interest; here, the empirical mean:

$$\mu^k_{males} = \frac{1}{n_{males}} \sum_{i=1}^{n_{males}} y^k_{males,i}$$

Now I have $K = 1000$ different values of μ_{males}.

Bootstrap: mean male height



In these data, $n_{males} = 161$

Compute the confidence interval by finding the *empirical quantiles*: sort the 1000 resampled statistics, then the 25th and the 975th entry in the sorted list will be the lower and upper bound 95% confidence interval.

In other words, 95% of the resampled statistics fell within this interval.

Bootstrap Methods:

- This is one version of bootstrap (case resampling method)
- Other variants of the bootstrap
  - Parametric bootstrap
  - Bayesian bootstrap
  - Smooth bootstrap
  - and more …

Bootstrap methods for estimating prediction error

- Ordinary bootstrap method
  - Given a data set D ={(X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n)}
    - X_i is the feature vector representing the i^{th} observation, and Y_i is the label for the i^{th} observation
  - Generate K (say 1000) bootstrap samples: D_1, D_2,…, D_k, each sample D_i has n (X, Y) observation pairs randomly sampled from D with replacement;
  - Build a model on each bootstrap sample D_i;
  - Test it on the original data set D.
- Q: Do you see any problems with this method?

Bootstrap methods for estimating prediction error

- Ordinary bootstrap method
  - Given a data set D ={(X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n)}
    - X_i is the feature vector representing the i^{th} observation, and Y_i is the label for the i^{th} observation
  - Generate K (say 1000) bootstrap samples: D_1, D_2,…, D_k, each sample D_i has n (X, Y) observation pairs randomly sampled from D with replacement;
  - Build a model on each bootstrap sample D_i;
  - Test it on the original data set D.
- Problems:
  - Some observations are used in both training and testing
  - Tends to underestimate the prediction error

Bootstrap methods for estimating prediction error

- Bootstrap Cross-Validation
  - Given a data set D ={$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$}
    - X_i is the feature vector representing the i^{th} observation, and Y_i is the label for the i^{th} observation
  - Generate K (say 1000) bootstrap samples: $D_1, D_2, \ldots, D_k$, each sample D_i has n (X, Y) observation pairs randomly sampled from D with replacement;
  - Perform a leave-one-out cross validation on each bootstrap sample D_i
- Q: Do you see any problems with this method?

Bootstrap methods for estimating prediction error

- Bootstrap Cross-Validation
  - Given a data set D ={(X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n)}
    - X_i is the feature vector representing the i^{th} observation, and Y_i is the label for the i^{th} observation
  - Generate K (say 1000) bootstrap samples: D_1, D_2,…, D_k, each sample D_i has n (X, Y) observation pairs randomly sampled from D with replacement;
  - Perform a leave-one-out cross validation on each bootstrap sample D_i
- Problems:
  - The leave-one-out learning set may overlap with the left out item
  - Tends to underestimate the prediction error

Bootstrap methods for estimating prediction error

- Leave-One-Out Bootstrap
  - Given a data set D ={(X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n)}
    - X_i is the feature vector representing the i^{th} observation, and Y_i is the label for the i^{th} observation
  - Generate K (say 1000) bootstrap samples: D_1, D_2,…, D_k, each sample D_i has n (X, Y) observation pairs randomly sampled from D with replacement;
  - Build a model C_i on each bootstrap sample D_i
  - Each observation X_i in the original data set D is predicted using each model C_j if C_j is built on a bootstrap sample D_j in which the particular observation X_i does not appear. (Note: each observation X_i is predicted multiple times.)
- Q: Do you see any problems with this method?

Bootstrap methods for estimating prediction error

- Leave-One-Out Bootstrap
  - Given a data set D ={(X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n)}
    - X_i is the feature vector representing the i^{th} observation, and Y_i is the label for the i^{th} observation
  - Generate K (say 1000) bootstrap samples: D_1, D_2,…, D_k, each sample D_i has n (X, Y) observation pairs randomly sampled from D with replacement;
  - Build a model C_i on each bootstrap sample D_i
  - Each observation X_i in the original data set D is predicted using each model C_j if C_j is built on a bootstrap sample D_j in which the particular observation X_i does not appear.
- Problems:
  - A bootstrap sample contains roughly 0.632n distinct observations from the original data set.
  - Tends to overestimate the true prediction error when the sample size n is small (the test set will be very small.)

5 R Packages for Imputing missing values

- MICE: (Multivariate Imputation via Chained Equations)
  - Assumes MAR
  - Create muptiple imputations
- Amelia
  - Assume MAR, and all variables in a data set have Multivariate Normal Distribution (MVN)
  - Create muptiple imputations
- missForest
  - builds a random forest model for each variable.
  - uses the model to predict missing values
- Hmisc
- mi

Some Classes in sklearn

- sklearn.impute.SimpleImputer
  - 4 imputation strategy: mean, median, most frequent, constant
- sklearn.impute.KNNImputer
- *class* sklearn.cross_validation.Bootstrap
  - Provides train/test indices to split data in train test sets while resampling the input n_bootstraps times:
  - each time a new random split of the data is performed and then samples are drawn (with replacement) on each side of the split to build the training and test sets.
- sklearn.utils.resample
  - Resample arrays or sparse matrices in a consistent way.
  - The default strategy implements one step of the bootstrapping procedure.

Sample code for sklearn.cross_validation.Bootstrap

```
>>> from sklearn import cross_validation
>>> bs = cross_validation.Bootstrap(9, random_state=0)
>>> len(bs)
3
>>> print bs
Bootstrap(9, n_bootstraps=3, n_train=5, n_test=4,
random_state=0)
>>> for train_index, test_index in bs:
...    print "TRAIN:", train_index, "TEST:", test_index
TRAIN: [1 8 7 7 8] TEST: [0 3 0 5]
TRAIN: [5 4 2 4 2] TEST: [6 7 1 0]
TRAIN: [4 7 0 1 1] TEST: [5 3 6 5]
```

Paper: Wide and Deep Learning for Recommender Systems

- Group discussion for COS524 in breakout rooms (~15 minutes)

- Share your opinions in the shared google doc:
  - Will send you the links in chat
  - You can focus on some of the questions

- Come back for precept wrap up

Wrap up for Precept6

- Assumptions on Missing Data
  - MCAR, MAR, and MNAR
- Imputation Methods
  - Mean imputation
  - Regression imputation
  - Matching methods
  - Last observation carried forward
- Bootstrapping
  - Bootstrap methods for estimating prediction error

Resources: (Some materials are taken from the following resources. Some Bootstrapping slides are from previous year by Prof. Engelhardt.)

- "Missing data" by Iris Eekhout
  - https://www.iriseekhout.com/missing-data/
- Imputation (statistics) from Wikipedia
  - https://en.wikipedia.org/wiki/Imputation_(statistics)
- Tutorial on 5 Powerful R Packages used for imputing missing values
  - https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/
- A Comparison of Six Methods for Missing Data Imputation
  - https://www.omicsonline.org/open-access/a-comparison-of-six-methods-for-missing-data-imputation-2155-6180-1000224.php?aid=54590
- Bootstrapping(statistics) from Wikipedia
  - https://en.wikipedia.org/wiki/Bootstrapping_(statistics)
- "Bootstrap Prediction Error Estimation" by Wenyu Jiang* and Richard Simon
  - https://linus.nci.nih.gov/techreport/prederr_rev_0407.pdf