Using Multiple Imputation, Survival Analysis, And Propensity Score Analysis In Cancer Data With A Large Amount Of Missing Data

Master's Thesis

Nathan Berliner [1]

[1]Department of Statistics
Rice University

Fill in Date here, 2015

Outline

Outline

In an ideal world

- We would have a large dataset
  - That was obtained from an RCT
  - That would help answer a clearly defined question
  - That had all the covariates of scientific interest
  - That contained no missing data

In Reality

- RCT's are expensive and often unethical
  - We often get retrospective observational data
  - Pulled from a database or historical records
- The questions we have may not be answerable from the data on hand
  - The data obtained often doesn't support the original question in mind
- The covariates collected are out of our control
  - Since often no control of experiment, no control over what is collected
- Lots of missing data
  - Since no control over how the data is collected, we can't guarantee that everything is collected
  - This issue is seemingly omnipresent in all types of data collection

Is This a Problem?

- Without an RCT, we can't be sure if differences in treatments is due to the treatment or something else
- Omitting important factors may bias our results
- With missing data, we will be throwing away data and biasing our results

The Solution

This thesis aims to fix some of these problems

- Fill in missing data via multiple imputation
- Create meaningful analytical models via survival analysis
- Get a causal interpretation from observational data

Motivation

- This thesis is motivated by cancer survival data with moderate missingness
- We will build the theory for dealing with this sitation
- And then apply it to a cancer data set

Abstract

In this thesis, multiple imputation, survival analysis, and propensity score analysis are combined in order to answer questions about cancer data with moderate missingness. While each of these fields have been studied individually, there has been little work and analysis on using the three in trio. Starting with an incomplete dataset, we aim to impute the missing data, run survival analysis on each of the imputed datasets, and then do propensity score analysis to observe causal effects. Along the way, many theoretical and analytical decisions are mode. I explain why each decision is made, and offer ample evidence for the other choices such that the interested reader may implement the methods if they so choose. I apply the methodology to a cancer survival dataset in a case study, but the methods used are general, and could be adapted for any type of data.

Outline

- Missing data happens when we intend to collect a piece of data but don't actually get it
- Historical approaches
  - Complete Case analysis: Throw away any record that is not complete
  - Available Case analysis: Use records so long as they are complete for the specific analysis in question

Imputation

Definition

The English verb "to impute" comes from the Latin imputo, which means to reckon, attribute, make account of, charge, ascribe. [1]

- In the 1930's, Allan, Wishart,and Yates laid framework for missing data
  - Idea: Fill in the missing value, deduct degrees of freedom to account for it
  - Issue: Dogmatic, and variance can't be estimates correctly

Multiple Imputation

Throughout the 70's and 80's Donald Rubin worked to improve on this

- Instead of imputing one value, lets impute it $m \geq 2$ times
- Draw the values from the missing datas posterior distibution

This idea is called Multiple Imputation (MI) and was formalized in 1987 [2]. It is the gold standard method for missing data currently.

How does MI work?

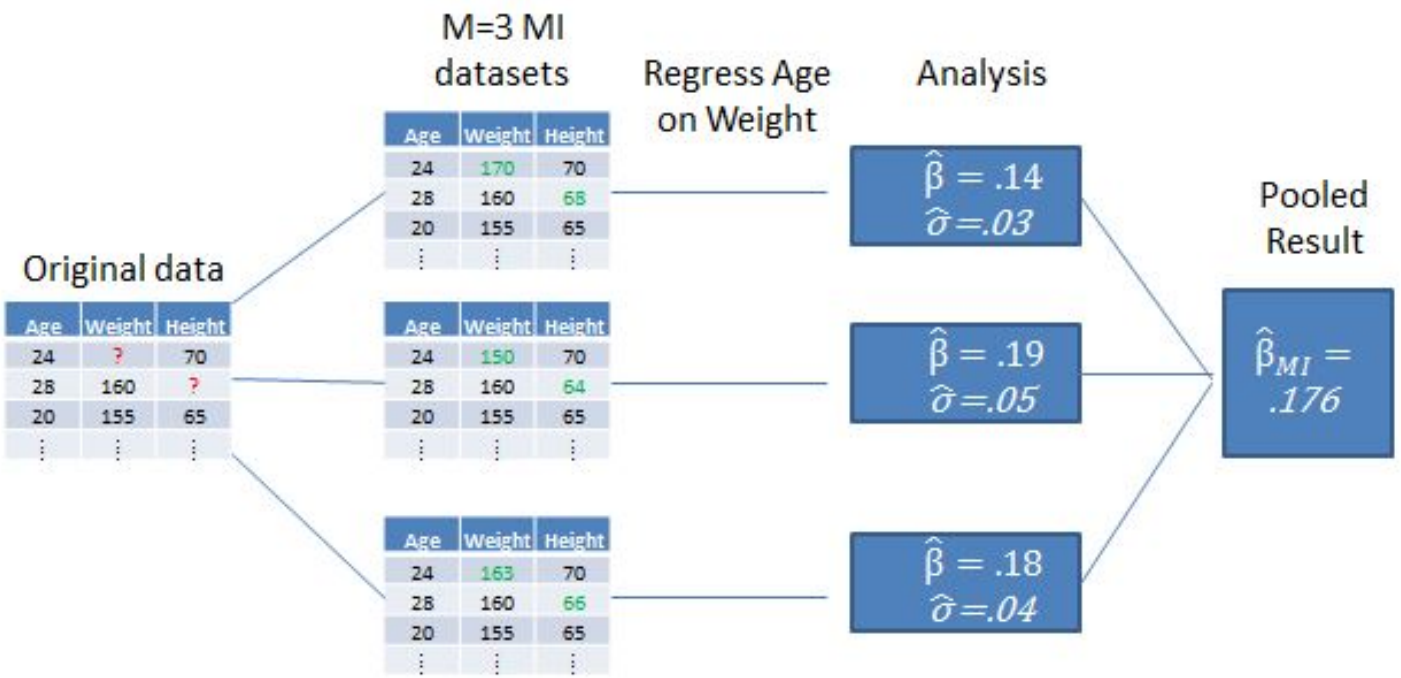

Figure: Visualization of MI data

Missingness is displayed by ?'s and the imputed data is shown as #'s. We then regress age on weight, get the results from the individual datasets, and then pool them together.

How to generate imputations

- FIll this in later

Rubin's Rules

Put them here

Outline

Survival Analysis

Survival Analysis

Survival analysis is a field of statistics concerned with analyzing time to event data, often in the face of censoring (not knowing the exact time of failure).

Examples:

- The survival of patients after a liver transplant in a hospital
  - Censoring: study ending, patients die before study starts, subject moves away
- The time until a child learns a new task
  - Censoring: Refuse participation, move away, don't recall the exact time they learned, already learned the task

Kaplan-Meier Estimator

- The survival function $S(t) = P(T > t)$ is estimated by the nonparametric Kaplan-Meier Curve

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

Where n_i is the number of subject in the risk set at time t_i and d_i is the number of deaths at time t_i

Log rank test

The log rank test compares two survival curves to see if from the same distribution

$$\frac{\sum_{j=1}^{J} w_j(O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{j} w_j^2 V_j}} \sim N(0,1)$$

- Where w_j is the weight of each observation (must be ≥ 0, we will set all to be 1)
- $N_j = N_{1j} + N_{2j}$ is the number at risk at time j(composed from deaths in each group)
- $O_j = O_{1j} + O_{2j}$ is the observed number of deaths at time j (composed from the observed deaths in each group)
- $E_{1j} = \frac{O_j N_{1j}}{N_j}$
- $V_j = \frac{O_j(N_{1j}/N_j)(1-N_{1j}/N_j)(N_j-O_j)}{N_j-1}$

Cox Regression

- Hazard is the instantaneous rate of event given that you have survived until time t, given by

$$h(t) = \lim_{\Delta t \to 0+} \frac{P[t \leq T < t + \Delta t | t \leq T]}{\Delta t}$$

- Cox regression models hazard by

$$h(t|Z) = h_0(t) \exp(\sum_{k=1}^{p} \beta_k Z_k)$$

- Where $h_0(t)$ is the baseline hazard
- Z_k is the k^{th} covariate
- β_k's are found by maximizing the partial likelihood function

The covariates act to multiply the hazard function.

Causal analysis

Suppose we have a new drug we want to test to see how efficacious it is.

- We would like to be able to say "The drug leads to better health"
  - But need an RCT to say this
  - We only have observational data
  - Thus differences could be attributed to the drug or another factor (like healthier people decided to take the drug)

  Idea: try to balance the covariates so the two groups seem identical at baseline

Counterfactual Model

Fill this in later.

Second Slide Title

- First item.

Second Slide Title

- First item.
- Second item.

Second Slide Title

- First item.
- Second item.
- Third item.

Second Slide Title

- First item.
- Second item.
- Third item.
- Fourth item.

Second Slide Title

- First item.
- Second item.
- Third item.
- Fourth item.
- Fifth item.

Second Slide Title

- First item.
- Second item.
- Third item.
- Fourth item.
- Fifth item. Extra text in the fifth item.

Outline

Blocks

Block Title

You can also highlight sections of your presentation in a block, with it's own title

Theorem

There are separate environments for theorems, examples, definitions and proofs.

Example

Here is an example of an example block.

Summary

- The first main message of your talk in one or two lines.
- The second main message of your talk in one or two lines.
- Perhaps a third message, but not more than that.

- Outlook
  - Something you haven't solved.
  - Something else you haven't solved.

References I


S. Van Buuren, *Flexible Imputation of Missing Data*. 2012.


D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. No. JOHN WILEY & SONS, 1987.