RICE UNIVERSITY

Put a cool title here

by

Nathan Karmazin Berliner

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Arts Statistics

Approved, Thesis Committee:

Rudy Guerra, Committee Chair
Professor of Statistics

Kenneth Hess, Thesis Director
Professor, MD Anderson Cancer Center

Yu Shen
Professor, MD Anderson Cancer Center

Marina Vannucci
Professor and Chair of Statistics

Houston, Texas

April, 2015

ABSTRACT

Put a cool title here

by

Nathan Karmazin Berliner

In this thesis, we will unify the theory of multiple imputation, survival analysis, and propensity score analysis. While each of these fields have been studied on their own intensely, there has been little work on using the three in trio. Starting out with an incomplete dataset, we aim to impute reasonable imputations, run survival analysis on each of the imputed datasets to get survival estimates, and then do propensity score analysis to observe causal effects. I apply the methodology to cancer survival in a case study, but the methods used are general, and could be used for any type of data.

Contents

Illustrations

Tables

Chapter 1

Introduction

1.1 Motivation

I want to develop methodology that can be used both by applied researchers and clinicians. I want it to be easy enough to describe to someone with a limited statistical background, but meaningful and valid so that the results obtained can be used in publication. The desire to have it this way stems from working on a project with both statisticians and clinicians.

Missing data is a major problem in both applied and theoretical statistics, however, it has not received attention proportional to its need. Survival analysis is a well studied field, but is relatively complete, so not much new research comes out of this field. Propensity score analysis will help us determine causal relationships when we dont have a completely randomized experiment. As one could imagine, all three of these fields are important to the applied statistician, as they will come across at least one at some point in their career. I will explain each of these three disciplines in detail before we dive into combining them.

1.2 Imputation

Imputation (specifically multiple imputation) is a way to fill in missing data, and it forms the base of our plan, all of the other analyses follow from it. Thus, we need a good understanding of it before we may proceed. Multiple imputation itself though is

a recent development, so we need to understand why it is used. At first, statisticians payed no attention to missing data, and happily discarded records for their analysis that were incomplete. This was known as complete case analysis. There are many problems with this paradigm. To begin with, you will lose a lot of statistical power when doing this, because you are literally throwing away records and thus decreasing your sample size. In addition, this can be costly to the researcher. If it costs a set amount to collect a single record, and you dont use this record, you are literally wasting money. As well, in some rare cases, incomplete data might be the only type we can get. Lastly, and most importantly, we will be biasing our estimates if we discard them. For example, if we have a random sample of people and are testing a drug, and want to run a regression on some collected covariates. Men are known to not want to give all of their information, so they leave them blank. In the analysis, we will need to discard the male samples because they are incomplete, leaving us only with women. Thus, we dont have a random sample anymore, and will get biased results. The next wave of statisticians wanted to improve upon this, so they developed what we now call today (single) imputation. In single method (such as regression, trees) is used one time to impute the missing value. While this is a little better than complete case analysis, it still has many drawbacks. Asserting that a single value is the true value is just foolish. There is always some amount of error involved, and we can in no way be 100% confident that our imputed value is correct. Furthermore, if I impute one value and you impute another, we may get totally different results from analysis on the data. This is obviously not desirable. In addition, imputing one time and calling it your data will artificially increase your sample size. You are in effect treating the imputed values as if they were real. If this was valid, we wouldnt ever need to collect samples, and we could just be armchair statisticians. While single

imputation certainly has its drawbacks, the idea of actually trying to fill in the data is an important one, and multiple imputation fills in the theoretical gaps. In 1987, Donald Rubin proposed multiple imputation as a solution [SAUCE]. The central idea is to make many datasets for the missing data, drawing the estimates for the missing data from the parameters posterior distribution. Once we have a sufficient number of datasets, we can run whatever analyses we would like on them, and then pool the results. We can get the standard errors by noting the within and between imputation variance. This is obviously much better than the first two methods because it allows is to not throw away data, as well as allowing us to quantify our uncertainty about imputing the missing values. The only real drawback of multiple imputation is that we still dont have true data, but we can be confident enough in our estimations to compensate for that. Multiple imputation use has been steadily increasing over the past 30 years, and it is now the standard for missing data.

1.3 Survival

Survival analysis is a huge field, and there have been many textbooks written about it. I only plan to introduce the topics that are relevant to my case study. For a much more detailed introduction, please see X Y Z Survival analysis on the whole is analysis of time to event data, often in the presence of censoring. The two main tools that we will be using are Kaplan-Meier estimate and Cox regression. The Kaplan-Meier estimate is a non-parametric estimate of the true survival function (the probability that you survive after a time t). Cox regression is a modelling tool that allows us to analyze the hazard ratio of a covariate. Using cox regression, we can make statements such as increasing the drug by one mg will increase its hazard by 30

1.4 Propensity Score Analysis

¡¡THIS NEEDS A LOT MORE WORK¿¿ In an ideal world, we would like to be able to do research and say that A causes B, not something along the lines of our study says that A is related to B. Using propensity score matching, we are able to get this interpretation, via Rubins causal model. Rubins causal model is a framework that allows us make causality statements, and propensity score is a tool that grants us access to the framework. The way propensity score analysis works is that we will match patients on their propensity to be in the treatment group. This is often done by logistic regression, although newer methods include regression trees and such.

Appendix A

Appendix

adfsdfsdf