METHODS TO CONTROL FOR OVERT AND HIDDEN BIASES IN COMPARATIVE STUDIES

by
Carrie A. Hosman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2011

Doctoral Committee:

Associate Professor Ben B. Hansen, Chair
Professor Robert W. Keener
Professor Edward D. Rothman
Associate Professor Michael R. Elliott

© Carrie A. Hosman 2011
All Rights Reserved

To my parents and grandparents,
for teaching me the value of education in all its forms

ACKNOWLEDGEMENTS

I could not have completed this work without the support and guidance I received throughout its development, and for that I am very grateful. I especially appreciate the mentorship and support I have received over the years from my advisor, Ben Hansen, who taught me so much about Statistics and the research process. Thank you also to the members of my committee – Ed Rothman, Mike Elliott, and Bob Keener–for guidance and feedback on my research. An additional thank you goes to Ed Rothman for being a supportive mentor from my earliest research experiences as an undergraduate through my time at CSCAR and the writing of my dissertation.

No journey is quite as enjoyable when traveled alone, and I think this has been very true of the development of this thesis. For being in the proverbial trenches with me, thank you to my fellow graduate students in the Statistics department. Whether you went before me and provided encouragement and advice when it was most needed, dropped in for a few years to bolster and entertain me, or stuck around the whole time helping me and supporting me from start to finish, I truly appreciate your presence on this journey. In particular, Matt, thank you for all of the above – encouragement, advice, entertainment, assistance – and a healthy dose of patience. I would offer you the “h” to express my gratitude, but I don’t think you will need any more letters after your name.

The development of this thesis was aided by financial support from the National Science Foundation (SES-0753164) and the National Security Agency (H98230-08-1-0089). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or National Security Agency.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

METHODS TO CONTROL FOR OVERT AND HIDDEN BIASES IN COMPARATIVE STUDIES

by
Carrie A. Hosman

Chair: Ben B. Hansen

When the goal of a comparative study is to ascertain the effect of some treatment condition, problems arise when it is not randomly assigned to units. In the absence of random assignment, units compared cannot be expected to be similar in terms of pretreatment covariates, yet the validity of resulting causal inferences relies on this equivalence. This thesis develops techniques that build upon existing methods to analyze comparative studies, lifting certain of their limitations. These methods focus on reducing bias due to nonequivalence of covariates across groups and can be easily combined with techniques that aim to reduce other biases, such as those that arise from a mismatch in the sample and target population.

To reduce bias in estimates from comparative studies, the best analysis ensures the likeness of the distributions of measured confounders across comparison groups. Methods such as matching or post-stratification on the measured covariates group similar units, and analysis is performed within subgroups. We apply this bias-

reducing idea to the Peters-Belson method, which assesses the existence of a disparity with regression models, to restrict comparisons to groups of units with similar covariate distributions. Propensity scores are a common way to organize units into groups. In practice, the propensity score is estimated by a parametric model, and the literature is divided regarding the selection of the best model. Consistent with one thread of the literature, we develop a method that improves the propensity score model by focusing it on covariates most relevant to an outcome of interest with the creation of a multidimensional prognostic score. By improving the propensity score model, units compared are more similar, and resulting analyses have greater validity.

While adjusting for measured confounders can sometimes suffice in the analysis of comparative studies, additional methods – broadly known as methods of sensitivity analysis – aim to quantify the potential impact of unmeasured confounders on the effect estimate. We introduce a method of sensitivity analysis for a linear regression model that is unique in its simplicity and ability to assess the impact of unmeasured confounders on the entire confidence interval, rather than only the point estimate.

CHAPTER I

Introduction

In order to assess the impact of influences as diverse as medical interventions, educational programs, and neighborhood effects, methods of analysis are necessary that can be used with or without a randomized experiment. With a randomized experiment, it is readily assumed that there are no systematic differences in groups that received various treatment conditions prior to the application of treatment. This assumption may be spurious in practice, especially in experiments with small sample sizes. In addition, with an observational study, this assumption of no systematic differences across groups is not so easily made. Frequently, observational studies are subject to such differences, but specific statistical techniques can raise awareness of this disparity and allow adjustments to be made to the data prior to the estimation of treatment effects. Many methods exist in the current literature to address the potential dissimilarities in covariate distributions that result from a lack of random assignment; however, these methods have limitations that should be addressed. The aim of this dissertation is to discuss several of these existing methods, while offering possible solutions to the shortcomings of the techniques.

1.1 Observational Studies

Cochran's often cited definition of an observational study describes it as having the objective "to elucidate cause-and-effect relationships" in an investigation in which "it is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures" (Cochran, 1965). This definition presents the prominent issue in parameter estimation in observational studies; without "controlled experimentation" – that is, the random assignment of units to treatments – standard techniques used to assess treatment effects in experiments are not necessarily valid. With the random assignment of treatments to units, we can follow the ideas of Fisher (1935) who argued that randomization was the "reasoned basis for inference" in experiments. With randomization, we can assume that there are no systematic differences between the group of subjects who received the treatment under study and those who did not; theoretically, the two groups should be comparable in terms of measured and unmeasured covariates, characteristics typically observed prior to receiving treatment (Rosenbaum, 2002). Randomization may, of course, create chance imbalances between the two groups, but standard statistical techniques can manage any uncertainty created by these imbalances.

When assessing information from an observational study, it is often difficult to disentangle the effects of pretreatment covariates from the effects of treatment itself as it is possible a certain effect can be caused by factors besides the particular treatment of interest. As Cornfield et al. (1959) explained, "a universe in which cause and effect always have a one-to-one correspondence with each other would be easier to understand, but it is obviously not the one we inhabit". Consequently, an analyst must make satisfactory adjustments to prevent pretreatment differences from impacting the estimation of the treatment effect parameter. Pretreatment differences can only be accounted for if they are known; it is impossible to know and measure every possible way in which units will differ prior to treatment. Methods of adjusting for imbalances in the measured, or observed, covariates can be defined as adjusting for *overt bias*. For unmeasured confounders, it is instructive to know the extent to which such confounders could bias our estimates. Methods of sensitivity analysis attempt to quantify the extent of this bias, and these sensitivity analyses comprise an assessment of *hidden bias*, or bias due to unmeasured confounders.

With observational studies that lack random assignment, it is important to check the similarity of treatment groups in order to compare groups with similar covariate distributions. Methods to adjust for this *overt* bias often involve some type of matching or stratification with the goal of placing treatment and control units into groups with similar covariate distributions. The adjustment methods can be viewed as data pre-processing methods; after the pre-processing steps are completed, under certain assumptions and with appropriate techniques, the data are analyzed as though they were from a block-randomized experiment. In this way, these adjustment methods aim to bring an observational study closer to an experimental ideal in which there are no systematic differences in covariates across treatment groups. Fundamentally, the goal of these pre-processing steps is to alert an analyst to situations in which groups with dissimilar distributions of covariate distributions are being compared and refocus treatment effect estimation on groupings of units in which covariate distributions appear to be alike. When situations arise in which careful design of a comparative study does not result in distributions of covariates that appear as though they could have been produced from a block-randomized experiment, proper analysis methods that bring attention to any differences are crucial. In these cases, the credibility of analysis relies on the ability of the adjustment methods to ensure the the similarity of the units being compared.

1.1.1 The potential outcomes framework

To better discuss effect estimation for observational studies, it is useful to understand the potential outcomes framework, which is discussed at length by Holland (1986). The general potential outcomes framework is applicable to experiments and observational studies, but this discussion focuses on a few details about the application to observational studies specifically. The potential outcomes framework begins with the set of potential outcomes for a population of units, U, and a particular cause or treatment of interest. If we allow Z to be the variable that denotes the level of treatment to which a unit u was exposed, for any u, we can have either $Z(u) = 1$ or $Z(u) = 0$, denoting that the unit u received either the treatment or control condition, respectively. Let Y represent the outcome where we denote $Y_t(u)$ and $Y_c(u)$ as the response observed for the same unit u under treatment and control conditions, respectively. Typically, we want to determine the causal effect for each unit u, which is simply the difference in the responses under each condition, or $Y_t(u) - Y_c(u)$.

The response actually observed can be written as $Y(u) = Z(u)Y_t(u) + (1 - Z(u))Y_c(u)$. Each unit u can only receive either the treatment or control condition, so only one of either $Y_t(u)$ or $Y_c(u)$ can be observed – not both simultaneously for the same unit. This conundrum was coined the Fundamental Problem of Causal Inference by Holland (1986) because, if we cannot know $Y_t(u)$ and $Y_c(u)$ simultaneously for one unit u, then we cannot observe the causal effect of the treatment on unit u. The statistical solution to this problem, which depends heavily on certain assumptions which will be discussed, makes use of information on many different units to establish an average causal treatment effect. Instead of simply observing $Y_t(u) - Y_c(u)$, we find a treatment effect, often referred to as the Average Treatment Effect (ATE)

$$\tau = E(Y_t(u) - Y_c(u)) = E(Y_t(u)) - E(Y_c(u)), \tag{1.1}$$

which allows us to look separately at those units for which $Z(u) = 1$ and $Z(u) = 0$ to estimate τ, as we do not have access to information from some parallel reality in which treated units are observed under the control condition and control units are observed under the treatment condition. In a randomized experiment, a difference in average outcomes observed under treatment and under control provides an unbiased estimate of the ATE. In an observational study, however, substantial bias can exist if the units for which $Z(u) = 1$ have a different distribution of pretreatment covariates that that of those units for which $Z(u) = 0$. An estimate of the ATE may capture a treatment effect or it may simply be a consequence of the difference in pretreatment covariates.

In order to obtain an estimate of the average causal effect of treatment, two assumptions must be made. First, it is assumed that the response of a particular unit varies depending on the treatment received by that unit, but it does not depend on the treatments assigned to any other units. This is what is known as the "stable unit treatment value assumption" or SUTVA. In addition, the treatment assignment is assumed to be strongly ignorable. Letting $\mathbf{x}$ be a matrix of observed covariates, it can be said that the treatment assignment is strongly ignorable if for a unit u, the vector of potential outcomes $(Y_t(u), Y_c(u))$ is independent of the treatment assignment $Z(u)$ conditional on $\mathbf{x}$, and if at each value of $\mathbf{x}$, there is a positive probability of receiving each treatment. In other words, if we take Y_t and Y_c to be vectors of potential outcomes for all units, treatment assignment is strongly ignorable if

$$(Y_t, Y_c) \perp\!\!\!\perp Z | \mathbf{X} = \mathbf{x} \tag{1.2}$$

and if,

$$0 < P(Z = z | \mathbf{X} = \mathbf{x}) < \mathbf{1} \tag{1.3}$$

for $Z = 1$ or $Z = 0$ and for all $\mathbf{x}$.

There are many methods to test the assumption of strong ignorability (see, for example, Rosenbaum (1984), Hong and Raudenbush (2006)), but in many analyses of observational studies, it is assumed to hold. If strong ignorability holds, the causal model described above can be used to represent an investigation of a cause and effect relationship in an observational study. The reliability of the assumption of strong ignorability forms the major difference between using the causal model in observational studies and experiments: in randomized experiments, treatment assignment is strongly ignorable, whereas in observational studies, ignorable treatment assignment is only an assumption (Rosenbaum, 2002)

Despite the difficulties that arise in analyzing observational studies, they are an essential source of information for researchers in many disciplines. With an understanding of the potential outcomes framework for observational studies and models to describe cause and effect relationships in them, issues that arise in using observational studies to estimate parameters can be discussed more concretely.

1.2 Limitations of existing methods and proposed improvements

Techniques to analyze observational studies encompass two themes: adjustment methods for observed information to avoid bias from comparing units that are dissimilar and sensitivity analyses to assess how sensitive conclusions are to bias from unmeasured confounders. These methods, and consequently the methods developed in this dissertation, pertain to internal rather than external validity. Although these new methods are illustrated on data for which external validity may be desired, this consideration is beyond the scope of this work. Methods to appraise and account for bias due to observed and unobserved confounders, however, are readily combined with statistical techniques that aim for external validity and generalizability, and the exploration of these connections form a related set of research (see, for example, Stuart et al. (2011)). Within each subsequent chapter of this dissertation, the literature on techniques to reduce and assess bias due to observed and unobserved confounders is explored to ground the discussion and illustrate the need to improve existing methods. There is a wealth of existing techniques that fall under both themes, and the developments in subsequent chapters build on these methods to improve upon them.

1.2.1 Drawing observational studies toward an experimental ideal

The first thread of the work discussed in this dissertation develops methods for accounting for overt bias in order to bring any comparative study toward the ideal of an experiment. Chapter II centers on a method used to assess a disparity in some outcome of interest between a majority and minority group. The technique, known as the Peters-Belson method (from the work of Peters (1941) and Belson (1956)), offers a clear framework to assess a disparity. When it is applied to observational data or an experiment in which resulting covariate distributions are not similar across comparison groups, the technique is subject to criticism due to the potential for problematic extrapolation. The method involves fitting a regression surface to a majority group, extrapolating the fit to a minority group, and estimating the average residual of that minority group to examine the effects of being a member of that minority group. To remedy problems due to differences in covariate distributions, the method outlined in Chapter II incorporates propensity score matched sets into the analysis of the data. This modification focuses analysis on comparing similar units in order to improve the resulting confidence interval estimates. The method is motivated by a simulation study and illustrated with a case study of the effects of neighborhood characteristics on health outcomes. In addition to the discussion of the use of propensity scores, a diagnostic is developed to help an analyst select the best outcome model.

Chapter III explores a dimension reduction technique for covariate information that aims to improve the propensity score (Rosenbaum and Rubin, 1983). The propensity score, or the conditional probability of receiving treatment given the observed covariates, is used to create groupings of units with similar covariate distributions; conditioning on the true propensity score, the distributions of covariates are the same across treatment and control groups (for more detail of this result, see Rosenbaum and Rubin (1983)). In practice, the propensity score is estimated, and this model can sometimes be led astray by variables that have little relationship to the outcome of interest (Brookhart et al., 2006). We develop a method to improve the standard propensity score estimation by using recently developed prognostic scores. Prognostic scores, introduced as an alternative or complement to propensity scores, extract the part of the covariate space most relevant to an outcome of interest (Hansen, 2008); in this way, it is a dimension reduction method. In Chapter III, a method for selecting multiple prognostic scores is developed and illustrated on a study of the effects of an educational intervention on standardized test scores. Additionally, the use of a diagnostic and associated theory are discussed to help an analyst select the optimal dimensions for the prognostic score. The multidimensional prognostic score addresses the criticism that propensity score estimation done in a standard manner may not yield the most useful propensity score for the ultimate goal of estimating the effect of treatment on the treated units. The procedure focuses a propensity score on covariates most relevant to an outcome of interest, leading to less biased estimates of treatment effects.

1.2.2 Adjusting estimates for omitted variable (or *hidden*) bias

In a randomized experiment, random assignment can, in theory, ensure that the distributions of both measured and unmeasured confounders are similar across comparison groups. Adjustment methods can account for differences in measured confounders, but the same methods cannot be applied to unmeasured confounders. Unmeasured confounders that are unable to be 'controlled' for in the same way as observed confounders could affect estimates of the effects of certain interventions, and methods for sensitivity analysis which derive from the early ideas of Cornfield et al. (1959) allow a researcher to assess whether potential unmeasured confounders could nullify or reverse key conclusions of a study. The second thread of research in this dissertation explores a method to assess sensitivity to bias due to unmeasured confounders.

Related to ideas presented in Marcus (1997), Lin et al. (1998), Frank (2000), and Imbens (2003), Chapter IV presents a sensitivity analysis for multiple linear regression in terms intrinsic to the fitting of multiple regression models. This sensitivity analysis is unique in that it provides easily applicable formulas that yield adjustments for confidence interval estimates of treatment effect – including both point estimates and standard errors. Many common formulations for sensitivity analysis ignore adjustments to the standard errors, and this oversight can be shown to be potentially problematic, particularly with smaller samples.

CHAPTER II

Reducing extrapolation in the application of the Peters-Belson method to an observational study

2.1 Overview

In order to assess the impact of some condition or treatment on an outcome of interest, many statistical techniques could be used. One such procedure, known as the Peters-Belson method due to its origination in the work of Peters (1941) and Belson (1956), offers a method for analyzing the a treatment effect in a clear manner with familiar regression models. When it is applied to an observational study, it suffers the drawback of inattention to potential disparities in covariate distributions across treatment groups. This paper offers a solution to this problem by combining the Peters-Belson method with propensity score matching. Propensity score matching restricts analysis to groupings of units in which the covariate distributions are alike to improve the estimation of treatment effects. To alert an analyst to cases when overfitting of the Peters-Belson model will create additional problems beyond extrapolation, a diagnostic is presented and illustrated.

2.2 The Peters-Belson method

For much of the past century, statisticians, economists, sociologists, and other scholars have attempted to explain disparities between groups of people. For example, to study a wage gap between the earnings of Ivy League graduates and public college graduates a natural question arises: how much would a graduate from a public college with certain qualifications and experience be expected to earn if he were in fact a graduate of an Ivy League college? Obviously, this quantity cannot be known with certainty for someone who is a public college graduate, but the estimation of it is central to the Peters-Belson (PB) method, which originated with the work of Peters (1941) and Belson (1956). The PB method, which is also known as the Blinder-Oaxaca method in Economics literature, was at one time a widely used method for the analysis of observational studies, and it is still applied to comparative studies, particularly for those analyses examining a disparity between a majority and minority group (Gastwirth and Greenhouse, 1995). It allows the use of standard regression models to assess these disparities, but it does not force an analyst to commit to some of the assumptions inherent in using regression models. In the present example of the effect of college type on earnings, the expected earnings of a public college graduate had he actually attended an Ivy League college can be estimated on the basis of observed covariate information with standard statistical models. Under the assumption of ignorability, a regression model is fit to the control group to model the relationship between the covariates and the outcome and predictions from the model offer estimates of expected earnings. By the ignorability assumption, the distributions of the covariates should be the same for the groups of public college graduates and Ivy League college graduates. The comparison of the estimated expected earnings obtained from the model to actual earnings allows an analyst to detect if a difference exists; if it does, there seems to be an effect on earnings of attending a public college as opposed to an Ivy League college.

2.2.1 Limitations of the PB method and proposed solutions

The general PB framework is subject to two limitations that are managed in this chapter. The central limitation discussed is the potential for the regression model to extrapolate, and it is especially problematic when the PB method is applied to observational studies. An additional limitation, one that is less central to this illustration but still important, arises when the PB method is applied to any type of study: variability of treatment effects can increase due to potential overfitting and sampling variability.

To help ameliorate difficulties due to extrapolation, we will illustrate how to combine the Peters-Belson method with optimal propensity score matching. In the context of a cluster-randomized experiment, simulations discussed in Hansen and Bowers (2009) illustrate that combining the PB method with randomization-based inference and adjustments for clustering yield good results in terms of the level and power of the resulting tests. Propensity-score matching in an observational study can be viewed as mirroring a randomized-block experiment that may have been done if feasibility and ethics allowed, as, within matched sets, observed covariates are distributed in such a way that treatment could have been randomly assigned. As a result, the properties of using the Peters-Belson method accounting for clustering should hold in observational studies with sets resulting from optimal matching on the propensity score.

To address the second, though not central, limitation that arises in the PB method of additional variability from overfitting or sampling variability in $\hat{\beta}$, Section 2.4.1 develops a diagnostic to alert an analyst to cases in which the additional variability may be problematic. The specification of the regression model used for the outcome analysis can be guided by this diagnostic.

2.2.2 Outline

In this chapter, the framework of the Peters-Belson method and proposed improvements to it are discussed in the context of experimental and observational case studies. The general PB method is explained in the context of an experimental case study in Section 2.3 to ground intuition before discussing the proposed improvements in the application of the method to an observational study. Through the development of a diagnostic for outcome model selection, Section 2.4 addresses a secondary limitation of the PB method: potential inflation of the variance of treatment effects. Though this limitation is not the central challenge addressed in this chapter, it can arise in the application of the PB method to any type of comparative study. Section 2.5 introduces an observational study of health disparities, which is analyzed by the PB method incorporating propensity score matched sets. To assess the performance of the diagnostic developed in Section 2.4, Section 2.6 describes a simulation study using the observational data to determine if enforcing the diagnostic criterion improves treatment effect estimates. Finally, in Section 2.7 the flexibility of the improved Peters-Belson method to examine doses of treatment rather than binary treatment scenarios is discussed.

2.3 The PB method in the context of an experimental case study

2.3.1 Data: The Milwaukee domestic violence arrest experiment

Before addressing the central problem this chapter aims to solve of resolving issues when the PB method is applied to an observational study, an experimental case study is discussed to introduce the PB method and a limitation of it when applied to any study. The experimental case study is the result of a series of studies of the effects of arrest on the subsequent behavior of individuals suspected of misdemeanor domestic violence. Specifically, this chapter uses data from the study that occurred in Milwaukee in 1987-1989, which randomized the action taken by police faced with a suspect accused of misdemeanor domestic violence for incidents and subjects that met certain eligibility criteria Sherman et al. (1992). If a suspect and incident met eligibility criteria, responding officers radioed headquarters to receive an assignment that was randomly selected from sealed envelopes. The three actions officers could take were to advise the suspect and not arrest him, arrest and promptly release the suspect, or arrest and hold the suspect. Recidivism rates were measured by rap sheets, domestic violence hotline calls, and subsequent victim interviews, so the initial publication of these findings (Sherman et al., 1992) uses many outcome measures. In addition, the publication explores many binary comparisons including arrest vs. no arrest, arrest and hold vs. no arrest, and arrest and hold vs. arrest and release. For purposes of illustrating the PB method, the focus in this section will be on comparing the two arrest conditions in which those suspects who were assigned to "arrest and release" are considered the control group, and those suspects assigned to "arrest and hold" are the treatment group. The outcome of interest is simply the number of subsequent arrests of the same suspect for any type of offense.

2.3.2 General formulation

Let $\mathbf{X}$ be a $n \times p$ matrix of covariates, Y be vector of observed outcomes, and Z be a vector of treatment assignments in which $Z_i = 0$ if unit i is in the control group and $Z_i = 1$ if unit i is in the treatment group. In the case of the Milwaukee experiment data, take $\mathbf{X}$ to be a 765×24 matrix of covariate information, Y to be the number of subsequent arrests recorded for each suspect, and Z to be a vector in which 376 units in the control group have a value of 0 and the 389 units in the treatment group have a value of 1. In the simplest sense, the PB method aims to test a weak null hypothesis and determine if the responses from those units in the treatment group differ, on average, from the responses that would have been expected had those units received the control condition instead.

To simplify initial discussions, consider a situation in which there is only one covariate of interest, or $p = 1$. For the Milwaukee experiment, one could consider the number of prior arrests as a single predictor for the number of subsequent arrests. As was indicated in the introduction, to perform any analysis with the PB method, an analyst must determine the expected response under the control condition for a unit that actually received the treatment condition. This can be accomplished by fitting a regression model, such as a standard ordinary least squares regression model, of Y_c on X in the control group (note that for a unit in the control group, $Y_c = Y$). From this model, a vector of estimated coefficients $\hat{\beta}$ allows estimation of predicted values of the response under control, denoted $\hat{y}_c(\hat{\beta})$, for the treatment group units based on their observed covariate information. With the Milwaukee experiment data, this amounts to creating a regression of the number of subsequent arrests on the number of prior arrests in the group that was arrested and released for which the number of prior arrests ranges from 0 to 23 prior arrests. The resulting model is $Y_c = 0.42 + 0.09X$, and this model is used to determine how many subsequent arrests we would expect for suspects who had been arrested and held supposing they had been arrested and released instead. Notationally, the model allows estimation of values of $\hat{y}_c(\hat{\beta})$ for those subjects who were arrested and held for which the number of prior arrests ranges from 0 to 19.

Denoting the size of the treatment group as n_t, a test of the weak null hypothesis of no average difference compares $Z'Y/n_t$ (the mean response observed in the treatment group) and $Z'\hat{y}_c(\hat{\beta})/n_t$ (the mean expected response of the treatment group, supposing the units had instead received the control condition). An estimate of treatment effect is given by

$$\bar{D} = \frac{Z'Y}{n_t} - \frac{Z'\hat{y}_c(\hat{\beta})}{n_t} = \frac{Z'(Y - \hat{y}_c(\hat{\beta}))}{n_t} \tag{2.1}$$

To test the weak null hypothesis of no difference between the responses observed in the treatment group and the responses that would have been expected had those units been in the control group, an additional assumption is required to conduct a t-test, where $t = \bar{D}/\sqrt{\mathrm{V}(\bar{D})}$ (Gastwirth and Greenhouse, 1995). In order to compute $\mathrm{V}(\bar{D})$ for this t-test, an analyst must assume that the residual variance of the outcome model for the treatment group is the same as the residual variance of the outcome model for the control group. (For further discussion of the computation of this variance, see Section 2.9.2.) While this assumption is in line with the idea of no difference between treatment and control groups, it is an assumption not required by other methods to test hypotheses that can be used with the PB method, as discussed later in this section. For the Milwaukee experiment data, when we consider only one covariate of prior arrests, this estimate of treatment effect is given by $\bar{D} = Z'Y/n_t - Z'\hat{y}_c(\hat{\beta})/n_t = 0.620 - 0.654 = -0.34$, so it appears that arresting and holding suspects rather than arresting and releasing them reduces the number of subsequent arrests by 0.34 arrests, on average. A t-test of the form described by Gastwirth and Greenhouse (1995) results in a test statistic of $t = -3.87$ and a p-value of 0.00, indicating that holding suspects after arrest does lead to a statistically significant reduction in the number of subsequent arrests.

It is rarely the case that an analyst has a single covariate, and when more covariates are considered, $Y - \hat{y}_c(\hat{\beta})$ will likely be nonzero in the control group. Thus, the estimate given by Equation 2.1 can be generalized to

$$\bar{D} = \frac{Z'(Y - \hat{y}_c(\hat{\beta}))}{n_t} - \frac{(1-Z)'(Y - \hat{y}_c(\hat{\beta}))}{n_c} \tag{2.2}$$

where n_c denotes the size of the treatment group. The test of the weak null hypothesis could proceed similarly, comparing $Z'(Y - \hat{y}_c(\hat{\beta}))/n_t$ to $(1-Z)'(Y - \hat{y}_c(\hat{\beta}))/n_c$ rather than $Z'Y/n_t$ and $Z'\hat{y}_c(\hat{\beta})/n_t$. With the Milwaukee data, if more covariates are considered, a new outcome model is created. The outcome model is chosen to be an OLS model with 24 mostly factor predictors describing the subject, the victim, the nature of the relationship between the victim and subject as well as their prior history for the outcome of total subsequent arrests of the same suspect for all offenses. Fitting the model and computing values of $\hat{y}_c(\hat{\beta})$, $Z'(Y - \hat{y}_c(\hat{\beta}))/n_t = 0.014$ and $(1-Z)'(Y - \hat{y}_c(\hat{\beta}))/n_c = -0.077$, the the resulting treatment effect estimate demonstrates that arresting and holding suspects leads to a significant increase of 0.91 subsequent arrests, on average, over arresting and releasing the suspects.

In addition to testing the weak null hypothesis of no average difference, the PB method can be used to test an array of strong null hypotheses, which revolve around the test that treatment has no effect whatsoever. Under the strong null hypothesis that treatment has no effect at all, all units should have the same response whether or not they are randomly assigned to the treatment or control condition. Note that in the language of potential outcomes, for a unit i for which $Z_i = 1$, $Y_i = Y_{ti}$ and for a unit i for which $Z_i = 0$, $Y_i = Y_{ci}$. Thus, under the strong null, we can infer that $Y_{ci} = Y_{ti} = Y_i$ for all i such that $Z_i = 1$. In the case of the strong null of no effect, Y_c values are implicitly inferred for all units:

$$Y_{ci} = \begin{cases} Y_i, & Z_i = 0 \\ Y_i, & Z_i = 1; \end{cases} \tag{2.3}$$

A strong null hypothesis could also posit a simple constant treatment effect τ, and Y_c values could be inferred for all units such that

$$(2.4) \qquad Y_{ci} = \begin{cases} Y_i, & Z_i = 0 \\ Y_i - \tau, & Z_i = 1; \end{cases}$$

For the strong null hypothesis that arresting and holding suspects decreases the number of arrests by 1 over arresting and releasing them, an analyst would adjust the observed number of subsequent arrests in the treatment group by adding 1 to the number of subsequent arrests; in effect, this adjustment aims to "undo" the effect of treatment. Other, more complex effects could be tested by inferring Y_c values for the treatment group in a similar manner.

To test these strong null hypotheses with inferred Y_c, as before, a model is fit to the control data so that values of the expected response under the control condition, or $\hat{y}_c(\hat{\beta})$ values, can be computed for all units. The inferential part of the PB framework, however, now aims to compare Y_c values to $\hat{y}_c(\hat{\beta})$ values to determine if responses under control inferred according to the null hypothesis are equivalent to those predicted based on covariate information and a model fit to the units that actually received control. If these are equivalent, the null hypothesis holds; if not, there is evidence that the stated null hypothesis does not hold.

As inference for a strong null hypothesis does not focus on a test of a difference in averages, a procedure like a standard t-test is not favorable. Instead, to have correct level tests of a strong null hypothesis, permutation tests are used. In order to perform inference with residuals defined as $e_i = Y_{ci} - \hat{y}_c(\hat{\beta})$ to test a strong null hypothesis, the use of a permutation test requires a permutation distribution against which the observation can be compared. This permutation distribution can be viewed as a collection all possible mean differences in treatment and control residuals, aggregated across blocks, assuming a null hypothesis. Under the assumption of no effect whatsoever (aside from any hypothesized effect that was accounted for in the process of inferring Y_c values for the treatment group) the distributions of the residuals should appear as though treatment were randomly assigned. Thus, to determine the permutation distribution, treatment assignment is randomly permuted, mean differences in treatment and control group residuals are computed, and these differences are collected to determine the reference distribution against which the observed value is compared. This permutation distribution offers the benefit of no additional parametric assumptions for the residuals as well as the correct level (α) for the test. The permutation test also readily incorporates an experiment with a blocking structure; to modify the basic permutation test, treatment assignment is randomly permuted within blocks and aggregating differences in averages across blocks leads to the reference distribution.

Fundamentally, the PB procedure allows the use of familiar regression models to assess the impact of some treatment in a randomized experiment; in contrast to a modeling strategy like regression with a dummy variable, the PB method does not force an analyst to make the same standard regression assumptions. The use of the PB method does, however, assume the sample is large enough so that there is little practical difference between an estimated vector of coefficients, $\hat{\beta}$, and a true underlying vector of coefficients, β_0. It is worth noting how the treatment effect estimated by the PB method and its variance are affected when this assumption does not hold and calculations use a $\hat{\beta}$ that potentially differs noticeably from β_0. In a randomized experiment, the PB method provides a consistent estimate of treatment effect even if the outcome model is misspecified, provided the sample size – particularly the sample size of the control group – is large enough. If the sample size is large enough, the estimated $\hat{\beta}$ should be close to the underlying model for the data described by β_0, as from the theory of regression models, $\hat{\beta}$ can be assumed to be a consistent estimate of β_0. As a result, if $\hat{y}_c(\cdot)$ is a continuous function of β and the sample is large, the estimate of treatment effect based upon $\hat{\beta}$ is consistent for the estimate based upon β_0. In a randomized experiment, the units that fall into the control group can be assumed to be a random sample of all units. Following formulas for regression estimation in survey sampling, a good estimate of the variance of the estimated treatment effect denoted by $\bar{D}$ is given by $\mathrm{Var}(\bar{D}) = {s_e}^2({n_t}^{-1} + {n_c}^{-1})$, where ${s_e}^2$ is the sample variance of the residuals computed according to some $\hat{\beta}$ in the control group, or ${s_e}^2 = \sum_{i=1}^{n_c}(Y_{ci} - \hat{y}_{ci}(\hat{\beta}))/(n_c - 1)$. Variability of $\hat{\beta}$ due to issues with the model and subset of units selected into the control group plays a role in the variability of the resulting residuals, but no component of the formula for the variance of $\bar{D}$ directly describes the variability in $\hat{\beta}$. A situation in which $\hat{\beta}$ is highly variable may result in highly variable residuals and a large value of ${s_e}^2$, but an analyst cannot parse out that part of ${s_e}^2$ due to variability in $\hat{\beta}$. In practice, if the variance of $\bar{D}$ calculated with $\hat{\beta}$ differs greatly from the variance calculated with β_0 due to excessive variability in $\hat{\beta}$, the variance formulas used with the PB method will not call attention to this discrepancy and cannot alert an analyst to potential downstream effects on the resulting inferences. To draw attention to scenarios in which a highly variable $\hat{\beta}$ could inflate the resulting estimate of the variance of the treatment effect, a diagnostic is developed in Section 2.4.

2.3.3 Preliminary Analysis

To analyze the Milwaukee experiment data, an analyst must specify an outcome model and hypothesized treatment effect. In this section, we use the OLS outcome model with all 24 predictors referred to in the previous section. If an analyst wants to test the strong null hypothesis that treatment has no effect whatsoever, Y_c values in the treatment group are inferred to be the observed number of subsequent arrests, following the result of Equation 4.5. For this test, the resulting p-value from the permutation test is 0.46, indicating no evidence of an effect of holding the suspect after arrest instead of releasing him. To test a strong null hypothesis that arresting and holding the suspect decreases the number of subsequent arrests by 1 arrest, the process of inferring Y_c values for the treatment group defines them as $Y_c = Y_t + 1 = Y + 1$, as the observed Y is also the potential outcome under treatment, or Y_t. For the test that treatment decreases subsequent arrests by 1, the resulting p-value from the permutation test is effectively 0, meaning that, after adjustment for the covariates in the outcome model, the data indicate that holding the suspect after arrest rather than releasing him does not reduce the number of subsequent arrests by even a single arrest.

In order to obtain interval estimates of treatment effect, a range of treatment effects is hypothesized, tests are conducted, and the tests are inverted to form confidence intervals. Point estimates and standard errors can also be determined from these series of tests. The results of this analysis are given by Table 2.1.

Table 2.1: Effect of treatment on subsequent arrests using PB with OLS model

Estimate	-0.061
Standard Error	0.082
90% CI	(-0.196, 0.074)

2.4 Diagnostic

While not the main limitation we aim to address in this chapter, the PB method as applied to any comparative study introduces a limitation: the method could inflate the variability of treatment effect estimates in the event of potential overfitting and sampling variability in the estimated coefficients that define the regression model. The details of how such additional variability could arise are detailed in this section. To reduce the effects of this additional variability, the regression model needs to be carefully chosen. In the context of a simple randomized experiment like the Milwaukee experiment, a diagnostic is developed to address this secondary limitation that arises in the use of the PB method with any comparative study. The diagnostic aids in the selection of a regression model, so a model less subject to these issues is used in subsequent analyses.

2.4.1 Motivation for the diagnostic

Consider the application of the Peters-Belson method to a our simple randomized experiment. For this subsection in which the diagnostic is motivated, consider the PB method under two simplifying assumptions. While not necessary for the theory or application of the diagnostic, the assumptions add clarity to the discussion of its purpose. First, assume the function $\mu_\beta(\cdot)$, where $\mathbf{E}(Y|\mathbf{X}) = \mu_\beta(\mathbf{X})$, is a known and correctly specified regression model. Additionally, add the assumption previously incorporated in the preliminary analysis of the experimental data in Section 2.3 that $\mu_\beta(\cdot)$ is a linear function of $\mathbf{X}$. Although this discussion focuses on the linear model context, the ideas behind the diagnostic apply to a broader class of models in which sampling variability and overfitting arise. The fitting of the outcome model presents two challenges to the PB method in this randomized experiment. First, the model must be estimated from the data, so it may differ from the underlying correct model. Also, when fitting a model to data, particularly with the aim of obtaining predictions for other data, the possibility of overfitting exists. In this section, we introduce a diagnostic to determine whether these challenges are great enough with a given model and dataset to impact treatment effect estimates.

In practice, residuals are determined by an estimated $\hat{\beta}$. In a general PB framework in which the covariates $\mathbf{x}$ are considered fixed constants, different values of $\hat{\beta}$ could arise if a different vector for Z or a different vector for Y_c is selected from all possible such vectors. The diagnostic developed in this section is built upon a conditioning set that conditions on the realized values of $Z = z$, so the randomness in $\hat{\beta}$ comes from randomness in Y_c values. With this conditioning set, the units in the control group will always be the same units with the same $\mathbf{x}$ values, but with randomly permuted Y_c values. The variability of $\hat{\beta}$ around the underlying β_0 potentially leads to bias and additional variability in treatment effect estimates that rely upon the values of the residuals, or $e(\hat{\beta})$.

To ground ideas about the impact of overfitting in conjunction with sampling variability, let n be the sample size and p be the number of parameters in β_0, and consider the following situations:

1. $n >> p$, so $\hat{\beta} \equiv \beta_0$ and sampling variability is approximately 0. Thus, in making predictions, in-sample prediction error and out-of-sample prediction error are similar.

2. n is small relative to p, so the model is saturated. Sampling variability in $\hat{\beta}$ is nonzero, and in-sample prediction error is very low relative to out-of-sample prediction error.

Under the assumptions of this subsection in which the model $\mu_\beta(\cdot)$ is assumed correctly specified and linear, in either of the two cases discussed above, there should be no bias in the residuals; consequently, there should be no bias in the resulting estimates of treatment effect (it is worth noting that bias, likely small relative to sampling error, would occur with a nonlinear model). Variability in residuals, however, (which would propagate to variability of corresponding effect estimates based upon residuals) differs across the two specified scenarios. In the first case, as prediction error will be similar for in or out of sample predictions, predictions for treatment and control groups will have similar errors. As the model is assumed to be correctly specified, these errors are likely to be small. Thus, $\mathrm{V}(Y_c - \mu_{\hat{\beta}}(\mathbf{x})) < \mathrm{V}(Y_c)$, or the residual variance will be small relative to the overall variability in Y_c values. In the second case, for the control group, predictions will have little or no error, so $\mathrm{V}(Y_c - \mu_{\hat{\beta}}(\mathbf{x}))$ is approximately zero, but it is likely to be the case that $\mathrm{V}(Y_c - \mu_{\hat{\beta}}(\mathbf{x})) > \mathrm{V}(Y_c)$ in the treatment group as prediction error is high for treatment units. Thus, it is difficult to determine the relationship between $\mathrm{V}(Y_c - \mu_{\hat{\beta}}(\mathbf{x}))$ and $\mathrm{V}(Y_c)$ overall for the second scenario, and in cases like the present study when treatment units outnumber control units, it may be that $\mathrm{V}(Y_c - \mu_{\hat{\beta}}(\mathbf{x})) > \mathrm{V}(Y_c)$ for all units. By this logic, sample estimates should follow a similar pattern. In the first case, as prediction error should be similar for treatment and control units, the sample estimate of the residual variance in the control group, or $s^2(Y_{ci} - \mu_{\hat{\beta}}(\mathbf{x_i}))$, should be a good estimate of the same quantity for all units and also a good estimate for $\sigma^2(Y_{ci} - \mu_{\hat{\beta}}(\mathbf{x_i}))$ for all units. In the second case, $s^2(Y_{ci} - \mu_{\hat{\beta}}(\mathbf{x_i}))$ computed from the control group will underestimate the same quantity for all units as well as its corresponding population quantity $\sigma^2(Y_{ci} - \mu_{\hat{\beta}}(\mathbf{x_i}))$.

Clearly, to have good estimates of treatment effect and corresponding variances based upon residuals, it would be more preferable to be in a situation like that of the first case described. In this section, the diagnostic developed aims to determine if the model and data at hand present a situation more like the preferable first scenario or the more problematic second scenario. While the diagnostic does not directly address the potential for bias (in fact, in the present discussion it is assumed away), the contribution of bias to the MSE of treatment effects is likely to be small, and controlling the variability may lessen some bias contributions. Thus, the diagnostic aims to decide if variance estimation is problematic due to sampling variability in $\hat{\beta}$ and overfitting; if so, adjustments can be made to the model to remedy the problem.

The previous discussion incorporating simplifying model assumptions indicated that if sampling variability in $\hat{\beta}$ is small and overfitting is not problematic, variance estimates from the control group are good estimates of overall variances. In the computations for the diagnostic, estimated variances will come from the control group as that is the group for which information is fully observed for the PB framework. Noting that, the previous discussion leads to our diagnostic, which aims to assess if sampling variability in the residuals due to sampling variability in $\hat{\beta}$ is small relative to overall variability in the residuals. If sampling variability in the residuals is large relative to overall variability, then the model and data at hand may be more like the saturated model of the second scenario. If the sampling variability in the residuals is small relative to overall variability, then sampling variability does not likely have large effects on predictions, and the situation at hand is closer to the more preferable one of the first case discussed. Thus, establishing that sampling variability is small relative to overall variability indicates that variance contributions to treatment effect estimates from overfitting in conjunction with sampling variability in $\hat{\beta}$ should not be cause for concern.

The motivation for the diagnostic in this subsection centers around the case of a simple randomized experiment with simplifying model assumptions and data that do not contain a blocking structure. The diagnostic developed in the following subsections is developed for a general experiment with or without a blocking structure in the design of the study. With the present experimental case study of the Milwaukee data, no blocking is present, so to apply the diagnostic to this study, b, or the number of blocks in the study design, is taken to be $b = 1$. Much of the diagnostic simplifies accordingly.

Preliminaries

For the diagnostic for the PB method with any type of regression model for the outcome model, define common notation. Let n_{bt} and n_{bc} be the number of treatment and control units, respectively, in block b. Further, define $(1/n_b) = [(1/n_{bc}) + (1/n_{bt})]/2$ and $n = \sum_b n_b$. In addition, let $\mathbf{1_b}$ be a vector of ones with $n_{bc} + n_{bt}$ terms. Define a conditioning set C to be

$$C := \{z_{bi}, \text{all } b, i; \text{for all } b, i \text{ and all } \beta, \text{ the order statistics of the control group} \\ (Y_{cbi} - \hat{y}_{cbi}(\beta)) : i = 1, ..., n_{bc}\}$$

By conditioning on the set C, the idea of having randomness in Y_c rather than z is maintained. Additionally, by conditioning on the order statistics, the distribution of a statistic of interest that is a function of Y_c conditional on C is a permutation distribution. As a result, inference falls into a permutation test framework with a random rather than fixed β. The permutation test framework also lends itself to the the consideration of a blocking structure as, in permutation tests, it is useful to consider relatively homogeneous strata to reduce problems due to excessive unit variability.

Write $\beta_0 = \mathrm{E}(\hat{\beta}|C)$ and assume $\hat{\beta}$ is a consistent estimate of β_0. Let $\bar{z}$ be a vector of length $\sum_b n_{tb} + n_{cb}$, $\bar{z} = (\frac{n_{1t}}{n_{1t}+n_{1c}}\mathbf{1_1}, ..., \frac{n_{bt}}{n_{bt}+n_{bc}}\mathbf{1_b})$, and for any β, $e(\beta) = Y_c - \hat{y}_c(\beta)$, where $\hat{y}_c(\beta)$ is a uniformly differentiable mapping (and thus a continuously differentiable mapping over the parameter space) from β to predicted Y_c values. Define the key statistic for the diagnostic as a quantity that examines treatment and control differences in residuals: $T_n(Y,\beta) = \frac{2}{n}(Z - \bar{z})'e(\beta)$.

Asymptotically, conditional on the set C, there should be no statistically discernible difference in the comparison of treatment and control residuals when the residuals are computed using $\hat{\beta}$ and β_0. In our notation, this amounts to saying that the difference between $T_n(Y,\hat{\beta})$ and $T_n(Y,\beta_0)$ scaled by the standard deviation of $T_n(Y,\beta_0)$, or $(T_n(Y,\hat{\beta}) - T_n(Y,\beta_0))/\mathrm{SD}(T_n(Y,\beta_0))$, is zero asymptotically, conditional on C. The diagnostic introduced in this section aims to determine if this characteristic exists in large finite samples for a specified model by focusing on the variability of the scaled difference between $T_n(Y,\hat{\beta})$ and $T_n(Y,\beta_0)$, which, by the previous argument, should be zero asymptotically. In large finite samples, the variability of this scaled difference should be small, provided the specified model is "good", where better models are less subject to problems resulting from overfitting in the presence of sampling variability. To evaluate a given model with the diagnostic, the ratio of the variability of $\sqrt{n}(T_n(Y,\hat{\beta}) - T_n(Y,\beta_0))$ to that of $\sqrt{n}T_n(Y,\beta_0)$ is examined. Conditional on the set C, it can be shown that the variability of $\sqrt{n}T_n(Y,\beta_0)$ converges to some positive finite value, so the denominator of this diagnostic ratio is finite. Throughout the remainder of this section, we will consider the specific case of the linear model in which $T_n(Y,\beta) = \frac{2}{n}(Z - \bar{z})(Y - \mathbf{x}\beta)$, but the ideas presented are generalizable to other types of models.

Sample to sample variability: variability in $T_n(Y,\hat{\beta}) - T_n(Y,\beta_0)$

To describe the variability in $T_n(Y,\hat{\beta}) - T_n(Y,\beta_0)$ conditional on C, use a Taylor series expansion and the intermediate value theorem to write

$$T_n(Y,\hat{\beta}) = T_n(Y,\beta_0) + (\beta - \beta_0)' \bigtriangledown_\beta T_n(Y,\beta)|_{\beta=b}, \tag{2.5}$$

where b is an intermediate value between β_0 and $\hat{\beta}$. Subtract $T_n(Y,\beta_0)$ from both sides:

$$T_n(Y,\hat{\beta}) - T_n(Y,\beta_0) = (\beta - \beta_0)' \bigtriangledown_\beta T_n(Y,\beta)|_{\beta=b}. \tag{2.6}$$

If we define $C_{\hat{\beta}}$ as $\lim_{n\to\infty} \mathrm{V}(\sqrt{n}(\hat{\beta} - \beta_0)|C)$ then by the Central Limit Theorem, as $\mathbf{E}(\hat{\beta}|C) = \beta_0$,

$$\sqrt{n}(\hat{\beta} - \beta_0) \to_d N(0, C_{\hat{\beta}})$$

Use the fact that $\mathbf{E}[(\hat{\beta} - \beta_0)' \bigtriangledown_\beta T_n(Y,\beta)|_{\beta=b}|C] = 0$ to find $\mathrm{V}((\hat{\beta} - \beta_0)' \bigtriangledown_\beta T_n(Y,\beta)|_{\beta=b}|C) = \mathbf{E}[((\hat{\beta} - \beta_0)' \bigtriangledown_\beta T_n(Y,\beta)|_{\beta=b})^2|C]$. In the linear model, $\bigtriangledown_\beta T_n(Y,\beta)|_{\beta=b} = 2n^{-1}\mathbf{x}'(Z - \bar{z})$ for any value of b, and an upper bound can be obtained for this variance:

$$\begin{aligned}
\mathrm{V}((\hat{\beta} - \beta_0)' \bigtriangledown_\beta T_n(Y,\beta)|_{\beta=b}|C) &= \mathbf{E}[(2n^{-1}(\hat{\beta} - \beta_0)'\mathbf{x}'(Z - \bar{z}))^2|C] \\
&= 4n^{-2}[(Z - \bar{z})'\mathbf{x}]\mathbf{E}[(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)'|C][\mathbf{x}'(Z - \bar{z})] \\
&= 4n^{-2}[(Z - \bar{z})'\mathbf{x}]\mathrm{V}(\hat{\beta}|C)[\mathbf{x}'(Z - \bar{z})] \\
&\leq 4n^{-2}[(Z - \bar{z})'\mathbf{x}]\mathrm{V}(\hat{\beta}|z_{bi},\ \text{all } b, i)[\mathbf{x}'(Z - \bar{z})]
\end{aligned}$$

where the final inequality is a result of estimating the variance of $\hat{\beta}$ by conditioning on a smaller conditioning set. By the relationship in Equation 2.6, it follows that $\mathrm{V}((\hat{\beta}-\beta_0)' \bigtriangledown_{\beta} T_n(Y,\beta)|_{\beta=b}|C) = \mathrm{V}((T_n(Y,\hat{\beta}) - T_n(Y,\beta_0))|C)$, which, in the case of a linear model, is dominated by $4n^{-2}[(Z-\bar{z})'\mathbf{x}]\mathrm{V}(\hat{\beta}|z_{bi},\text{ all } b,i)[\mathbf{x}'(Z-\bar{z})]$. $\mathrm{V}(\hat{\beta}|z_{bi},\text{ all } b,i)$ can be estimated robustly from a linear model fit using a sandwich estimator or other fitting methods that estimate standard errors of estimated coefficients more robustly.

Translating notation, we can view $2n^{-1}\mathbf{x}'(Z-\bar{z})$ as the difference between two $p \times 1$ vectors of covariate means where one vector comes from the treatment group and the other from the control group. If we consider the case in which units are randomly assigned to treatment groups, the imbalances, or differences in covariates across treatment and control groups, should follow normal distributions with a mean of zero (see Hansen and Bowers (2008) for a discussion of these issues). As the sample size grows to infinity, these normally distributed imbalances should converge to zero, or, in other words, the variance of the imbalances should converge to zero allowing the imbalances to converge to their mean. Stating earlier arguments in terms of our notation, under random assignment, imbalances in the treatment and control group in terms of average $\mathbf{x}$, which can be written as $2n^{-1}\mathbf{x}'(Z-\bar{z})$, would have a normal distribution with an expectation of zero. With a sample size increasing to infinity, the imbalances would approach zero, so $\sqrt{n}(2n^{-1}\mathbf{x}'(Z-\bar{z})) = O_p(1)$. Thus, under random assignment or a comparable design that makes a study not unlike one with random assignment, which we created with stratification, it follows that $2n^{-1}\mathbf{x}'(Z-\bar{z}) = O_p(n^{-1/2})$, which entails that $2n^{-1}\mathbf{x}'(Z-\bar{z}) \rightarrow 0$.

From earlier results, we know $\hat{\beta}-\beta_0 = O_p(n^{-1/2})$ as $\sqrt{n}(\hat{\beta}-\beta_0) = O_p(1)$, which implies $(\hat{\beta}-\beta_0)'2n^{-1}\mathbf{x}'(Z-\bar{z}) = O_p(n^{-1})$. Thus, it follows that $\sqrt{n}(\hat{\beta}-\beta_0)'2n^{-1}\mathbf{x}'(Z-\bar{z}) = O_p(n^{-1/2})$, or $\sqrt{n}(\hat{\beta}-\beta_0)'2n^{-1}\mathbf{x}'(Z-\bar{z}) \rightarrow 0$. By the continuous mapping theorem,

$$(\sqrt{n}(\hat{\beta}-\beta_0)'2n^{-1}\mathbf{x}'(Z-\bar{z}))^2 = 4n^{-1}[(Z-\bar{z})'\mathbf{x}](\hat{\beta}-\beta_0)^2[\mathbf{x}'(Z-\bar{z})] \to 0$$

By dominated convergence, it follows that

$$\mathbf{E}(4n^{-1}[(Z-\bar{z})'\mathbf{x}](\hat{\beta}-\beta_0)^2[\mathbf{x}'(Z-\bar{z})]|C) = 4n^{-1}[(Z-\bar{z})'\mathbf{x}]\mathbf{E}[(\hat{\beta}-\beta_0)^2|C][\mathbf{x}'(Z-\bar{z})] \to 0$$

Thus, convergence of the sample to sample variance is a necessary consquence:

$$\mathrm{V}(\sqrt{n}(T_n(Y,\hat{\beta})-T_n(Y,\beta_0))|C) \leq 4n^{-1}[(Z-\bar{z})'\mathbf{x}]\mathrm{V}(\hat{\beta}|z_{bi},\ \text{all } b,i)[\mathbf{x}'(Z-\bar{z})] \to 0$$

For the purposes of a diagnostic, we aim to leverage this relationship to show that an estimate of $4n^{-1}[(Z-\bar{z})'\mathbf{x}]\mathrm{V}(\hat{\beta}|z_{bi},\ \text{all } b,i)[\mathbf{x}'(Z-\bar{z})]$ is small in finite samples, where "smallness" is defined as a fraction of the overall variability as measured by the variability in $T_n(Y,\beta_0)$.

Overall variability: variability in $T_n(Y,\beta_0)$

We want to estimate the variability in $T_n(Y,\beta_0)$ conditional on C, and, in order to make comparisons to the variability in $T_n(Y,\hat{\beta})-T_n(Y,\beta_0)$ established previously, we will use an additional scaling factor of $\sqrt{n}$. It is not straightforward to obtain $\mathrm{V}(\sqrt{n}T_n(Y,\beta_0))$ as β_0 is unknown. For purposes of making comparisons in a finite sample based on asymptotic results, if we cannot obtain $\mathrm{V}(\sqrt{n}T_n(Y,\beta_0))$, it should be sufficient to make our finite sample comparison using a quantity with the same asymptotic value as $\mathrm{V}(\sqrt{n}T_n(Y,\beta_0))$.

For some fixed β, define

$$v_n(\beta) = \mathrm{Var}(\sqrt{n}T_n(Y,\beta)|C) = \frac{2\sum_b n_b s^2[\{e_{bi}(\beta):i\}]}{\sum_b n_b} \tag{2.7}$$

where $s^2[\{e_{bi}(\beta) : i\}]$ is the estimated variance of $e(\beta)$ for stratum b.

This equation arises from the following calculation:

$$\begin{aligned}
\mathrm{V}(T_n(Y,\beta)|C) &= \mathrm{V}(2n^{-1}(Z-\bar{z})'e(\beta)|C) \\
&= \frac{\sum_b \mathrm{V}(2(Z_b-\bar{z}_b)'e_b(\beta)|C)}{n^2} \\
&= \frac{\sum_b {n_b}^2(n_t^{-1}+n_c^{-1})s^2[\{e_{bi}(\beta) : i\}|C]}{n^2} \\
&= \frac{\sum_b 2n_b s^2[\{e_{bi}(\beta) : i\}|C]}{n^2}
\end{aligned}$$

As β_0 is a fixed value, $v_n(\beta_0) = \mathrm{V}(\sqrt{n}T_n(Y,\beta_0))$. In an ideal world, we would like to use $v_n(\beta_0)$ in our diagnostic for a finite sample comparison, but this cannot be obtained. Note that with the conditioning set C, the values of $e(\beta)$ are random, as Y_c is a random variable, so for $\hat{\beta}$, $v_n(\hat{\beta}) \neq \sqrt{n}(T_n(Y,\hat{\beta}))$ due to randomness in $\hat{\beta}$. For asymptotic results, define $v(\beta)$ as the limit of $v_n(\beta)$ for a fixed β. Assume $v_n(\beta)$ converges uniformly to $v(\beta)$.

In a linear model, for some β, $e(\beta) = Y - \mathbf{x}\beta$ is a continuous function, so the variance of $e(\beta)$ is continuous provided the stratum sample sizes are greater than 1 (to have a nonzero denominator). Thus, $v_n(\cdot)$ is a continuous function. As uniform convergence preserves continuity, then v is a continuous function.

By the triangle inequality,

$$|v_n(\hat{\beta}) - v(\beta_0)| \leq |v_n(\hat{\beta}) - v(\hat{\beta})| + |v(\hat{\beta}) - v(\beta_0)| \tag{2.8}$$

For any $\epsilon > 0$, it must be established that $\lim_{n\to\infty} P(|v_n(\hat{\beta}) - v(\beta_0)| \geq \epsilon) = 0$. Rewriting the inequality of Equation 2.8 in these terms,

$$\begin{aligned} P(|v_n(\hat{\beta}) - v(\beta_0)| \geq \epsilon) &\leq P(|v_n(\hat{\beta}) - v(\hat{\beta})| + |v(\hat{\beta}) - v(\beta_0)| \geq \epsilon) \\ &\leq P(|v_n(\hat{\beta}) - v(\hat{\beta})| \geq \epsilon/2) + P(|v(\hat{\beta}) - v(\beta_0)| \geq \epsilon/2) \end{aligned}$$

By the continuous mapping theorem and the consistency of $\hat{\beta}$, $P(|v(\hat{\beta}) - v(\beta_0)| \geq \epsilon/2) \to 0$ as $n \to \infty$ for any $\epsilon > 0$. Additionally, as $v_n(\cdot)$ converges uniformly to $v(\cdot)$, for any input value, say a, for every $\epsilon > 0$, there is an N such that for $n \geq N$, $P(|v_n(a) - v(a)| \geq \epsilon/2) = 0$. Thus, $P(|v_n(\hat{\beta}) - v(\hat{\beta})| \geq \epsilon/2) \to 0$ as $n \to \infty$, and it follows that $v_n(\hat{\beta}) \to_p v(\beta_0)$. Thus, as we cannot directly compute $\mathrm{V}(\sqrt{n}T_n(Y, \beta_0)) = v_n(\beta_0)$, as our comparison draws on asymptotic results, it is sufficient to use $v_n(\hat{\beta})$, as they both achieve the same value as $n \to \infty$.

Thus, for the purposes of our diagnostic, we aim to find the square root of the ratio of $4n^{-2}[(Z - \bar{z})'\mathbf{x}]\mathrm{V}(\hat{\beta}|z_{bi},\ \text{all } b, i)[\mathbf{x}'(Z - \bar{z})]$ to $n^{-2}\sum_b 2n_b s^2[\{e_{bi}(\beta) : i\}|C]$ to demonstrate that the former is small relative to the latter. After simplification, the diagnostic relies on assessing the magnitude of

$$\sqrt{\frac{2[(Z - \bar{z})'\mathbf{x}]\mathrm{V}(\hat{\beta}|z_{bi},\ \text{all b, i})[\mathbf{x}'(Z - \bar{z})]}{\sum_b n_b s^2[\{e_{bi}(\beta) : i\}|C]}} \qquad (2.9)$$

A rule of thumb might be to aim for a ratio of less than 0.2.

2.4.2 Application of the diagnostic to Milwaukee experiment

If an ordinary linear model is used to model the relationship between subsequent arrests and 24 mostly factor variables that describe the suspect and victim's demographic information and past history for the 382 control suspects, the ratio of interest is 0.47 (as show in Table 2.2), well above our rule of thumb ratio. By our diagnostic, choosing this outcome model may lead to highly variable treatment effect estimates.

Table 2.2: Estimates of variabilities and resulting ratio for the computation of the diagnostic under different model specifications for the experimental case study

	Overall Standard Error	Standard Error of Difference	Ratio
	estimated by $v_n(\hat{\beta})^{1/2}$	$(T_n(Y,\hat{\beta}) - T_n(Y,\beta_0))$	Eq. 2.9
1. Linear model	1.51	0.72	0.47
2. Bayesian linear model	1.51	0.66	0.44
3. Bayesian linear model with scale parameter of 0.1	1.49	0.44	0.29
4. Bayesian linear model with scale parameter of 0.03	1.54	0.26	0.17

Methods such as penalized regression and Bayesian regression have the property that variance estimates better account for issues of collinearity and unusual observations than standard linear model estimation techniques. Gelman (2004) notes that problems due to overfitting are of less concern with "reasonable" prior distributions, and this quality is attractive given the aim of the diagnostic. With a Bayesian linear model, the additional information from any prior other than a noninformative prior can act like additional data to help with potential collinearity. In addition, a Bayesian model with a prior for the coefficients of a t-distribution with a small degrees of freedom parameter can effectively downweight unusual observations in the data, not allowing these points to drive the model fit.

In line with these ideas, we choose a Bayesian linear model with a t-prior with four degrees of freedom for all coefficients, where four is chosen to be small enough to allow for thick tails without introducing additional difficulties (for example, choosing $df \leq 2$ allows for infinite mean and variance). With this model, we find the variability of $T_n(Y, \beta_0)$ estimated by $v_n(\hat{\beta})$ is the same as that obtained from the non-Bayesian linear model, but the Bayesian linear model reduces variability in the difference $T_n(Y, \hat{\beta}) - T_n(Y, \beta_0)$, reducing the overall diagnostic criterion ratio from 0.47 to 0.44. Specifying a smaller value for the scale parameter of the t-prior for all coefficients while keeping the mean of zero reduces the variability of $T_n(Y, \hat{\beta}) - T_n(Y, \beta_0)$ further, but with a consequence that resulting parameter estimates will become less stable. With careful consideration of this consequence, further adjustments to the prior could be made or variables could be removed from the model to allow the variance ratio to decrease even more, if desired, to meet or exceed the diagnostic rule of thumb. By reducing the scale parameter, the estimates of the diagnostic criterion approach and fall below the rule of thumb ratio of 0.2. To incorporate the diagnostic into the analysis of this experimental case study, an analyst may choose to use the fourth model in Table 2.2 in the analyses in Section 2.3. When the model chosen by the diagnostic is used, the effect estimate decreases substantially, though it is still not significant, and the standard error increases slightly as compared to the estimates in Table 2.1.

Table 2.3: Effect of treatment on subsequent arrests: comparing OLS model and model chosen by the diagnostic

	OLS model	Bayesian model chosen by diagnostic
Estimate	-0.061	-0.039
Standard Error	0.082	0.083
90% CI	(-0.196, 0.074)	(-0.176, 0.097)

2.5 Using the PB method to analyze observational data

When the PB method is applied to an experimental study in which distributions of covariates are balance across treatment groups, the only limitation is that of additional variability managed by the diagnostic. When the method is applied to observational data or an experiment without such covariate balance, adjustments are required due to an additional limitation.

2.5.1 The central limitation of the PB method: Problematic extrapolation of the outcome model

While a straightforward way of assessing disparities, the PB method's assumption of ignorability can be called into question when the framework is applied to an observational study or experiments that similarity of covariate distributions across treatment groups. This opens the Peters-Belson method to the critique of extrapolation of the fitted outcome model. When fitting a regression surface to one group for use in estimating values for another group, an analyst may not be aware of a lack of overlap in the covariate distributions across the two groups. In the example of education-based wage disparities, it may be more typical for state college graduates in a certain field to be younger and have less experience. Perhaps for workers with less experience, the functional form of the regression model differs, but fitting and extrapolating a regression model will not call attention to this disparity.

Simulation studies, detailed in Appendix 2.9.1, offer information about the performance of the Peters-Belson method when the covariate overlap between comparison groups in an observational study is not what it would be in a properly randomized experiment. These simulation studies illustrate that poor estimates of the level and lower levels of power are obtained when the regression model in PB must extrapolate across disparate treatment and control groups to a greater extent. Solutions to the problem of extrapolation in the Peters-Belson method have been proposed in the literature. One proposed solution (Ñopo, 2008) is to create "synthetic" observations, or artificially generated counterfactual observations, without a regression model. Instead, these "synthetic controls" are weighted combinations of several existing control observations. The method also imposes a restriction of estimation to a region of common support, a region of the covariate space in which there are both treatment and control observations.

Addressing the problem of extrapolation more generally, one could focus on covariate overlap across groups. Propensity scores (Rosenbaum and Rubin, 1983) were designed to help an analyst to be aware of and avoid extrapolation that could occur with a regression model. The propensity score is the conditional probability of receiving treatment given a unit's observed covariates. If we match or stratify on the true propensity score, then conditional on this true propensity score, treated and untreated units should have the same distributions of covariates within subclasses Rosenbaum and Rubin (1983). When fitting a standard regression model to the control group in the application of the Peters-Belson method, an analyst can have a list of predictors in a model and not readily know how the groups compare on these predictors. By reducing the data down to an easily plotted one dimensional score, propensity scores afford an analyst the opportunity to clearly see if covariate distributions do not overlap. Thus, propensity score estimation makes it apparent if any further analyses such as the estimation of treatment effect would be subject to errors of extrapolation. In addition, by grouping units with similar propensity scores into matched sets or strata and only making comparisons between treatment and control units within these groups, an analyst avoids making comparisons between units that were not comparable before treatment was applied.

To take advantage of these properties of propensity scores, another solution proposed to reduce problems due to extrapolation in the PB model specifically is to use inverse propensity score weighting in the outcome model (DiNardo, 2002). Weighting methods can have problems in a practical context when the fitted propensity scores, probabilities, are very close to 1 or 0. In simulation results, Frölich (2004) finds that estimates obtained from pair matching tend to have smaller values of MSE than those obtained from weighted regression models. Additionally, Hansen (2004) finds that optimal full matching, where restrictions on the number of treated and control units in a matched set are relaxed and not set to be exactly 1 to 1 ratios, typically outperforms pair matching.

As alluded to in Section 2.2.1, to lessen problems of extrapolation for the fitted outcome model, our proposed modification to the PB method begins by making any comparative study look in terms of covariate distributions like a block-randomized study by forming optimally matched propensity score sets. Analysis proceeds by incorporating this blocking structure into the analysis as though the blocks were a feature of experimental design. Thus, the method proceeds as in the analysis of a randomized experiment, but the permutation test restricts permutation of the treatment assignment to matched sets.

2.5.2 Observational case study: Chicago Community Adult Health Study

The benefits of supplementing a PB analysis of a disparity with propensity scores is illustrated on data from the Chicago Community Adult Health Study (CCAHS). The CCAHS was designed with the objective of increasing the understanding of the role residential context, in conjunction with individual and household factors, plays in a variety of health outcomes. Data were collected between May, 2001 and March, 2003, during which 3105 Chicago adults aged 18 and older were interviewed and some direct physical health measures were made including body size measurements, weight, and blood pressure. The 3105 adults were sampled from 343 neighborhood clusters (NCs), which were previously defined by the Project on Human Development in Chicago Neighborhoods (PHDCN) conducted by Sampson et al. (1997). People in 80 focal areas previously defined by PHDCN were sampled at twice the rate of people in nonfocal areas, and one individual was interviewed per household.

In the CCAHS data, there are many variables along which a researcher could analyze a disparity between a treatment group (individuals with a certain attribute) and a control group (individuals absent the certain attribute) due to the nature of the city of Chicago. Measures of various composite neighborhood-level aspects, such as affluence, disadvantage and urbanicity were constructed with a factor analysis of measured neighborhood-level variables.

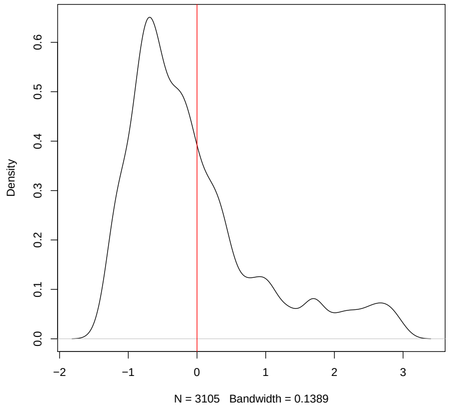
For this case study, we aim to examine if the impact of neighborhood affluence on blood pressure. Residents of high affluence neighborhoods will be considered the control group and those in low affluence neighborhoods will be considered the treatment group, making the treatment, in effect, the lack of affluence. It may seem odd to have the lack of a trait be the treatment condition, but, in this case, it is justifiable due to the nature of the effects of affluence. A control condition should be relatively homogeneous in its application; it stands to reason that neighborhoods on the upper end of the affluence scale are more homogeneous as the effects on health outcomes of, for example, an “upper middle class” or “wealthy” neighborhood should not differ much. With a less affluent neighborhood, however, there could be much more variability in the affluence of these neighborhoods and its particular impact on the lives of the residents of these neighborhoods. For this variability, it makes sense to consider the lack of affluence to be the treatment condition. While we begin by considering a binary treatment of low neighborhood affluence and high neighborhood affluence, Section 2.7 will present ways of modeling the heterogeneity in the “dosage” of treatment.

As the effects of neighborhood affluence on blood pressure likely depend on other individual and neighborhood level control variables, attempts to estimate the effects of affluence will control for demographic factors, health characteristics (including eating, sleeping, smoking, drinking, and exercise behaviors), and other neighborhood-level information.

2.5.3 Fitted propensity score with the CCAHS data

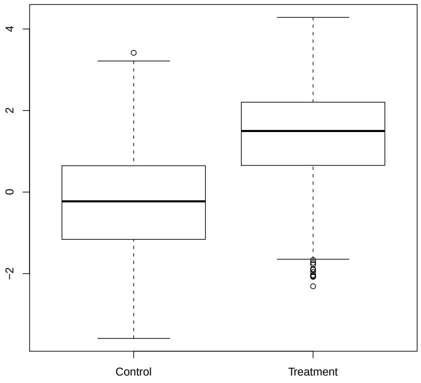
The modification to the standard PB procedure discussed in this paper requires a propensity score in order to create a blocking structure to mimic a block-randomized experiment. We aim to fit a propensity score for neighborhood affluence as the treatment variable, which is a continuous measure in the CCAHS data. Standard propensity scores are built around a dichotomous treatment variable, so a division is made in the standardized version of the continuous neighborhood affluence measure to obtain a binary treatment variable. While it may be more clear to have a treatment variable that is naturally binary, a binary variable can be created in this manner, preferably with guidance by a subject matter expert to make a meaningful division. Figure 2.5.3 shows a density plot of the affluence measure and notes where this cutoff was made in order to create two groups defined by their exposure to neighborhood affluence.

Figure 2.1: Measure of neighborhood affluence and cutoff point to create two treatment groups



The propensity score is estimated by a logistic regression model; treatment assignment is modeled as a function of many person-level and neighborhood-level pretreatment variables such as age, sex, race, parental status, marital status, immigration status, as well as the other neighborhood-level composite factors aside from affluence. Before matching, a balance test in the form of those described by Hansen and Bowers (2008) find that the treatment group, or the group that does not live in an affluent neighborhood, and the control group, those who live in a more affluent neighborhood, have a statistically significant imbalance overall in terms of the distributions of all measured variables included in the propensity score model. This finding is echoed by the boxplots in Figure 2.5.3. The fitted values of this propensity score will be used to create matched sets and strata in subsequent discussions.

Figure 2.2: Distribution of the linear propensity score in the control (high affluence) and treatment (low affluence) groups



2.5.4 Application of the diagnostic to CCAHS data

While the diagnostic in Section 2.4.1 was developed for a randomized experiment, it can also be applied to the use of the PB method when analyzing a comparative study incorporating propensity score groupings. If units are placed into groups based on their propensity scores in such a way that there are no longer statistically significant covariate imbalances across the treatment and control groups within strata, then within a stratum, covariate imbalances should not be very different from the normally distributed imbalances that would have been seen under randomization. As a result, within these created blocks, our intuitions can be guided by those formed considering the case of random assignment because our study design looks in terms of balance in covariate distributions as it would appear under random assignment. For purposes of the diagnostic, if a grouping structure is necessary to approximate a block-randomized experiment, propensity score strata indexed by b are used rather than propensity score matched sets. To compute variances of residuals within optimally matched sets, due to the small number of units in each set, potential responses under control would need to be hypothesized for treatment units in order to have more than one unit for which a variance could be computed. By using larger "blocks", or groupings of units, in the form of propensity score strata, the dependence on a specific hypothesis is minimized for the purposes of the model diagnostic. Using balance test routines developed by Hansen and Bowers (2008), the number of strata is selected to be the smallest number of strata such that there is not a statistically significant imbalance, which will be important for later results.

When the diagnostic is applied to the CCAHS data, the best choice for the number of strata is to divide the data into nine strata on the basis of the propensity score. When nine strata are chosen, there are no significant covariate imbalances within strata. Thus, if the stratification is taken into account, the design of the study of the CCAHS data makes covariate distributions within strata similar to those of a block-randomized experiment. This allows the arguments made in the outline of the diagnostic to be used. The results for systolic and diastolic blood pressure were nearly identical, so, for simplicity, the diagnostic is discussed in the context of the outcome of systolic blood pressure and several outcome models.

If an ordinary linear model is used to model the relationship between systolic blood pressure and 17 demographic, health-related, and neighborhood-level variables, the ratio of interest is 0.2 (as show in Table 2.4), which falls at our rule of thumb ratio. Modeling the data with such a linear model, however, neglects to account for clustering effects by neighborhood; for this reason, a multilevel model with a random effect for neighborhood may be more appropriate. The fit of the multilevel model reduces the overall variability, or the estimated $v_n(\hat{\beta})$, but it does not decrease the variability in $T_n(Y, \hat{\beta}) - T_n(Y, \beta_0)$, leading to a larger ratio of 0.22.

Table 2.4: Estimates of variabilities and resulting ratio for the computation of the diagnostic under different model specifications for the observational case study

	Overall Standard Error estimated by $v_n(\hat{\beta})^{1/2}$	Standard Error of Difference $(T_n(Y, \hat{\beta}) - T_n(Y, \beta_0))$	Ratio Eq. 2.9
1. Linear model	27.63	5.40	0.20
2. Multilevel model	25.05	5.40	0.22
3. Bayesian linear model	27.63	5.30	0.19
4. Bayesian linear model with scale parameter of 0.1	27.53	4.69	0.17

Following the logic described with the application of the diagnostic to the experimental case study, a Bayesian linear model with a t prior with 4 degrees of freedom is used for all coefficients. As with the Milwaukee experiment data, we find the variability of $T_n(Y, \beta_0)$ estimated by $v_n(\hat{\beta})$ is the same as that obtained from the non-Bayesian OLS model, but the Bayesian linear model reduces variability in the difference $T_n(Y, \hat{\beta}) - T_n(Y, \beta_0)$. As a result, the Bayesian linear model with standard scale results in a diagnostic ratio of 0.19, the smallest obtained so far. Adjusting the scale parameter as was done with the experimental case study, the sample to sample variability decreases, leading to a corresponding decrease in the diagnostic criterion. After looking at these results, one may choose to use a Bayesian multilevel model, as this might combine the abilities of these models to separately reduce the overall variability and variability of the difference in T_n values. For purposes of our illustrative analysis, we will use the Bayesian linear model with the $t(4)$-prior and adjusted scale parameter of 0.1, as a primary goal is to keep the variance contribution due to sampling variability small.

2.5.5 Assessing the effects of a binary affluence measure on blood pressure

Without any additional covariate adjustment or adjustments for the similarity of comparison groups, the effects of a lack of affluence on blood pressure can be assessed by a standard t-test across the two levels of affluence. This analysis yields a difference in means of 3.2 points and a p-value of 0.0001, so there appears to be a highly significant relationship between the lack of affluence and blood pressure. Going further, and controlling for demographic, health-related, and other neighborhood-level covariate information with a linear regression model, the effect of a lack of affluence on blood pressure appears to be an increase of 1.6 points, which is significant at $\alpha = .10$. While both of these methods obtain significant results, neither accounts for the disparity in covariate distributions across treatment and control groups demonstrated by the boxplots in Figure 2.5.3. ANOVA and OLS regression methods may not be the best methods of analysis, but they are commonly used to analyze data like the CCAHS data. The present analysis applying the PB method with and without the incorporation of propensity matched sets shows the need to not only incorporate covariate adjustment, but also account for the disparity in covariate distributions across treatment and control groups.

Based on a strong null hypothesis Y_c values are inferred for the treatment group units as discussed in Section 2.3.2. Using the diagnostic of Section 2.4.1 to guide the outcome model chosen for the PB analysis, a Bayesian linear model with a t-prior with $df = 4$ and a reduced scale for all coefficients is used for the analysis for the responses of both systolic and diastolic blood pressure. After an outcome model is chosen, predicted $\hat{y}_c$ values are computed to obtain expected blood pressure values for all individuals supposing they lived in a high affluence, or control, neighborhood. With the Y_c values and $\hat{y}_c$ values for all units, residuals can be obtained. A permutation test is conducted to decide whether or not, after adjusting the Y_c values for the particular strong null hypothesis, it can be determined that there was no (additional) effect of treatment whatsoever. In order to obtain point estimates, standard errors, and 90% confidence intervals for both systolic and diastolic blood pressure, a range of treatment effects is hypothesized in a series of strong null hypotheses for both systolic and diastolic blood pressures. PB analysis is performed and the resulting hypothesis tests are inverted to form confidence intervals. The resulting point estimates, standard errors, and confidence intervals are presented in Table 2.5 both with and without adjustment for propensity matched sets.

Table 2.5: Estimated treatment effect of affluence on both systolic and diastolic blood pressure

	SYSTOLIC BP		DIASTOLIC BP	
	without matching	with matching	without matching	with matching
point estimate	1.10	0.69	0.77	0.87
standard error	0.75	0.87	0.46	0.52
90% Conf Int.	(-0.13, 2.34)	(-0.73, 2.12)	(0.02, 1.52)	(0.01, 1.73)

Table 2.5 illustrates that the incorporation of propensity matched sets may inflate standard errors, as would be expected with a matching procedure, but it does not shift treatment effect estimates in a given direction (e.g. toward zero or away from zero); incorporating the matched sets resulted in a lower estimated treatment effect for systolic blood pressure, but a higher estimated treatment effect for diastolic blood pressure. With systolic blood pressure, however, the estimated treatment effects are not significant. A lack of affluence may seem to increase both types of blood pressure

with the t-test and OLS regression results, but with the PB method with propensity matched sets only one of the two results is statistically significant at $\alpha = .10$.

2.6 Simulation study of diagnostic performance

By selecting a model that obeys the diagnostic presented in Section 2.4.1, we aim to control the variability in treatment effect estimates that could arise from sampling variability in $\hat{\beta}$ and/or the presence of overfitting to achieve sharper estimates of the treatment effect. In this section, a simulation study examines to what extent fixing problems with the variance estimation by imposing the diagnostic leads to improvements in the estimation of the treatment effect. The performance of the diagnostic is assessed in a randomized setting in this simulation as the primary motivation of the diagnostic is to ensure that when used in the setting of an experiment, it improves subsequent treatment effect estimation. We conjecture that controlling the variance by choosing a model that meets the diagnostic criteria would mainly translate to improvements in the variance of the estimated treatment effect estimates.

2.6.1 Methodology

This simulation study uses the CCAHS data and permutes data within propensity score strata, incorporating the assumption that the study design is such that some values within a stratum are as though they were randomly assigned to units. For the simulation, the effect of neighborhood affluence on the outcome of systolic blood pressure will be examined. The simulation is structured so that the true treatment effect is no effect. The design of the simulation respects the conditioning set C defined in Section 2.4.1 by permuting values of the residuals within propensity score strata, which maintains the order statistics of the observed residuals. In addition, to respect the other part of C, z will be held fixed, so the same units will fall into the treatment and control group in every iteration of the simulation. The simulation study assesses the performance of the diagnostic by considering three different scale parameters for the $t(4)$-prior and examining the bias and the variance of the estimated treatment effect that corresponds to the model for which the diagnostic criteria was met.

First, the data are prepared for the simulation according to the following process:

1. Y_c values are inferred for all units under the null hypothesis of no effect. For an observed response Y, $Y_c = Y$ if $z = 0$ and $Y_c = Y - \tau$ if $z = 1$. τ is taken to be the treatment effect estimated in Section 2.5.5 to uphold the null of no effect, so $\tau = 0.69$.
2. The outcome model of a Bayesian linear model with a t-prior with 4 degrees of freedom and a reduced scale parameter of 0.1 for all coefficients is fit to the control group, and predicted values, or $\hat{y}(\hat{\beta})$, are estimated.
3. From the $\hat{y}(\hat{\beta})$, residuals are computed such that $e_i = Y_{ci} - \hat{y}(\hat{\beta})$ for all units.

Then, the simulation randomly permutes residuals within strata and performs analysis according to the following process, storing estimates of treatment effect and values of the diagnostic ratio in each iteration:

1. Residuals are randomly permuted within propensity score strata to obtain a vector of permuted residuals e_p, and values of ${Y_c}^*$ are constructed: ${Y_c}^* = e_p + \hat{y}(\hat{\beta})$.
2. Three outcome models, all Bayesian linear regression models with a t-prior of 4 degrees of freedom for all coefficients, are fit, allowing the scale parameter to be the default value of 2, the reduced value of 0.1, and a further reduced value of 0.05. Define these models in terms of a value $\hat{\beta}_j^*$, where $j = 1, 2, 3$.
3. For each of the three outcome models, residuals e^* are computed: $e^* = {Y_c}^* -$

$\hat{y}(\hat{\beta}_j^*)$.

4. Analysis proceeds as in Section 2.5.5 and permutation test estimates of treatment effect are obtained for each of the three models. These values, which can be called $\hat{\tau}_1$, $\hat{\tau}_2$, and $\hat{\tau}_3$ are stored for each run of the simulation.

5. Additionally, for each of the three outcome models, the value of the diagnostic ratio is computed and stored. These values will be referred to as d_1, d_2, and d_3.

6. For comparison purposes, a simple estimate of treatment effect is also computed. This estimate is computed by taking a weighted average of the difference in the Y_c^* values in the treatment and control group across propensity strata.

For the estimated treatment effects from the simulation, bias and variance are computed to evaluate the performance of the diagnostic. As stated, the value of the treatment effect should be zero under this design, so bias can be determined as the average deviation from zero across all estimated treatment effects. The variance of the average treatment effect is estimated by the variance in the simulated treatment effects.

2.6.2 Results

Table 2.6 presents the results of the simulation study for three models with different scale parameters for the prior distribution. In line with previous conjectures, bias in the treatment effect estimates is a very small contribution to the MSE of the treatment effect estimates; variance provides a much larger contribution, so it is sensible that the diagnostic seeks to control this component. As compared to the weighted average estimate, the PB method performs better in terms of bias, variance, and, as a consequence, MSE, regardless of prior chosen. Clearly, adjustments for covariates and overlap in covariate distributions is beneficial for the estimation of treatment effects.

Examining the results by model, achieving a smaller diagnostic value by tightening the scale on the prior of the coefficients results in greater control of the variance but also an increase in bias. This bias-variance tradeoff is seen in the change in the MSE: moving from a scale parameter of 0.1 to 0.05 changes the diagnostic value and variance, but it barely changes the MSE. In this case, both diagnostic values are below 0.2, so our rule of thumb would have declared either to be an acceptable model, and, in terms of MSE, they are comparable. Studying the two models that straddle the 0.2 threshold for the diagnostic, the model with the prior with the larger scale has better performance in terms of bias, but the smaller variance for the model with a tighter prior distribution leads to smaller MSE. Thus, controlling the variance component by enforcing the diagnostic criterion seems to be beneficial in terms of variance and MSE, as we conjectured. From this simulation, it appears that the diagnostic performs as desired in controlling potential problems with estimating treatment effects and corresponding variances.

Table 2.6: Assessing the diagnostic via simulation in which all prior distributions for the PB models follow a $t(4)$-distribution

	Avg. Diagnostic Value	Bias	Variance	MSE
PB Model with scale=2	0.211	0.019	0.897	0.897
PB Model with scale=0.1	0.190	-0.025	0.888	0.889
PB Model with scale=0.05	0.163	-0.090	0.880	0.888
Weighted average estimate		0.142	1.000	1.020

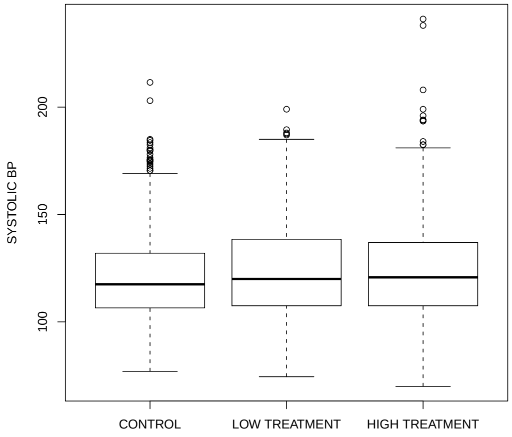
2.7 Extending the PB and propensity score framework to doses of treatment

When assessing the effect of some treatment on an outcome of interest, it is frequently the case that the effect of the treatment varies by the amount of treatment received. In the case of the CCAHS data, it is reasonable to imagine that the effects of living in a disadvantaged neighborhood may differ by the extent of the disadvantage. Perhaps the effects of severely disadvantaged neighborhoods differ more than those of more moderately disadvantaged neighborhoods. The Peters-Belson method discussed in Section 2.5.5 considers the effect of treatment to be homogeneous. In this section, a method to assess a hypothesis of these “dosage” effects is discussed.

As with the decision between treatment and control groups, with the CCAHS data, a the continuous outcome measure of neighborhood affluence must be cut to define doses of treatment. To make this decision, several other covariates are plotted against neighborhood affluence. When these plots are made, there is repeatedly a break in this standardized neighborhood affluence measure around -0.45, so this is chosen to separate a higher level of treatment from a lower level of treatment. As a simple analysis that does not consider additional covariate adjustment, standard one-way ANOVA procedures with post-hoc tests can be used to determine if systolic blood pressure differs across groups. This preliminary analysis shows an overall significant difference but no difference between high and low levels of treatment and a significant difference in systolic blood pressure between both the group with a high level of treatment and the control group as well as the group with a low level of treatment and the control group. These differences are readily seen by the boxplot of doses presented in Figure 2.3.

Although the preliminary analysis and boxplot may not indicate a need for the use of a “dose” effect of treatment, an analyst should confirm that these preliminary analyses do not present a dose effect due to the lack of adjustment for confounders. To make this confirmation, the Peters-Belson method incorporating propensity matched sets can be applied considering doses of treatment. Our modifications to the Peters-Belson procedure have two key parts: optimally matched propensity score sets and

Figure 2.3: Systolic Blood pressure different levels of treatment



an outcome model. Both of these components must be adjusted to accommodate the consideration of doses.

To adjust the matched sets for the consideration of doses, two propensity scores are estimated in addition to the standard propensity score with a single treatment group and control group: one additional score describes the probability of being in the highest dose of treatment relative to the control group, and the other describes the same probability in terms of the lower dose of treatment relative to the control group. In standard optimal matching on propensity scores, a distance between any two units is computed as the pairwise difference in the linear predictor of the estimated propensity score. The pairwise differences are placed in a distance matrix of dimension $n_t \times n_c$ where n_t is the number of treatment units and n_c is the number of control units. This distance matrix determines the arrangement of units within sets. With the two propensity scores for different levels of treatment, separate distance matrices of dimension $n_{dose} \times n_c$ (where n_{dose} is the number of treatment units in that particular dose group) are created for each dose of treatment. These matrices are formed by determining the Mahalanobis distance of the linear predictor of the estimated overall propensity score and the linear predictor of the dose-specific propensity score. Matched sets are formed by applying an optimal matching algorithm to the $n_t \times n_c$ aggregate of the dose-specific matrices.

Following the determination of matched sets, an outcome model from which the residuals are computed must be selected. The general form of the outcome model can follow the choices made in Sections 2.5.5 and 2.4.1, but it must be decided how the dose effect will be incorporated into the analysis. If the analysis should go beyond a simple treatment effect to a dose effect, should the dose effect be linear? Should the dose effect be quadratic? In order to make this determination, a comparison of nested models incorporating these different treatment effects will be made. The nested models will consider the same covariates as the outcome model used in the basic Peters-Belson analysis, but, to test the models, they incorporate a fixed effect for matched set and the specified treatment effect. The effects considered are no treatment effect of any kind, a simple treatment effect with no dosage effect, a linear dosage effect, and a quadratic dosage effect. When these nested models are compared, it appears that an analyst should indeed use a simple treatment effect (p-value $=$ 0.004), but nothing beyond a simple treatment effect (a linear dose effect results in a p-value $= 0.15$). This confirms the findings of the preliminary analysis that there is no evident dose effect of neighborhood affluence on blood pressure. If, however, there were dosage effects discovered, the outcome model selected by the above procedure could be used to modify the standard Peters-Belson analysis accordingly in order to obtain dose-specific estimates of treatment effect.

2.8 Summary

The Peters-Belson method was developed as a technique to assess the effect of a disparity on a particular outcome of interest. It offers a clear framework to describe these effects with standard regression models in the case of a randomized experiment. Although the framework is useful, the extrapolation of regression models required by the technique poses a difficulty when the method is applied to an observational study in which treatment was not randomly assigned. If treatment were not randomly assigned, the covariate distributions in the treatment and control group could differ substantially, and this presents problems when using a model fit to the control units to predict values for treated units.

This chapter presents a solution to reduce problems due to extrapolation by incorporating optimally matched sets on the basis of the propensity score. Performing the analysis within propensity matched sets improves the estimates of treatment effects by ensuring that the extrapolation of the regression model is performed for units with similar covariate distributions, aside from treatment status. While the use of propensity matched sets does increase standard errors of estimates, the protections against extrapolation have the potential to greatly reduce bias. In addition to potential problems due to extrapolation, the outcome model required for analysis can be subject to additional problems in the case of overfitting, regardless of the type of study to which the Peters-Belson method is applied. With greater degrees of overfitting, additional contributions to the variability of the treatment effect estimate could be problematic. The diagnostic presented in Section 2.4.1 offers a guideline to determine when an analyst should consider adjusting a particular outcome model, and this guideline seems to be useful as the associated simulation in Section 2.6 indicates. This chapter illuminates the central drawbacks of the Peters-Belson method, especially when it is applied to an observational study, and presents solutions and assessments of these drawbacks; thus, this modified Peters-Belson method and diagnostic can be applied to any type of study to assess the effect of some treatment of interest.

2.9 Appendix

2.9.1 Simulation studies to illustrate problems with extrapolation

To craft simulations illustrate the difficulties when the PB regression model must extrapolate, a study of SAT coaching by (Powers and Rock, 1999) is used. In this study, the treatment is taken to be SAT coaching and the outcome of interest is the math post test score. From the control group only, hypothetical treatment and control groups will be selected by a biased randomization procedure and called the pseudo-treatment and pseudo-control group. In order to randomly allocate units to the pseudo-treatment and pseudo-control groups, a propensity score is estimated. For the simulation, many propensity models are considered and evaluated on the difference from true propensity score model (as determined by the fit to the observed data) and the expected separation in the treatment and control groups on the basis of the fitted propensity score. The four propensity score models that represent the combinations of the highest and lowest values of these two measures are used in the simulation as four different versions of the true propensity model. After restricting our attention to a randomly chosen subset of the control group data (800 of the approximately 3500 control group subjects), each version of the true propensity score is used as the probability of selection into the pseudo-treatment group. By restricting our attention to the control group, none of these units actually received the treatment under study, which, in this case, is SAT coaching. Thus, it follows that the true effect of treatment across both the pseudo-treatment and pseudo-control groups should be zero, and the null hypothesis for the simulation study is the hypothesis that treatment has no effect on the outcome. For each iteration of the simulation, a new pseudo-treatment and pseudo-control group are selected. Within each iteration, the Peters-Belson treatment effect comparing the pseudo-treatment and pseudo-control groups is estimated along with a corresponding standard error. Standard errors are estimated in two ways: the standard method as incorporated into the analysis by Gastwirth and Greenhouse (1995) and a permutation test standard error (the ideas of which are discussed in, among other sources, Hansen and Bowers (2008)).

Under these specifications, the simulation aims to assess the level and power of the test of the null hypothesis of no effect. To assess the level, the proportion of times the null hypothesis of no effect is rejected when it should be true is recorded for each version of the propensity score truth. The nominal level is 5%. To assess power, a treatment effect of 5 points in either direction is assumed, and the outcome values of the pseudo-treatment group are adjusted accordingly to allow power to be tracked and averaged over the assumed positive and negative treatment effect. In Table 2.7, it is clear for all propensity score "truths" the model-based standard errors undershoot the nominal level, while having very low levels of power. In contrast, the permutation-based standard errors inflate the level over the nominal level, while yielding better power results. Neither formulation, in terms of the level and power, seems especially promising for this data set which has a lack of covariate overlap across comparison groups. The different standard errors used are discussed in the next section.

Table 2.7: Assessing level and power of hypothesis tests based on the Peters-Belson method via simulation

Model	Difference from true propensity score	Expected treatment-control group separation	SE Version	Level (%)	Power (%)
1	high	high	Model	2.8	9
			Permutation	5.7	17
2	low	low	Model	5.4	15
			Permutation	10.3	19
3	high	low	Model	4.3	12
			Permutation	7.8	19
4	low	high	Model	2.8	11
			Permutation	7.1	17

2.9.2 A tale of two standard error estimates

For the simulations discussed in Section 2.9.1, two standard error estimates were used. Regression-based standard errors of Gastwirth and Greenhouse (1995) and the permutation standard errors, which are like those discussed in Hansen and Bowers (2008), led to different results for coverage and power. The permutation-based standard errors are consistently smaller than the model-based standard errors. To illustrate why this is the case, one can examine the formulas and sample calculations based on one run of the simulation study.

If Z defines the vector of treatment assignment, e defines the vector of residuals, and $\mathbf{X_c}$ and $\mathbf{X_t}$ are the covariate matrices for the control and treatment groups, respectively, the treatment effect $\bar{D}(e)$ can be computed as in Equation 2.2. The permutation-based estimate of the null variance is defined as

$$\text{Var}_P(\bar{D}(e)) = {s_e}^2({n_t}^{-1} + {n_c}^{-1}) \tag{2.10}$$

while the model-based estimate of the null variance is defined as

(2.11)

$$\text{Var}_M(\bar{D}(e)) = {s_{e|Z=0}}^2({n_t}^{-1} + {n_c}^{-1} + (\bar{X}_t - \bar{X}_c)'[(X_c - \bar{X}_c'\mathbf{1})'(X_c - \bar{X}_c'\mathbf{1})]^{-1}(\bar{X}_t - \bar{X}_c)).$$

In one run of the simulation with the SAT data and a simulated sample size of $n = 800$ of which $n_t = 160$ and $n_c = 640$, a treatment effect estimate, essentially an estimate of bias as the treatment effect should be zero, is -0.40. In the same run, ${s_e}^2 = \text{Var}(e) = 3564$ and ${s_{e|Z=0}}^2 = \text{Var}(e|Z = 0) = 3426$.

Thus, for the permutation-based standard error,

$$\text{Var}_P(\bar{D}(e)) = 3564(160^{-1} + 640^{-1}) = 3564(128^{-1}) = 27.85$$

and for the model-based standard error,

$$\text{Var}_M(\bar{D}(e)) = 3426(160^{-1} + 640^{-1} + .0032) = 3426(128^{-1} + .0032) = 37.70$$

which accounts for the differences in coverage and power for the two versions of the standard error.

CHAPTER III

Using multidimensional prognostic scores to make valid comparisons and inferences in observational studies: Diagnostics and application

3.1 Overview

Even with the most careful design and execution of an observational study, attempts to obtain causal inferences from the study must account for the lack of randomization to treatment conditions. Methods using matching or post-stratification aim to ensure comparability prior to the application of treatment. These adjustments can be done for one or several covariates or some score that reduces the dimensionality of the data such as a propensity score (Rosenbaum and Rubin, 1983). A recently introduced technique to help an analyst restrict comparisons to similar units or groups of units is the prognostic score, which was developed as a reduction of multivariate data to a score of lesser dimension in Hansen (2008). In this paper, we illustrate an extension of the prognostic score to multiple dimensions where the dimensions result from fits of hypothesized models. The incorporation of a matching adjustment that relies on the multidimensional prognostic scores will lead to improved comparisons to estimate causal effects of treatment.

3.2 Introduction

In observational studies, systematic differences between treatment and control groups can arise prior to the application of treatment due to the lack of random assignment. As a result, it can be difficult to disentangle the effects of treatment from these pretreatment differences. To reduce the impact of these pretreatment differences, methods such as subclassification – matching or post-stratification – on a single covariate or a univariate summary measure of the data have been used in a variety of settings. These methods aim to ensure comparisons are made between groups with covariate distributions similar to what could have been observed under random assignment. When these adjustments are required for multivariate data, summary scores of a much reduced dimension can be used in place of a larger covariate matrix. While the propensity score (Rosenbaum and Rubin, 1983) is commonly used as a summary measure of multivariate data into a single dimension, the prognostic score (Hansen, 2008) was introduced as an alternative dimension reduction of multivariate data, though not necessarily to a univariate score. In contrast to propensity scores which reflect the treatment assignment as a function of observed covariates, the prognostic score models the relationship between the outcome and covariates to reduce the data to a smaller dimension. Prognostic scores can be viewed as extracting the most important part of the covariate space for predicting an outcome of interest. We will say that this extracted score is a "prognostically relevant" part of the covariate space with respect to an outcome of interest. By extracting a part of the data that is the most prognostically relevant, prognostic scores aim to improve the comparability of groups in terms of the covariates that are most important to the outcome of interest; this comparability can be referred to as the "relevant similarity" of groups with respect to this outcome. That is, groups with high levels of comparability in terms of the covariates most predictive of some outcome can be said to be relevantly similar.

In this paper, the multidimensional prognostic score presented allows for different choices of prognostic model specification to be considered, which provides freedom in the method of extracting the most prognostically relevant linear combination of covariates from the data. Section 1 provides background information to anchor our discussion. In Section 2, we introduce an observational study of an educational intervention on which we illustrate our method. Section 3 outlines the logic of a multidimensional prognostic score and introduces diagnostics to assess which dimensions of the prognostic score should be used in a further analysis. In Section 4, we describe the use of multidimensional prognostic scores to improve the propensity score model and thus offer an improvement in the relevant similarity of the comparison groups with respect to the outcome of interest. Finally, Section 5 discusses the ability of our pre-processing procedure with multidimensional prognostic scores to accommodate studies with multiple outcomes of interest.

3.2.1 Methods for design and treatment effect estimation

As detailed in Imbens (2004), methods for treatment effect estimation in observational studies include regression methods, matching methods, propensity score methods, or some combination of these three techniques. By simultaneously adjusting for covariate differences and estimating treatment effects for all units, the use of regression models alone has the potential to confound the comparison between treatment and control groups. In addition, regression models depend on the model specification to determine the functional form of the relationships between the covariates and outcome. In contrast, matching methods, propensity score methods, or some combination are a type of data nonparametric pre-processing in the language of Ho et al. (2007): they are applied to the data in order to diagnose and help with group comparability, but by themselves, they do not directly act as methods for treatment effect estimation.

In the analysis of an observational study, pre-processing can serve as part of a "design" stage prior to the estimation of treatment effects. The concept of designing an observational study has been discussed by, among others, Cochran (1965), Rubin (2008), and Rosenbaum (2010). An observational study should be designed with careful consideration of the underlying randomized experiment that would have been conducted had it been ethical or feasible to do so. As part of a design process, an analyst must be clear on the treatment assignment mechanism, or $\mathrm{P}(Z|\mathbf{X}, Y_c, Y_t)$, and the propensity score is an attempt to recreate this mechanism. In addition to understanding how treatment was assigned, it is best if there is some known and understood exclusion criteria for excluding a unit from the experiment; units need to be included or excluded from the study on the basis of the observed covariates rather than on the treatment assignment. After treatment is assigned, which is the state of data seen in an observational study, it may be clear which units belong to a treatment group, but the selection of a comparison group from among many possibilities can present a challenge that must be carefully considered. The choice of comparison group is important as an analyst should aim for treatment and comparison groups to be as similar as possible in terms of observed information prior to the application of treatment. When a comparison group is chosen to relate to the treatment group, an analyst needs to ensure a level of similarity across these groups that could have been obtained in a randomized experiment. Preprocessing steps like matching can be viewed as a way to mirror the underlying experimental structure; whereas in an experiment, one might block on certain characteristics and randomize within blocks, in an observational study, matching creates blocks where the distribution of covariates across groups appears as though treatment had been randomly assigned.

As they are part of a “design” stage analogous to that of an experiment, pre-processing steps exclude either all outcome data or all outcome data for the treatment units. These adjustments largely focus on the observed covariate information. Early work in this type of pre-processing (Cochran, 1968) introduced the idea of matching or stratifying on the observed covariates, $\mathbf{X}$, by focusing on a single covariate. It was established that stratifying on one normally distributed covariate with equal variances across populations being compared can remove at least 90% of the bias on that covariate. After using matching or stratification, treatment effects can be estimated most simply using weighted averages. When adjusting for more than one observed covariate, a researcher will encounter the so-called “curse of dimensionality”. With multiple covariates, the covariate space is divided into such fine partitions that many subclasses have one or no units, making comparisons between treatment groups within these subclasses impossible.

To allow for covariate adjustment when there are multiple covariates that require adjustment, Rosenbaum and Rubin (1983) developed the propensity score. The propensity score is the conditional probability of receiving treatment given a unit’s observed covariates. If we subclassify on the true propensity score, then conditional on this true propensity score, treated and untreated units should have the same distributions of covariates within subclasses. After matching on propensity scores, treatment effects could again be estimated by weighted averages.

Often, regression methods, matching methods, and propensity score methods are used in combination. For example, an analyst could stratify on the estimated propensity score and use a regression model to estimate the average treatment effect within strata. As an alternative, an analyst could stratify or block on one or two covariates, match on estimated propensity score, and find treatment effects as a weighted average across matched pairs. These methods either alone or in combination are among the more commonly used methods of adjustment and estimation.

3.2.2 Assessing the quality of a match

Before estimating treatment effects for settings in which matching was used as a pre-processing step, it is helpful to assess how well the matching accounted for pretreatment differences in covariates. Within sets of a good match, distributions of observed covariates within groups should resemble those that could have been obtained from a randomized experiment. This quality assessment is crucial whether matching was done on covariates or a function of covariates like the propensity score. As was established in the discussion of design, the goal of making treatment groups similar with respect to observed covariates should guide the choice of subclassification or matching procedures, but how does one decide how to choose the stratification or matching that is most appropriate? An analyst could make several choices that impact the quality of a match. An analyst can choose a variety of distances on which to match such as Euclidean or Mahalanobis distance. In addition, an analyst can consider different forms of matching such as greedy matching (where matches are made without consideration of future matches) or optimal matching (where the best set of all matches is selected from all possible arrangements). When matching, restrictions can be imposed on the number of units allowed from each treatment group in matched set. For a discussion of these variations and their potential ramifications see Stuart (2010). Further, when matching is performed on the basis of a propensity score, Rosenbaum and Rubin (1983) detail the theoretical balancing properties of a true propensity score: when subclassifying on the true propensity score, distributions of covariates across treatment groups should be the same Rosenbaum and Rubin (1983). When using adjustments based on the propensity score, a model must be specified, and we cannot be certain that the model specified is the true model. Diagnostics have been developed to assess both the the stratification or matching chosen and, if a propensity score is used, the specification of the propensity score model. In this paper, we focus on the diagnostics collectively called balance tests.

To assess comparability of treatment groups, an analyst could use simple t-tests to compare means across groups, but these t-tests are prone to the problems detailed by Imai et al. (2008), namely, that the focus is on the means alone. Recent research has focused on creating balance tests that better compare the distributions of covariates across treatment groups for data that has been stratified or matched in some way (Lee (2008), Austin (2009), and Hansen and Bowers (2008)). If the distributions of the covariates are judged as similar enough across treatment groups, it can be argued that there are no crucial problems in either the specification of the propensity score or the stratification scheme chosen. The balance assessment of this stratification or matching scheme is crucial, as the stratification or matching design on which balance is tested will be later used to estimate causal effects of treatment.

The balance tests proposed in Hansen and Bowers (2008) rely on randomization-based inference. These tests draw on the advantages of performing inference with randomization-based methods; namely, they allow the data to be viewed as is and not as a sample from some superpopulation. As a result, these tests do not make appeals to distributional assumptions. To perform these balance tests for propensity score subclassification, the randomization distribution is computed and a measure of the difference in the covariate values between treatment and control groups is compared to this distribution. The randomization distribution is computed by randomly permuting the treatment assignments within subclasses while assuming the treatment has no effect on the covariate values; in other words, the distribution of standardized differences can be found by collecting the standardized differences that would be computed under all possible realizations of the treatment assignment vector within each subclass. With this distribution, it can be assessed whether the observed standardized differences are significant, which implies that the particular covariate being examined appears to differ in its distribution across treatment groups. An overall balance test combines the results across all covariates to obtain a χ^2 test statistic and corresponding p-value to obtain an aggregate measure of the relative lack of comparability in covariates across treatment groups.

3.2.3 Prognostic scores

The prognostic score (Hansen, 2008) is a dimension reduction technique used to improve the similarity of comparison groups in terms of covariates most important to the outcome. Prognostic scores reduce a potentially large covariate matrix to a few dimensions that have the capacity to deconfound Y and Z. In other words, after adjustment for a prognostic score, Z can be viewed as independent of Y, allowing treatment effects to be estimated as in a randomized experiment. To illustrate the idea of a prognostic score, suppose X_1 is some subset of $\mathbf{X}$ and that adjusting for X_1 is sufficient to deconfound an observational study by removing associations between Y and Z. Also suppose X_2 is some subset of $\mathbf{X}$ such that $X_2 \subseteq X_1^C$ and adjusting for X_2 in addition to X_1 does not further remove associations between Y and Z. Thus, adjustments for X_2 are irrelevant to the broader goal of deconfounding an observational study with a particular outcome of interest to make it appear more like a randomized experiment. If a reduction of $\mathbf{X}$ were defined as the set (X_1, X_2)

then, by the above discussion, X_1 could be said to be *prognostically relevant*, while X_2 can be viewed as *prognostically irrelevant.* The prognostically relevant reduction selected as the prognostic score need not be a minimal or unique reduction of the data. Adjustments for the prognostic score aim to achieve similarity in units to be compared. Clearly, units that are alike in terms of all $\mathbf{X}$ will be similar in terms of the part of $\mathbf{X}$ that is prognostically relevant, which, in the previous illustration, is X_1. However, units that share similarities on X_1 may not be alike for all parts of $\mathbf{X}$. Thus, while accounting for X_1 would ensure similarity of units on that prognostically relevant subset of $\mathbf{X}$, X_1 need not be the only such prognostically relevant subset for which this could hold.

When two units are similar in terms of the prognostically relevant part of $\mathbf{X}$, we say that those units have *relevant similarity* with respect to some outcome of interest and the particular reduction considered. In order to determine the prognostically relevant part of $\mathbf{X}$ and assess relevant similarity, an analyst must choose one set of potential outcomes – Y_t values or Y_c values – on which to base determinations. It can be reasoned that the estimate of interest is the effect of treatment on only the treated units (ETT) rather the average treatment effect (ATE) for all units. An overall estimate of treatment effect would assume a constant average treatment for all units, whether or not they could reasonably receive the treatment. As Heckman (1997) argues, a researcher interested in improving policy is probably not concerned with the effects of, for example, a job training program on a millionaire; it is more reasonable to examine the effects of a job training program on the the subjects who utilized the program, as they are individuals like those who would reasonably be affected by any policy related to such a program. By obtaining an estimate of the ETT, an analyst can answer the question: what is the benefit of the job training program to those who participated in comparison to what they would have experienced without participation in the program? Often, this is the far more relevant question for interventions examined by observational data. Further, focusing the estimation procedure on potential outcomes under control links the ideas of prognostic scores to older techniques in the literature used to estimate treatment effects when treatment is some level of disadvantaged or minority status. This method, which originated with the work of Peters (1941) and Belson (1956), became known as the Peters-Belson procedure and focuses on predicting potential outcomes under control for treatment units supposing they were members of the control group. Unlike prognostic scores, the Peters-Belson procedure aims to estimate treatment effects; prognostic scores are a pre-processing method. In addition to these reasons for preferring to use Y_c values over Y_t values to assess similarity, the control group is typically larger, thus, there are more measured values of Y_c.

To formalize the ideas of prognostic scores defined with respect to potential outcomes under control, define $\Psi(\mathbf{X})$ as some reduction of $\mathbf{X}$. The function $\Psi(\mathbf{X})$ could be defined through any standard dimension reduction methods; a basic form would be the linear combination of covariates determined by a linear regression model. We can say $\Psi(\mathbf{X})$ is a prognostic score if and only if conditioning on it removes associations between the covariates and potential outcomes under control such that $Y_c \perp \mathbf{X}|\Psi(\mathbf{X})$ for any $X \in A$, where A is a measurable set (Hansen, 2008). When associations between covariates and potential outcomes are removed, there should be no systematic association between uncontrolled variation in covariates and the outcomes. Thus, any resulting links between treatment conditions and outcomes are more likely to be the result of the treatment itself and less likely to be the result of uncontrolled background information. In the absence of bias due to unobserved information that was not included in the covariate matrix $\mathbf{X}$, $Y_c \perp Z|\Psi(\mathbf{X})$ for any $X \in A$, where A is a measurable set (Hansen, 2008). In other words, conditioning on the prognostic score removes associations between the potential outcomes under control and the treatment assignment. Specifically, provided there is no level of the prognostic score at which units receive the treatment with certainty, an ETT estimate can be computed. If $Y_c \perp z|\mathbf{X}$ and with probability one, $\mathbf{P}(z = 1|\Psi(\mathbf{X})) < 1$, then

$$\begin{aligned}
\tau_{ETT} &= \mathbf{E}(Y_t - Y_c|z = 1) \\
&= \mathbf{E}[\mathbf{E}(Y|z = 1, \Psi(\mathbf{X})) - \mathbf{E}(Y|z = 0, \Psi(\mathbf{X}))|z = 1]
\end{aligned}$$

Results on the deconfounding properties of the prognostic score and the ETT estimate will be important to ground the discussion of the multidimensional prognostic score in Section 3.4. In the present development of a multidimensional prognostic score, however, rather than conditioning directly on a prognostic score to estimate treatment effects, a function of it will be used. Namely, estimation will condition on a propensity score formed with prognostic score dimensions, or a prognostic-propensity score, as discussed in Section 3.5. While the discussions in this paper will focus on prognostic scores as a pre-processing step in the design of an observational study, the ultimate aim of the pre-processing is an ETT estimate similar to that defined as τ_{ETT}.

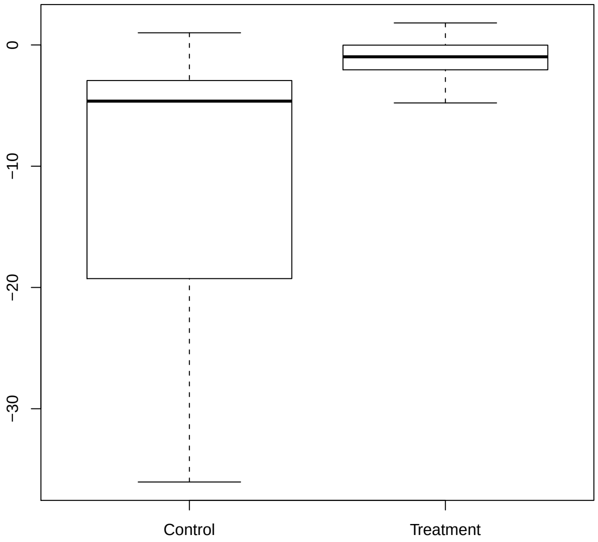
3.3 Study of an educational intervention

To illustrate our extension of the prognostic score to multiple dimensions, we use a study of an educational intervention in schools in Texas. The initial reviews of the program compared a subset of Texas schools that elected to adopt the program to others that did not. As adoption of the program was by choice, the problem inherent to observational studies arises: the absence of random assignment of schools to receive or not receive the program may lead to systematic imbalances in the treatment and control groups. To examine the effectiveness of the intervention by assessing student progress, data from the Texas Assessment of Knowledge and Skills (TAKS) is used, which is publicly available data from the Texas Education Agency.

The data set contains measurements aggregated to the school level for demographic characteristics and test outcomes. For each school, multiple years of data are included: measurements were taken in 2005, 2006, and 2007. There are 1475 control schools, schools that did not receive the intervention, which provide 4425 observations across the three years of data collection. In addition, there are 177 treatment schools that were involved in the program for one, two, or three years of the period of the study. These treatment schools provide 322 observations. With different observations for the same school across several years, the observations in the data set can be viewed as school-year observations, a combination of data for both the school and year. The data provides 11 outcome measures to assess student progress, including one aggregate outcome measure for Algebra 1 performance; our analyses focus on this aggregate outcome measure.

If we fit a standard propensity score model with logistic regression to the data, we can use the estimated propensity scores to provide intuition regarding the overlap in the covariates of the treatment and control groups. Side-by-side boxplots, appearing in Figure 3.1 illustrate the extent of the overlap an analyst might see if she fit propensity scores to the large set of covariate information available without attention to which covariates may be more prognostically relevant. The overlap is minimal and seems to indicate problems for further inference. Some have suggested that situations such as these are hopeless for further data analysis, and do not consider potential adjustments to the propensity score specification (Rubin, 2007). This paper develops techniques to improve this standard pre-processing of observational data through an enhanced propensity score model.

Figure 3.1: Boxplots comparing the linear propensity scores of all control (left) and treatment (right) schools



3.4 The multi-dimensional prognostic score

In contrast to the single dimension of propensity scores, prognostic scores can have many dimensions. Analysts comfortable with propensity score methods may view the potential for multiple dimensions as a drawback; an advantage of the propensity score machinery is its ability to transform a wealth of covariate information into a univariate representation, ameliorating problems due to the so-called curse of dimensionality. The extra dimensions in a prognostic score can arise for two reasons: the continuous nature of Y_c and the consideration of multiple models.

Propensity scores capture the relationship between a typically binary Z and multidimensional $\mathbf{X}$, while prognostic scores capture the relationship between a continuous Y_c and multidimensional $\mathbf{X}$. The single dimension of a propensity score estimates the relationship $\mathbf{E}[Z|\mathbf{X}] = \mathbf{P}(\mathbf{Z} = \mathbf{1}|\mathbf{X})$. As we can write $\mathbf{P}(Z = 0|\mathbf{X})$ as $1-\mathbf{E}[Z|\mathbf{X}]$, then computing $\mathbf{E}[Z|\mathbf{X}]$ provides all the information that could be obtained for the conditional distribution of Z given $\mathbf{X}$. With a continuous Y_c, however, modeling $\mathbf{E}[Y_c|\mathbf{X}]$ as the prognostic score does cannot provide us with the same completeness of information about the conditional distribution of Y_c given $\mathbf{X}$. Using additional dimensions for the prognostic score may help add to the ability of the prognostic score to capture the conditional distribution of Y_c given $\mathbf{X}$.

Aside from the difference between a binary Z and a continuous Y_c, the multiple dimensions of a prognostic score help account for model misspecification. With a propensity score, one commits to a single dimension and a single model for a propensity score, but it is difficult to know how to select the best propensity score model. Many agree that standard variable selection techniques are not adequate because the goal of a propensity score model should not be to best predict treatment (Brookhart et al., 2006). With prognostic scores, each dimension can be the result of a different model without choosing an absolutely "correct" specification. Each model represents a unique way of extracting the most prognostically relevant part of the covariate space. As many specifications are possible, the fitted prognostic score is less subject to the criticism of an incorrect model specification. Empirical results suggest the additional dimensions greatly improve the ability of an analyst to achieve this comparability.

3.4.1 Diagnostics for prognostic score specifications

With prognostic scores, diagnostics provide goodness of fit tests of the proposed prognostic score dimension(s). The goal of prognostic scores is to extract the linear combination of known covariates that is prognostically relevant, or predictive of an outcome of interest, based on a model fit in the control group. Prognostic score diagnostics are assessed for the proposed dimensions for the control group alone to determine if the proposed prognostic scores capture all of the prognostically relevant part of $\mathbf{X}$. For prognostic scores, we define a concept called prognostic balance: prognostic balance is a measure of association between covariates and outcomes for control group units with a small distance on the estimated prognostic score. An assessment of this prognostic balance provides an indication as to how well the chosen prognostic dimensions fit the data, making it an important diagnostic step in selecting the best prognostic score for use in later adjustments and analyses. The machinery of diagnostic tests for prognostic balance builds on that discussed for propensity score diagnostics in Section 3.2.2

The assessment of prognostic balance relies on comparisons within matched pairs of control units. If there is no significant association between the outcomes and the covariates in the control group as determined by examining these matched pairs, then theoretical results in Hansen (2008) indicate that further pre-processing steps can be taken with the proposed prognostic dimensions. To match pairs in the best way possible, the diagnostic uses optimal matching, which is a matching routine that considers all possible collections of matches among all units and selects the best collection of all possible. A method for performing an optimal full match for pairs in a nonbipartite setting – a setting in which any unit can, in theory, be matched to any other unit other than itself – is provided by Lu et al. (2001), which applies the algorithm devised by Derigs (1988). In order to measure the “distance” between any two units, a matrix of all possible pairwise distances is constructed. Between any two units, their separation is defined by a Mahalanobis distance on the estimated prognostic score dimensions. From these distances, the optimal matching routine chooses the best matching arrangement for all units that minimizes the sum of the matched pair distances across pairs. Although other methods could be used in a diagnostic for prognostic scores, optimal matching in this way allows an analyst to rule out problems with the matching procedure as a cause of troublesome prognostic balance results. In nonbipartite matching, to disallow very poor matches from being made, dummy observations are included in the matching routine. As a result, school-year observations that are very unlike any other observation in terms of measured covariates, such as demographic composition or pretest scores, are not included in the assessment of prognostic balance. In addition to using these dummy observations in the nonbipartite matching, for computational ease, control subjects are matched on the prognostic score within quintiles of the first dimension of the prognostic score (or on the prognostic score itself if the score is unidimensional).

Following the creation of matched pairs of control units, hypothesis tests are used to assess the prognostic balance, which can be viewed as a measure of the “goodness of fit” of a prognostic score specification. In this situation, goodness of fit refers to how well the prognostic score pairings account for the association between the outcome and a given covariate. If the prognostic score is a good fit, within matched pairs, there should be no association between the Y_c value and $\mathbf{X}$ values. If we suppose we have a single covariate X and Y_c were binary so that each pair contained one unit with each value of Y_c, matched set differences on X across the two levels of Y_c could be computed. The differences could be aggregated across sets with a weighted average to provide a statistic, but a distribution would need to be chosen to conduct a hypothesis test. Permutation tests allow this assessment to be conducted in the absence of any additional distributional assumptions. By permuting the values of Y_c between the X values in a matched pair and computing a collection of aggregated pairwise differences across all pairs, a reference distribution can be obtained to which we can compare the realized value of the statistic. It is not the typical case for an outcome of interest to be binary, but similar logic can be applied for a continuous outcome. Rather than computing differences across the two values of Y_c, with a continuous outcome, measures of correlation with a covariate in the form of regression coefficients are used. These measures of correlation can be added across sets to obtain a statistic. The reference distribution is computed similarly to that with a binary Y_c, except all possible sums of correlations – rather than differences – are collected for a reference distribution.

The measure of prognostic balance can be defined as the coefficient of Y_c in a regression of X_i on Y_c and S, with S defined to be a factor variable of subclass membership. Another way of defining this prognostic balance for a single covariate X_i is

$$\hat{\beta}_{Y_c X_i} = \frac{\sum_S (X_{iS} - \bar{X}_{iS})'(Y_{cS} - \bar{Y}_{cS})}{(Y_c - \bar{Y}_c)'(Y_c - \bar{Y}_c)} \tag{3.1}$$

where X_{iS} is a vector of values of X_i in stratum S, Y_{cS} is a vector of Y_c values in stratum S, and $\bar{X}_{iS}$ and $\bar{Y}_{cS}$ are the means of these vectors, respectively.

In the present study, more than one covariate is of interest, so a summary measure that incorporates all covariates is needed. The permutation test gives rise to covariances between balance coefficients, which are denoted as $\text{Cov}(\hat{\beta}_{Y_c X_i}, \hat{\beta}_{Y_c X_j})$ for any i and j. These can be collected into a covariance matrix V where the $(i,j)^{th}$ entry is $\mathrm{Cov}(\hat{\beta}_{Y_c X_i}, \hat{\beta}_{Y_c X_j})$. If we let $\hat{\beta}_{Y_c}$ be the vector of imbalance coefficients defined in Equation 3.1 for all covariates or columns of $\mathbf{X}$ the measure of imbalance across all covariates can be defined as a statistic d where $d = \hat{\beta}'_{Y_c} V \hat{\beta}_{Y_c}$, using results developed by Hansen and Bowers (2008). The statistic d follows an approximate χ^2 distribution, which allows a test of the null hypothesis that the prognostic score dimensions chosen are a good fit for the data. This balance assessment provides appraisals of the prognostic score dimensions across all covariates of interest, but offers little guidance regarding the potential problems that may lead to a lack of fit of selected prognostic score dimensions. Investigating covariate imbalance by matched pair with the balance plots introduced in Section 3.4.3 provides additional insights.

3.4.2 Choosing the dimensions of a multi-dimensional prognostic score

A prognostic score aims to reduce the dimensionality of the covariates by modeling the response as a function of the covariates for the control group with the aim of extracting the linear combination of covariates most highly related to the outcome. In this way, the prognostic score can be viewed as a dimension reduction technique for the covariates fit to the control group alone. Many statistical analysis procedures could be used to execute this dimension reduction, and different model considerations comprise candidates for prognostic score dimensions. In many instances, several models for data are justifiable, and an analyst must choose between them. When the data suggest multiple valid models to determine the most prognostically relevant part of the covariate space, they can be considered candidates for dimensions of the prognostic score. In this case study, the following dimensions will be considered for reasons which will be subsequently explained:

1. Ordinary least squares model

2. Ridge regression model
3. Weighted least squares regression
4. Inverse of the sample size of each school

From one perspective, the linear combination of covariates that seems to matter most is the linear combination determined by an OLS model. Although OLS regressions are easy to implement and widely understood, they have a few obvious limitations and may not always serve as the best model. In the education intervention data, when the model includes all the covariates that might be deemed relevant for a prognostic score model, singularity in the design matrix prevents the OLS model from computing some coefficients. To accommodate the additional information without removing covariates from the model, the linear combination of covariates that results from a ridge regression could be considered as a candidate for a prognostic dimension. The ridge regression model, fit using the generalized cross validation estimate of the penalty provides another, yet related, set of fitted values that could be viewed as the estimated prognostic scores.

An additional candidate dimension for the prognostic score that augments the single dimension fit by either OLS or ridge regression imagines data in which, rather than aggregating information to the school level due to convenience or privacy concerns, individual data nested within schools was obtained. If this were the case, the data would have included individual-level and school-level information. This would lead to a multilevel modeling framework with a school-level random effect. In this framework, the error term in the model has a variance from both levels: the error term for the model would have variance of $\sigma_2^2 + \sigma_1^2/m_s$ where σ_2^2 is the variance of the errors for the students and σ_1^2 is the variance for the errors for the school effect with m_s the school size in number of students. Viewing the data in this way, an analyst might wish to couple the OLS fitted values with the reciprocal of the school size (as the reciprocal of school size is bounded) to incorporate this background multilevel model idea with the given information. The reciprocal of school size alone or in combination with the linear model fitted values provides another possible dimension of the multidimensional prognostic score to investigate.

An equally logical but distinct way of obtaining the most prognostically relevant linear combination of covariates builds on the OLS model by weighting school-year observations and fitting a weighted least squares, or WLS, model. Thinking in terms of the multilevel modeling framework outlined above, the variance of the errors is defined by

$$\begin{aligned}\mathrm{var}(\epsilon) &= \sigma_2^2 + \sigma_1^2/m_s \\ &= \sigma_2^2\left(1 + \frac{\sigma_1^2}{\sigma_2^2}\frac{1}{m_s}\right)\end{aligned}$$

For the control schools, the data contains 3 years of information for each school. Using this information, an estimate of the variability between schools and within schools can be obtained and used as estimates of σ_1^2 and σ_2^2, respectively. Thus, it is natural to select $(1 + \frac{\sigma_1^2}{\sigma_2^2}\frac{1}{m_s})^{-1}$ as our weight, substituting the estimates of the two variances in place of the true values. When a WLS model is fit using these weights, the variance structure of the speculated multilevel model is incorporated into the standard regression model. The fitted values from this model provide another candidate dimension for the prognostic score.

If an analyst had other justifiable models to extract the most prognostically relevant part of the covariate space, he could continue to obtain an even larger set of candidate models. Rather than deciding between the models or devising some system of weights to take a weighted average across the candidate models, several models with attractive features can be considered. Using various combinations of these candidate dimensions, the prognostic balance is assessed to determine the goodness of the fit of different prognostic scores; the diagnostic uses a measure of prognostic balance to determine which combination of these candidate models will provide the dimensions to the chosen prognostic score. Table 3.1 provides results for the balance assessments of many one, two, three, and four dimensional prognostic scores. The columns of Table 3.1 present the proportion of controls matched by the nonbipartite matching routine, the χ^2 statistic that results from the balance assessment, and the corresponding p-value.

The prognostic balance appraisals presented in Table 3.1 indicate that if an analyst includes multiple dimensions in the prognostic score, the prognostic score will be a good fit for the data. The analyst who otherwise would have struggled to select only one model or may have devised some scheme for combining models to select only one dimension finds a benefit by utilizing multiple candidate dimensions. The p-value for the balance assessment for all four candidate dimensions is a robust 0.49; that of the standard OLS fit alone is a paltry 0.013. Adding in just one more dimension to the dimension fit by OLS, specifically the fitted ridge regression estimates, improves the fit of of the prognostic model: for this specification of a multidimensional prognostic score, the p-value is 0.77, which is, in fact, the highest p-value across all specifications. Not all two-dimensional prognostic scores fit the data as well as the combination of the OLS and ridge regression fitted values. Among the three-dimensional prognostic score specifications presented in the table, there is comparatively little difference across the three fits. While it appears that the additional dimension may fit worse than the best two-dimensional prognostic score, the results with the educational intervention data show the goodness of fit of the prognostic model is far less dependent on the specific dimensions selected when an analyst uses more dimensions.

For purposes of illustration, much of the current discussion will focus on the four-dimensional prognostic score. For this specification of the prognostic score, we can further assess the fit through balance plots.

Table 3.1: Results of prognostic balance across pairs when performing a non-bipartite matching with sinks on the prognostic scores indicated

Prognostic score dimensions used	% controls matched	χ^2 statistic ($df = 110$)	p-value
One dimension:			
OLS	92	146	0.012
Ridge Regression	91	149	0.008
WLS	90	254	0.000
Reciprocal of school size	96	716	0.000
Two dimensions:			
OLS and Ridge	96	99	0.77
OLS and WLS	96	126	0.14
OLS and reciprocal of size	93	115	0.35
Three dimensions:			
OLS, Ridge, and WLS	97	101	0.72
OLS, Ridge, and reciprocal of size	97	110	0.48
OLS, WLS, and reciprocal of size	96	106	0.59
Four dimensions:			
OLS, Ridge, WLS, and reciprocal of size	97	110	0.48

3.4.3 Balance plots to further assess prognostic balance

In addition to examining the measure of prognostic balance for a covariate X across all subclasses, we can plot the cumulative contribution to the prognostic balance measure across subclasses. To introduce these plots, we focus on pair matching; thus, subclasses are pairs of units. If we plot the contribution of each pair to the measure with pairs ordered in some manner – such as by propensity score averages or prognostic score differences – we can gain a better understanding of the sources of the imbalance. The endpoint of each cumulative sum of prognostic balance aligns with the z-statistics provided using the balance tests of Hansen and Bowers (2008). The overall imbalance can be decomposed into the sum of the imbalance across pairs or strata as in Equation 3.1. Thus, for each matched pair denoted by S, the contribution to the measure of prognostic balance on X_i for that pair is one piece of the summation in 3.1:

$$\frac{(X_{iS} - \bar{X}_{iS})'(Y_{cS} - \bar{Y_{cS}})}{(Y_c - \bar{Y_c})'(Y_c - \bar{Y_c})}. \tag{3.2}$$

The plots present prognostic balance against ordinal scales – rather than interval scales – of either prognostic score distance or propensity score average within pairs; that is, the pairs are ordered by increasing distance or average and the values are labeled at regularly spaced intervals in terms of the number of pairs. In addition, to help an analyst better focus on the prognostic balance in control units, the measures are weighted in a manner inversely proportional to the probability of being in the control group (or, $1 - E(Z|\mathbf{X})$ where $E(Z|\mathbf{X})$ is the estimated propensity score). This weighting magnifies any lack of prognostic balance in school-year observations in the control group that seem to be most strongly like treated school-year observations on the basis of the observed covariate information. Without the weights, an analyst would still have an accurate picture of prognostic balance; the weighting is an additional tool to indicate potential problems.

Plots of prognostic balance by propensity score average, for example, can indicate if more of the problems with prognostic balance are due to regions with large or small propensity scores. Figure 3.2 presents prognostic imbalance across pairs as a function of average propensity scores. In Figure 3.2, the plots shown present two racial composition variables: the percent white and the percent black in a given school. Neither would yield a measure so large to be ruled a statistically significant lack of fit; however, while the plot for the percent black has similar features to random fluctuation with no problematic trends, the plot for the percent white shows trending in the rightmost section of the plot. This trending can indicate a potential problem as a pattern similar to random fluctuation is what one should expect if any departures from prognostic balance were not systematic across pairs. The region of the trending on this plot indicates that pairs with the highest average propensity score, units that are most like treatment units among all the control units, seem to have a systematic imbalance in the percentage of the student body that is white. A variable of this nature could be addressed by additional covariate adjustment in the process of making estimations of causal effects, which will be discussed in Section 3.5.

In a similar manner, plots of this prognostic balance by prognostic score distance within pairs can show if more imbalance comes from poorly matched pairs (i.e. pairs with far apart units). The plots of Figure 3.3 show the balance in two covariates ordered by matched pair distances on prognostic score, ordering pairs from best matched to most poorly matched. In Figure 3.3, the pretest score for the Algebra 1 measurement, A1SS.py, is clearly imbalanced as the sum of the standardized imbalance measure exceeds the line indicating a corresponding p-value of less than 5%. The other variable presented shows no real problem with imbalance.

The assessment of prognostic balance with the balance plots offers some guidance to the researcher about next steps to take or issues to be keenly aware of when moving on to the next steps in adjustment and effect estimation. Significantly imbalanced covariates can be included in the prognostic-propensity model along with a prognostic score of any dimension to offer additional adjustment for these problematic covariates

Figure 3.2: Two selected balance plots with pairs arranged by average propensity score within the matched pair

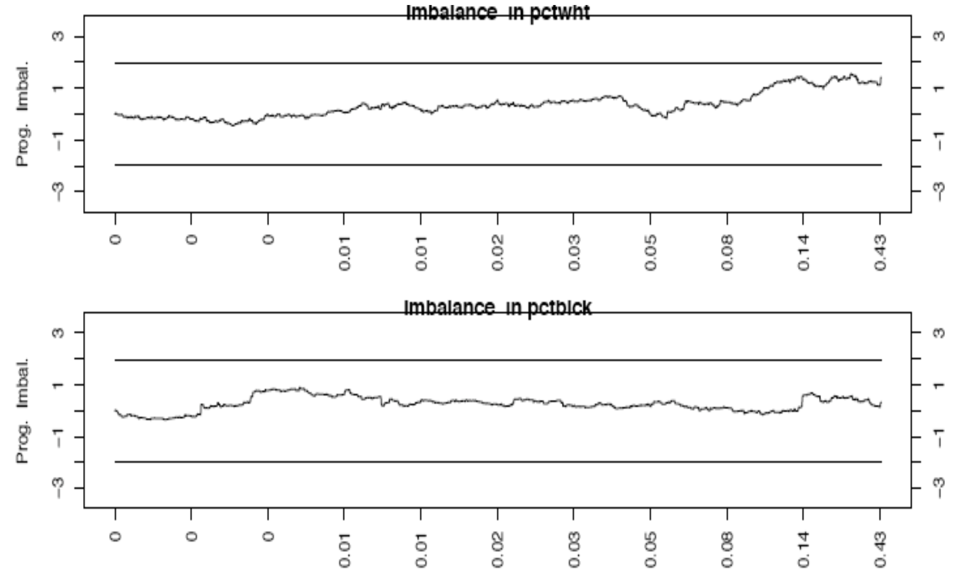
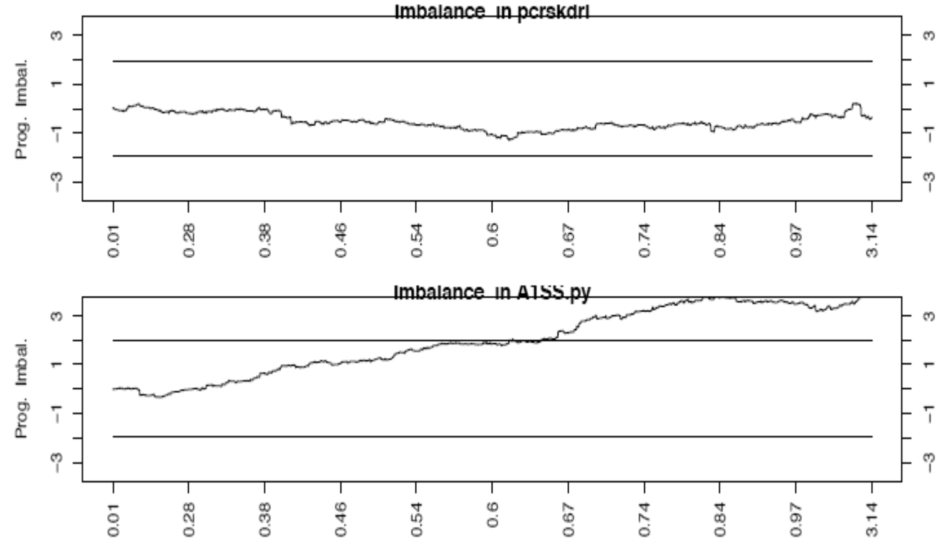


Figure 3.3: Two selected balance plots with pairs arranged by matched Mahalanobis distance on the estimated prognostic score



as will be discussed in Section 3.5.

3.5 Using the extended prognostic score in a prognostic-propensity score

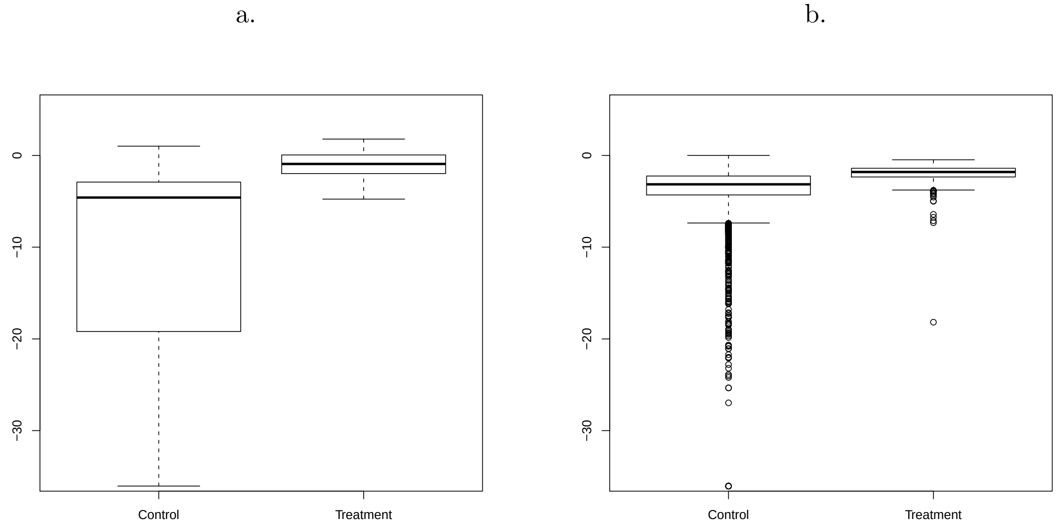
Investigations into propensity scores and diagnostics examine which variables are most crucial to include in a propensity score to get the best estimates of treatment effects. Rubin and Thomas (1996) suggest including all variables thought to be related to the outcome variable whether or not those same variables are related to the treatment variable. In fact, including variables related to the treatment but unrelated to the outcome can decrease the efficiency of the estimate of treatment effect based on a propensity matching; if a variable has some relationship to the outcome, the increase in bias due to omitting such a variable may outweigh any decreases in efficiency (Rubin, 1997). More recent simulation results by Brookhart et al. (2006) confirm these assertions. A propensity score model, typically a logistic regression model, that is optimized to predict treatment assignment need not be optimal in terms of mean squared error of subsequent treatment effect estimates. Standard variable selection routines one might perform on a propensity score model would eliminate covariates that do not aid in predicting the treatment value. From their simulation studies, Brookhart et al. (2006) find that including covariates unrelated to the treatment or those related in some small, chance way, removes nonsystematic bias due to chance associations between the treatment and particular covariate. Further, they recommend not excluding covariates unless there is strong prior evidence that a particular covariate is unrelated to the outcome of interest.

Many sources indicate it is advisable to include covariates most relevant to the outcome of interest in the propensity score model. In order to do that clearly, it must be decided which covariates are related to the outcome of interest. If this assessment is made using all information available about the study population, as is the case in Brookhart et al. (2006), it is as though the analyst is using results of the outcome analysis in the creation of the propensity score, which is part of a pre-processing step and should be separate from outcome analyses. The prognostic dimensions discussed in Section 3.4.2 are, by construction, strongly related to the outcome of interest and incorporate information from the covariates. Forming a propensity score with the prognostic score dimensions produces a propensity score focused on a set of linear combinations of the covariates most relevant to the outcome of interest – precisely what previous research has indicated yields the best propensity score – without previewing the results of the outcome analyses. We call this type of propensity score a prognostic-propensity score following Hansen (2008).

To create a prognostic-propensity score using these dimensions, we first compute the fitted values of the prognostic score for the treatment group using the models previously estimated using the control group. In the educational intervention data, this amounts to using the model fit on the 4425 control units to obtain predictions for the 322 treatment units. For the dimension that is the reciprocal of the school size, the data from the treatment group can be used directly. As 117 of the control units were matched to dummy observations for the prognostic score diagnostics, we can reason that they are very unlike the other units under study in terms of observed covariates. As a result, when creating the prognostic propensity score, only 4308 controls, or 4630 total units are used. With the prognostic dimensions for the 4630 units, we can form a propensity score using the dimensions as the predictors in the propensity score model. A logistic regression model is fit using these predictors and the treatment indicator as the response.

The propensity scores estimated by this model can be used to improve our estimates of treatment effects in any of the standard methods that utilize propensity scores. The boxplots in Figure 3.4 present the distribution of the propensity scores and prognostic-propensity scores for the four-dimensional prognostic score in the control group as compared to the treatment group. There is little overlap in the distribution of linear propensity scores estimated in the standard way across treatment and control groups, but there is substantial overlap in the distribution of the linear prognostic-propensity scores as is apparent in the boxplots of Figure 3.4. It is this overlap that forms the foundation for any causal inferences an analyst desires to make. The use of the prognostic-propensity score improves the accuracy of the estimation of causal effects as a consequence of the improvement in the relevant similarity of the comparison groups with respect to the outcome of interest – in this case, test scores. Assessing overlap across treatment groups on the covariates most relevant to the outcome, where these covariates are defined by the fitted prognostic score dimensions rather than the true prognostic scores, is justified by the results presented in the Appendix.

Figure 3.4: Boxplots comparing the distributions of the linear propensity scores for only the units that could be matched in the prognostic score diagnostic for a) the standard propensity score, and b) the prognostic-propensity score fit with the four-dimensional prognostic score. In both plots, the control group is on the left and the treatment group, the right.

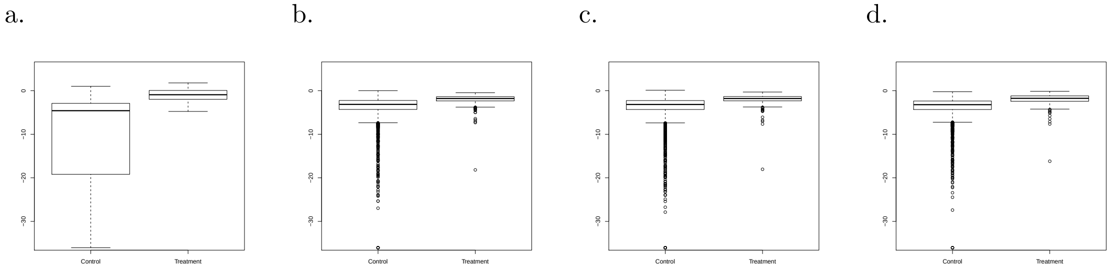


As indicated in Section 3.4.3, individual covariates strongly out of balance in the prognostic score balance diagnostics can be included along with the fitted prognostic dimensions to further improve balance prior to making causal inferences. We can identify problematic covariates of two types. A covariate can be problematic if its z-statistic, the sum of standardized imbalance across pairs, is significant. In the the illustration with the data, two pretest scores – one for a specific object and the overall Algebra 1 pretest score – have a significant lack of prognostic balance. They are included along with the four prognostic score dimensions to form a modified prognostic-propensity score. The distribution of the fitted values for this prognostic-propensity score are shown in the boxplots in the third panel of Figure 3.5 (the first and second panel present the same figures from Figure 3.4 for comparison).

Balance plots provide another method of identifying a problematic covariate to add to the predictors in the prognostic-propensity score model. Even if a covariate is not deemed to have a statistically significant lack of prognostic balance, it is a candidate to include in the prognostic-propensity score if the balance plots indicate its imbalance is systematic in the regions of the plot corresponding to pairs with close matches or relatively large propensity scores. If we include a covariate of this nature, such as the percentage of students in a school who are white based on the balance plots in Figure 3.2, in combination with the two imbalanced pretest measures, a third prognostic-propensity score can be estimated. The distribution of the fitted values from this propensity score is presented in the final panel of Figure 3.5, and indicates changes in the balance between the distributions of the fitted values across treatment and control groups as compared to the two previous prognostic-propensity scores.

The use of prognostic scores in combination with propensity scores aims to focus the propensity score on the covariates most relevant to the outcome of interest to

Figure 3.5: Boxplots comparing the distributions of the linear propensity scores for only the units that could be matched in the prognostic score diagnostic for a) the standard propensity score, and b) the prognostic-propensity score fit with the four-dimensional prognostic score c) the prognostic-propensity score with significantly imbalanced covariates, and d) the prognostic-propensity score with significantly imbalanced covariates and a problematic covariate diagnosed by the balance plots. All plots present the control group on the left and the treatment group on the right.



achieve comparability across treatment and control groups. Using these variations of the prognostic-propensity score an analyst can greatly improve upon the overlap in the distributions of propensity scores for units being compared, and thus also improve the overlap in the most prognostically relevant covariates as indicated to the across treatment and control groups. This prognostic-propensity score can be used in much the same way as a standard propensity score; for example, an analyst can stratify or match on the propensity score and estimate treatment effects as a weighted average across matched sets or strata. To assess improvements to the overlap on the constructed covariates most relevant to the outcome, the prognostic score dimensions, we can apply balance tests in the way they are applied to propensity-matched or propensity-stratified data sets. Table 3.2 presents balance test p-values across different scenarios to assess covariate balance between treatment and control groups. The p-value in the first column is small if statistically significant imbalances exist between treatment and control groups for either all the variables in the propensity score model by default or all the variables in the prognostic-propensity score model, if a prognostic-propensity score model was fit. The p-value in the second column is small if statistically significant imbalances exist between treatment and control groups for the four prognostic score dimensions selected in Section 3.4.2.

Table 3.2: Balance comparisons between treatment and control groups without and with stratification on several propensity and prognostic-propensity scores

	All variables in the model for propensity or prognostic-propensity score	Four prognostic score dimensions
No stratification	0.00	0.00
Stratified on standard propensity score	0.03	0.02
Stratified on prognostic-propensity score with 4 dimensions	0.77	0.77
Stratified on prognostic-propensity score with 4 dim. and 2 covariates	0.59	0.46
Stratified on prognostic-propensity score with 4 dim. and 3 covariates	0.45	0.28

The first row of Table 3.2 indicates, in the absence of stratification on any estimated score, highly statistically significant imbalances exist between treatment and control groups for all variables included in the propensity score model and the four prognostic score dimensions. The remaining rows of Table 3.2 incorporate stratification of units into quintiles on the basis of either the propensity score or a prognostic-propensity score, using a standard recommendation for removing bias in comparative studies. The balance test p-values then indicate if statistically significant imbalances in covariates exist within strata; consequently, if treatment effect estimation is performed within strata, and there is no statistically significant imbalance in covariates within strata, estimates of treatment effect should be less biased than corresponding estimates not accounting for the stratification. Based on the results in the table, stratification on the standard propensity score still leaves statistically significant imbalances within strata between treatment and control groups. This conclusion seems logical based on the overlap that appears in plot (a) of Figure 3.5. While stratifying on one of the prognostic propensity-scores creates strata in which there is no statistically significant imbalance on the four prognostic score dimensions, the prognostic-propensity scores with additional covariates added do not indicate any additional improvement in covariate balance in this case study. The balance results of Table 3.2, viewed in conjunction with the plots in Figure 3.5, provide evidence of the benefit to a comparative study in terms of covariate balance from incorporating multidimensional prognostic-propensity scores. As a consequence of these improvements in covariate balance between treatment and control groups, using the prognostic-propensity scores discussed in this section in place of the standard propensity score should produce improved estimates of treatment effects that are less subject to systematic differences in the distributions of the covariates most related to the outcome across treatment groups due to the lack of random assignment.

3.6 The case of multiple outcomes of interest

In the preceding sections, the argument for multidimensional prognostic scores is illustrated on data with one primary outcome of interest. Frequently, an analyst is faced with multiple outcomes of interest and believes the data pre-processing should make comparison groups relevantly similar with respect to all outcomes. With prognostic scores alone, achieving this relevant similarity is difficult or impossible; however, the prognostic-propensity score allows for data pre-processing considering multiple outcomes. This section presents a brief illustration of how the techniques of multidimensional prognostic scores and prognostic-propensity scores can be employed for studies with multiple outcomes of interest.

The data offers an aggregate measure of Algebra I performance in addition to ten individual measures, or objects, that aggregate to the composite measure. The individual measures are highly related to the aggregate measure; a regression of the composite measure on the ten separate dimensions yields an R^2 of 0.988. Suppose an analyst, faced with ten separate outcome measures for mathematics performance, wants to use prognostic scoring for an adjustment method. As the different outcomes are related in this case, one possibility would be to create an aggregated outcome measure and create prognostic scores for that outcome as described in Section 3.4.2. This method would seem illogical if the outcomes had no relation to each other; for example, if the different outcomes were for reading, mathematics, and writing scores, an analyst may have difficulty composing an aggregated outcome measure without the input of education experts who could devise a weighting for the three scores. We view the creation of prognostic scores as part of data pre-processing, so an analyst need not struggle to obtain one outcome measure. Prognostic scores from multiple outcomes can be created according to the process described in Section 3.4.2, and all dimensions can be included as predictors in the prognostic-propensity score model.

As a brief illustration, prognostic scores were fit using prognostic scores fit for both the aggregate outcome and the ten separate outcomes with OLS regression alone (the prognostic score for the aggregate outcome is identical to the OLS fit described in Section 3.4.2). For each of the ten separate outcomes, prognostic balance was assessed with respect to the outcome used as the response variable in the prognostic score model. For the prognostic score fit to the aggregated measure of Algebra performance, prognostic balance was assessed with respect to the aggregated measure as well as the ten separate outcomes. Table 3.3 summarizes the results. Prognostic balance with respect to a given object or outcome is always better for the prognostic score fit to the specific outcome rather than the prognostic score created from the aggregated measure. Prognostic balance with respect to the aggregated outcome for the score created from it is much worse than the prognostic balance for the individual outcome scores. Based on prognostic balance results, if an analyst is choosing dimensions of a prognostic score for her prognostic-propensity score, it is clear that the scores from the individual outcomes are the better choice. After deciding to use a variety of individual outcomes, an analyst can follow the logic detailed in Section 3.4.2 to refine the prognostic score selections, possibly adding additional scores for outcomes on the basis of different modeling perspectives.

Table 3.3: Results of prognostic balance across pairs with multiple outcomes for three selected outcomes

Response in prognostic score model	Balance assessed with respect to	% controls matched	χ^2 statistic ($df = 110$)	p-value
Object 8	Object 8	91%	103	0.67
A1SS	Object 8	92%	165	0.00
Object 9	Object 9	91%	86	0.96
A1SS	Object 9	92%	283	0.00
Object 10	Object 10	90%	112	0.43
A1SS	Object 10	92%	192	0.00
A1SS	A1SS	92%	146	0.01

After an analyst settles on which individual outcomes to include and how to model these outcomes, further pre-processing steps can be taken. Prognostic-propensity scores can be fit for both the single dimensional prognostic score fit to the aggregated outcome and the ten-dimensional prognostic score fit to the separate object measures. On the basis of the resulting prognostic-propensity scores, the observations can be stratified into quintiles for a quick test of the comparability of covariates in the treatment and control groups. For these prognostic-propensity scores fit from simplistic OLS model-based prognostic scores, the covariates are quite imbalanced regardless of which prognostic-propensity score is used. When the prognostic-propensity score with the ten-dimensional prognostic score is used, however, the p-value for the balance test described in Hansen and Bowers (2008) is roughly twice that of the p-value for the balance test with the prognostic-propensity score for the single dimensional prognostic score. Thus, with the resulting prognostic-propensity score, the multidimensional prognostic score in which the dimensions are defined by separate outcomes improves comparability of the treatment groups more than that created using an aggregated outcome measure.

3.7 Discussion

Prognostic scores prove useful in their contributions to the pre-processing step of balancing the distributions of covariates in an observational study. They provide an analyst with the advantage of blinding himself to the results of the outcome analysis when creating the prognostic score and subsequent adjustments that are part of the pre-processing of the data. From a design perspective, it is advantageous for the estimation of prognostic scores to remain part of the pre-processing of the data and not use any outcome data from the treatment group. As has been discussed, adjustment methods for observational studies aim to establish comparability of groups prior to the assessment of treatment effects, and this objective would be lost if the estimation of prognostic scores is performed outside of a design stage. From an estimation perspective, fitting prognostic score models to the control group helps to estimate the effect of treatment on the treated units, without assuming a constant treatment effect across groups.

When the distributions of pretreatment covariates are similar across comparison groups within matched sets, the estimation of causal effects of treatment is improved. By determining the most prognostically relevant part of the covariate space, estimated prognostic scores can improve the fitting of the standard propensity score model. The present extension of the prognostic score allows the consideration of several different model specifications to create a multidimensional prognostic score. As a result, an analyst can consider a variety of models to establish the relationship between the responses for the control group and the corresponding covariates without determining that one is necessarily the true model. Diagnostics in the form of tests and plots to assess prognostic balance provide appraisals of the chosen dimensions of the prognostic score. These estimated prognostic scores can be included as predictors in a prognostic-propensity model with or without the presence of additional covariates for further covariate adjustment to provide an improved propensity score that is formed from covariates most relevant to an outcome of interest. In turn, the estimated prognostic-propensity scores can be used like standard propensity scores to aid in the design of observational studies. The routine of estimating a multidimensional prognostic score and the corresponding prognostic-propensity score offers an improvement to the standard pre-processing of observational data via propensity scores alone.

3.8 Appendix: Justification for balance checking on estimated prognostic scores

In the discussion of the prognostic-propensity score in Section 3.5, a standard notion of covariate distribution overlap across treatment groups is assessed with boxplots. For further study of this propensity balance on prognostic scores, an analyst could apply a version of the tests introduced in Hansen and Bowers (2008). Rather than assessing similarity in covariate distributions, an analyst would be assessing similarity in the distributions of the linear combinations of covariates deemed to be most relevant to the outcome of interest – the prognostic score dimensions. These tests could be conceptualized as randomization t-tests of the treatment coefficient in a regression of the covariate in question on the treatment variable and a fixed effect for pair or stratum membership.

Ideally, this comparability would be ascertained on the true prognostic scores, the exact values of combinations of covariates most relevant to the outcome, rather than estimates of these values. With a dimension fitted by a linear model, this is the difference between assessing similarities on dimensions defined by $X\beta_0$, where β_0 is the vector defining the true most prognostically relevant linear combination of covariates, rather than an assessment using $X\hat{\beta}$.

In Proposition III.1 and its proof, we establish the similarity between checking balance on an estimated and actual prognostic score via an asymptotic argument. We argue conditionally on the number of treated units per block, or $\sum_{i=1}^{n_b} Z_{bi} = m_b$, for $b \in [1, B]$, assuming there are n_b units in block b. Z, a random variable, is the vector of treatment assignments and is uniform on all {0,1}-valued vectors such that $\sum_{i=1}^{n_b} Z_{bi} = m_b$. x is a matrix of covariates, which are considered to be fixed values. The asymptotic argument is made not for a single sequence of observations, but for a sequence of experimental populations, $\nu = 1, 2, ...$ in which increasing ν corresponds to increasing the number of observations. This follows the structure of proof used in Hansen and Bowers (2008).

3.8.1 Definitions

Define

$$(3.3) \qquad d_Z(x_\nu \beta) = \frac{\sum_{b=1}^{B} h_{\nu b}\{(Z'_{\nu b} x_{\nu b}\beta/m_{\nu b}) - (1 - Z'_{\nu b})x_{\nu b}\beta/(n_{\nu b} - m_{\nu b})\}}{\sum_{b=1}^{B} h_{\nu b}}$$

where $h_{\nu b}$ is the harmonic mean of $n_{\nu b} - m_{\nu b}$ and $m_{\nu b}$, or, respectively, the number of control units and the number of treatment units in block b for population ν. Multiplying $h_{\nu b}$ through, it can be seen that

$$\begin{aligned} d_Z(x_\nu\beta) &= \frac{\sum_{b=1}^B h_{\nu b}\{(Z'_{\nu b}x_{\nu b}\beta/m_{\nu b}) - (1 - Z'_{\nu b})x_{\nu b}\beta/(n_{\nu b} - m_{\nu b})\}}{\sum_{b=1}^B h_{\nu b}} \\ &= \frac{\sum_{b=1}^B\{Z'_{\nu b}x_{\nu b}\beta - \frac{m_{\nu b}}{n_{\nu b}}(Z'_{\nu b}x_{\nu b} + (1 - Z'_{\nu b})x_{\nu b})\beta\}}{\sum_{b=1}^B h_{\nu b}} \\ &= \frac{\sum_{b=1}^B\{Z'_{\nu b}x_{\nu b}\beta - m_{\nu b}\frac{\sum_{i=1}^{n_{\nu b}} x_{bi}}{n_{\nu b}}\beta\}}{\sum_{b=1}^B h_{\nu b}} \\ &= \frac{\sum_{b=1}^B\{Z'_{\nu b}x_{\nu b}\beta - \mathbf{E}(Z'_{\nu b}x_{\nu b}\beta)\}}{\sum_{b=1}^B h_{\nu b}} \end{aligned}$$

Define $S_{\nu bxx}$ as the as the sample covariance matrix of x_ν in block b. Then,

$$\mathrm{V}_0(\beta) = \frac{\beta'(\sum_{b=1}^B h_{\nu b}S_{\nu bxx})\beta}{(\sum_{b=1}^B h_{\nu b})^2} \tag{3.4}$$

If $\hat{\beta} = 0$, define $\frac{\mathrm{V}(\beta_0)}{\mathrm{V}(\hat{\beta})} = 0$.

For fixed β, $\mathrm{V}(d_Z(x_\nu\beta)) = \mathrm{V}_0(\beta)$.

$$\begin{aligned} \mathrm{V}(d_Z(x_\nu\beta)) &= \mathrm{V}\left(\frac{\sum_{b=1}^B\{Z'_{\nu b}x_{\nu b}\beta - \mathbf{E}(Z'_{\nu b}x_{\nu b}\beta)\}}{\sum_{b=1}^B h_{\nu b}}\right) \\ &= \frac{\mathrm{V}(\sum_{b=1}^B\{Z'_{\nu b}x_{\nu b}\beta\})}{(\sum_{b=1}^B h_{\nu b})^2} \\ &= \frac{\beta'\mathrm{V}(\sum_{b=1}^B\{Z'_{\nu b}x_{\nu b}\})\beta}{(\sum_{b=1}^B h_{\nu b})^2} \\ &= \frac{\beta'(\sum_{b=1}^B h_{\nu b}S_{\nu bxx})\beta}{(\sum_{b=1}^B h_{\nu b})^2} \\ &= \mathrm{V}_0(\beta) \end{aligned}$$

Proposition III.1. *For measures of balance denoted by $d_Z(x\beta)$, it can be shown that*

$$\frac{d_Z(x_\nu\hat{\beta})}{\mathrm{V}_0^{1/2}(\hat{\beta})} - \frac{d_Z(x_\nu\beta_0)}{\mathrm{V}_0^{1/2}(\beta_0)} = O_p(\nu^{-1/2}) \tag{3.5}$$

3.8.2 Conditions

1. x is a matrix of finite constants with a fixed number of columns (described by p).
2. Over all populations indexed by ν, $\frac{m_{\nu b}}{n_{\nu b}}$, or the fraction of units in the treatment group in block b, is bounded away from 0 and 1.
3. For $S_{\nu bxx}$ as defined previously, we assume $\frac{\sum_{b=1}^{B} h_{\nu b} S_{\nu bxx}}{(\sum_{b=1}^{B} h_{\nu b})^2} \rightarrow_p S_{xx}$ as $\nu \rightarrow \infty$, where S_{xx} is non-negative definite.
4. β_0 is finite and $\beta_0{}' S_{XX} \beta_0 > 0$.
5. Suppose $f(y_i, x_i, \beta)$ is a possible likelihood for the data (though possibly misspecified) and $logf(y_i, x_i, \beta)$ is differentiable. Define $\hat{Q}_\nu(\beta) = \nu^{-1} \sum_{i=1}^{\nu} logf(y_i, x_i, \beta)$ and let $\hat{\beta}_\nu$ be the maximizer of $\hat{Q}_\nu(\beta)$. Assume $\hat{Q}_\nu(\beta)$ is a concave function, a property that can be established for the log-likelihoods of many ordinary regressions and generalized linear models.

   If we suppose there exists a function $Q_0(\beta)$ such that $\hat{Q}_\nu(\beta) \rightarrow_p Q_0 \beta$ for all β. Define β_0 to be the unique maximizer of $Q_0(\beta)$, where β_0 is an element of the interior of the convex set that describes the parameter space. Then, $\hat{\beta}_\nu$ exists with probability approaching one and $\hat{\beta}_\nu \rightarrow \beta_0$ as $\nu \rightarrow \infty$.

3.8.3 Proof

We can begin by writing

$$\frac{d_Z(x_\nu \hat{\beta}_\nu)}{\mathrm{V}_0(\hat{\beta}_\nu)} = \left\{ \frac{d_Z(x_\nu(\hat{\beta}_\nu - \beta_0))}{\mathrm{V}_0^{1/2}(\beta_0)} + \frac{d_Z(x_\nu \beta_0)}{\mathrm{V}_0^{1/2}(\beta_0)} \right\} \sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}}$$

which implies the LHS of Equation 3.5 can be written as

$$\frac{d_Z(x_\nu(\hat{\beta}_\nu - \beta_0))}{\mathrm{V}_0^{1/2}(\beta_0)} \sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}} + \frac{d_Z(x_\nu \beta_0)}{\mathrm{V}_0^{1/2}(\beta_0)} \left(\sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}} - 1 \right)$$

First, it can be shown that as $\nu \rightarrow \infty$

$$\sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}} \rightarrow_p 1$$

By the continuous mapping theorem, it is sufficient to show in the above that the square of the LHS converges to the square of the RHS.

By definition,

$$\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)} = \frac{\beta_0{}'(\sum_{b=1}^{B} h_{\nu b} S_{\nu bxx})\beta_0/(\sum_{b=1}^{B} h_{\nu b})^2}{\hat{\beta}_\nu'(\sum_{b=1}^{B} h_{\nu b} S_{\nu bxx})\hat{\beta}_\nu/(\sum_{b=1}^{B} h_{\nu b})^2}$$

By conditions 3, 4, 5, and another application of the continuous mapping theorem,

$$\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)} = \frac{\beta_0{}'(\sum_{b=1}^{B} h_{\nu b} S_{\nu bxx})\beta_0}{\hat{\beta}_\nu'(\sum_{b=1}^{B} h_{\nu b} S_{\nu bxx})\hat{\beta}_\nu} \rightarrow_p \frac{\beta_0{}' S_{xx}\beta_0}{\beta_0{}' S_{xx}\beta_0} = 1$$

as $\nu \rightarrow \infty$.

Thus, $\sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}} \rightarrow_p 1$ and $\left(\sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}}\right) - 1 \rightarrow_p 0$.

Hansen and Bowers (2008), Section 3.2, summarizes conditions under which we can establish the following asymptotic distribution result, taking weights defined as $w_b = \frac{h_b}{\sum_{b=1}^{B} h_b}$ (normalized harmonic mean weights):

$$\frac{d_Z(x_\nu\beta_0)}{\mathrm{V}_0^{1/2}(\beta_0)} = \frac{d_Z(x_\nu\beta_0)}{\mathrm{V}^{1/2}(d_Z(x_\nu\beta_0))} \rightarrow_d N(0,1)$$

as $\nu \rightarrow \infty$.

By an application of Slutsky's Theorem, as $\nu \rightarrow \infty$:

$$\frac{d_Z(x_\nu\beta_0)}{\mathrm{V}_0^{1/2}(\beta_0)}\left(\sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}} - 1\right) \rightarrow_d 0$$

It remains to show

$$\frac{d_Z(x_\nu(\hat{\beta}_\nu - \beta_0))}{\mathrm{V}_0^{1/2}(\beta_0)}\sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}} = O_p(\nu^{-1/2})$$

Write

$$\begin{aligned}\frac{d_Z(x_\nu(\hat{\beta}_\nu-\beta_0))}{\mathrm{V}_0^{1/2}(\beta_0)} &= \frac{d_Z(x_\nu(\hat{\beta}_\nu-\beta_0))}{\mathrm{V}^{1/2}(x_\nu\beta_0)}\\ &= \frac{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}-\mathbf{E}(Z'_{\nu b}x_{\nu b}))}{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\beta_0-\mathbf{E}(Z'_{\nu b}x_{\nu b}\beta_0))\}}(\hat{\beta}_\nu-\beta_0)\end{aligned}$$

$\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}-\mathbf{E}(Z'_{\nu b}x_{\nu b}))\mathrm{V}^{-1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\beta_0-\mathbf{E}(Z'_{\nu b}x_{\nu b}\beta_0))\}$ is a $1\times p$ vector for a given value of ν or a $\nu\times p$ matrix for all ν. For any $k\in[1,p]$, we can define $\tilde{\beta}$ such that the kth term of $\tilde{\beta}$ equals 1, and for all i not equal to k, the ith term of $\tilde{\beta}$ equals 0. Thus, $x_k=x\tilde{\beta}$ and $x_{bk}=x_b\tilde{\beta}$.

For a selected k, terms of a column of length ν of this $\nu\times p$ matrix are defined:

$$\frac{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu bk}-\mathbf{E}(Z'_{\nu b}x_{\nu bk}))}{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu bk}-\mathbf{E}(Z'_{\nu b}x_{\nu bk})\}}\frac{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\tilde{\beta}-\mathbf{E}(Z'_{\nu b}x_{\nu b}\tilde{\beta}))\}}{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\beta_0-\mathbf{E}(Z'_{\nu b}x_{\nu b}\beta_0))\}}$$

Noting $\mathbf{E}(\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu bk}-\mathbf{E}(Z'_{\nu b}x_{\nu bk})))=0$, by the Central Limit Theorem,

$$\frac{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu bk}-\mathbf{E}(Z'_{\nu b}x_{\nu bk}))}{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu bk}-\mathbf{E}(Z'_{\nu b}x_{\nu bk})\}}\rightarrow_d N(0,1)$$

as $\nu\rightarrow\infty$.

For all $k\in[1,p]$, define σ_k as a positive constant such that $\tilde{\beta}'S_{xx}\tilde{\beta}/\beta_0{}'S_{xx}\beta_0={\sigma_k}^2$. We can show that as $\nu\rightarrow\infty$,

$$\frac{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\tilde{\beta}-\mathbf{E}(Z'_{\nu b}x_{\nu b}\tilde{\beta}))\}}{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\beta_0-\mathbf{E}(Z'_{\nu b}x_{\nu b}\beta_0))\}}\rightarrow_p \sigma_k$$

It is equivalent to show its square converges to ${\sigma_k}^2$. The square of the LHS can be written as

$$\begin{aligned}
\left(\frac{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\tilde{\beta}-\mathbf{E}(Z'_{\nu b}x_{\nu b}\tilde{\beta}))\}}{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\beta_0-\mathbf{E}(Z'_{\nu b}x_{\nu b}\beta_0))\}}\right)^2 &= \frac{\mathrm{V}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\tilde{\beta}-\mathbf{E}(Z'_{\nu b}x_{\nu b}\tilde{\beta}))\}}{\mathrm{V}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\beta_0-\mathbf{E}(Z'_{\nu b}x_{\nu b}\beta_0))\}} \\
&= \frac{\tilde{\beta}'[\sum_{b=1}^{B}h_{\nu b}S_{\nu bxx}]\tilde{\beta}}{{\beta_0}'[\sum_{b=1}^{B}h_{\nu b}S_{\nu bxx}]\beta_0} \\
&\rightarrow_p \frac{\tilde{\beta}'S_{xx}\tilde{\beta}}{{\beta_0}'S_{xx}\beta_0} \\
&= {\sigma_k}^2
\end{aligned}$$

Combining previous results, for some $k \in [1,p]$ with Slutsky's Theorem, as $\nu \rightarrow_p \infty$, it can be established that each column of

$$\frac{d_Z(x_\nu)}{\mathrm{V}^{1/2}(d_Z(x_\nu\beta_0))}\sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}}$$

converges to a normal distribution:

$$\frac{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu bk}-\mathbf{E}(Z'_{\nu b}x_{\nu bk}))}{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\beta_0-\mathbf{E}(Z'_{\nu b}x_{\nu b}\beta_0))\}}\sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}} \rightarrow_d N(0,{\sigma_k}^2)$$

which implies

$$\frac{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu bk}-\mathbf{E}(Z'_{\nu b}x_{\nu bk}))}{\mathrm{V}^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\beta_0-\mathbf{E}(Z'_{\nu b}x_{\nu b}\beta_0))\}}\sqrt{\frac{\mathrm{V}_0(\beta_0)}{\mathrm{V}_0(\hat{\beta}_\nu)}} = O_p(1)$$

By the Central Limit Theorem and results for general regression models,

$$\sqrt{\nu}(\hat{\beta}_\nu-\beta_0) \rightarrow_d N(0,A^{-1}BA^{-1})$$

where $B=\mathbf{E}[\epsilon^2x'x]$, where $\epsilon = Y - x'\beta_0$ and $A=\mathbf{E}[x'x]$.

Thus, $(\hat{\beta}_\nu-\beta_0)=O_p(\nu^{-1/2})$. It also follows that for any $k \in [1,p]$, we can look at any term k in both vectors $\hat{\beta}_\nu$ andβ_0 and see that

$$\sqrt{\nu}(\hat{\beta}_{\nu k}-\beta_{0k}) \rightarrow_d N(0,A^{-1}BA^{-1}{}_{[k,k]})$$

implying $(\hat{\beta}_{\nu k} - \beta_{0k}) = O_p(\nu^{-1/2})$.

If we notice

$$\frac{d_Z(x_\nu(\hat{\beta}_\nu - \beta_0))}{V_0^{1/2}(\beta_0)}\sqrt{\frac{V_0(\beta_0)}{V_0(\hat{\beta}_\nu)}}$$

is equivalent to

$$\sum_{k=1}^{p}(\hat{\beta}_{\nu k} - \beta_{0k})\frac{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu bk} - \mathbf{E}(Z'_{\nu b}x_{\nu bk}))}{V^{1/2}\{\sum_{b=1}^{B}(Z'_{\nu b}x_{\nu b}\beta_0 - \mathbf{E}(Z'_{\nu b}x_{\nu b}\beta_0))\}}\sqrt{\frac{V_0(\beta_0)}{V_0(\hat{\beta}_\nu)}} \tag{3.6}$$

and also, by previous results, each of the k terms in (3.6) is $O_p(\nu^{-1/2})$. Thus, the sum of the k terms is also $O_p(\nu^{-1/2})$, which ultimately implies (3.5).

CHAPTER IV

The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder

4.1 Overview

Omitted variable bias can affect treatment effect estimates obtained from observational data due to the lack of random assignment to treatment groups. Sensitivity analyses adjust these estimates to quantify the impact of potential omitted variables. This paper presents methods of sensitivity analysis to adjust interval estimates of treatment effect — both the point estimate and standard error — obtained using multiple linear regression. Taking into account the impact an omitted variable could have on the standard error of the treatment effect is important to inferences obtained from multiple regression; the narrowing or widening of a confidence interval can alter inferences made about treatment in a different manner than a shift in the point estimate. The methods are demonstrated on data from Connors et al.'s (1996) study of right heart catheterization and health outcomes.

4.2 Introduction

In analyzing data to determine the causal effect of a treatment with a multiple linear regression model, the estimated coefficient of the treatment assignment indicator is often taken to be an estimate of the treatment effect. This approach is valid provided the standard linear model assumptions are met and that there are no unaccounted for differences in the treatment and control groups. In the case of an observational study in which there is no random assignment to treatment, there is potential for an unmeasured confounder to exist that would bias the effect estimate as measured by the fitted regression coefficient. Adjusting for as many covariates as an analyst has available can greatly reduce the potential impact of omitting a single variable, but the possibility still exists that the omission could still have a considerable effect. Rather than viewing the conclusions from observational data analyzed with multiple regression as invalid, a better approach is to quantify the potential change an omitted confounder could have on the treatment effect estimate if it were able to be measured and included in the regression model. For this purpose, a sensitivity analysis can be formulated.

Methods of sensitivity analysis quantify the degree of bias from omitted confounders that would change or nullify conclusions of a study. Types of sensitivity analyses for various models and data structures have been discussed by Cornfield et al. (1959), Seber (1977), Rosenbaum and Rubin (1983), Rosenbaum (1988), Rosenbaum (1991), and Copas and Li (1997), Robins et al. (2000), Marcus (1997), Lin et al. (1998), Frank (2000) and Imbens (2003), among others. Early literature on sensitivity analysis examines the effect of an omitted variable on point estimates. Much of the current literature on sensitivity analysis directly addresses the effect of omitted variables on inferences; however, these effects are examined in the context of relatively specialized methods. The methods of sensitivity analysis presented in this paper have many advantages over prior methods, but a key advantage is that our methods apply to one of the most widely used multivariate analytic techniques – multiple regression. The typical procedure of the existing methods for sensitivity analysis for the omission of a confounder in a multiple regression model aims to describe relationships between omitted variables and variables present in the model and use these relationships to adjust point estimates, confidence limits, or p-values. The method of sensitivity analysis presented here follows this familiar outline, but it has several advantages over the existing methods. The procedure proposed in this paper describes the relationship between omitted and measured confounders in terms intrinsic to multiple regression with two simple quantities (Section 4.3), so additional regression fits can calibrate intuition about these relationships in practice when the omitted confounder is unknown (Angrist and Imbens (2002) discuss this calibration in a related context). Two quantities are used so one can track associations between the omitted variable and the treatment assignment, while the other quantity can track the relationship with the response variable, controlling for all of the measured confounders. Depending on the confoundedness of an omitted variable with both the treatment and outcome, the interval estimate of treatment effect could shift at its center, widen, or narrow, or some combination of changing in width and point estimate. Importantly, our method, in contrast to many methods of sensitivity analysis, adjusts both the point estimate and standard error of the treatment effect, allowing for completely adjusted inferences. While sensitivity analyses are important to help understand observational data as best as we are able, sensitivity analyses for omitted variables are not restricted to observational data, sensitivity analyses are useful in any analysis in which there is a possibility for an unobserved variable to affect the results.

Section 4.2 continues with an introduction to a case study and associated data. Section 4.3 details the two key quantities underlying our sensitivity analysis and how they are calibrated and used to adjust the point estimate and standard error of treatment effect. Section 4.4 presents a key theoretical results underlying the modification of confidence limits to account for an omitted confounder. In Section 4.5, we show that our method is not limited only to the context given in the previous sections: it is readily adaptable to situations in which multiple confounders are omitted and situations in which an analyst desires to combine propensity score matching with ordinary multiple regression.

4.2.1 A case study

The methods introduced in this paper are applied to data from the much-debated Connors et al. (1996) study of a critical-care procedure known as pulmonary artery or right heart catheterization (RHC). RHC is a procedure that was first used in 1970 to perform continuous measurements of right heart pressures and blood flow in critically ill patients and became the standard in the treatment of such patients, without first being tested by a randomized controlled trial. Observational studies that preceded the Connors et al study (e.g. Gore et al. (1987) and Zion et al. (1990)) determined that RHC did not have the effects physicians expected; successful outcomes were not more likely in patients who received the RHC. These studies were criticized due to the potential of an omitted variable to lead to a systematic bias in the results; it may be that patients who received RHC differed in terms of this omitted variable from those whose care was not managed with the procedure. Connors et al. (1996) used a large sample and considerable adjustments for pre-treatment confounders found that not only did the procedure not appear to be beneficial, receiving the RHC procedure seemed to worsen patient outcomes. Since the Connors et al study, five randomized clinical trials have been conducted (Rhodes et al. (2002); Sandhan et al. (2003); Shah and Stevenson (2004); Richard and et al (2003); Harvey et al. (2005)), which supported most findings of the observational findings of Connors et al. (1996). The procedure is still widely used despite these findings. In our analysis, we quantify the effect an omitted confounder could have on the effect of RHC on the length of a patient's hospital stay.

4.2.2 The SUPPORT data

The data analyzed in Connors et al. (1996) is from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). The study collected data at five centers on the decision-making and outcomes of seriously ill, hospitalized adult patients. Patients included in the study had to meet certain entry criteria and particularly a predefined level of severity of illness. These patients were studied in two phases. Phase I was a prospective observational study of 4301 patients designed to describe the process of decision making and care of critically ill patients. Phase II was a cluster RCT of an intervention (the delivery of patient preferences regarding treatment to the physicians) to improve decision making and outcomes for severely ill patients. Intervention was found not to affect physicians' decisions or patient outcomes, so phase I and phase II data sets were combined and all 5735 SUPPORT patients who were admitted to an ICU in the first 24 hours of study were analyzed together, regardless of phase status.

For all SUPPORT patients, initial disease categories upon admission to the hospital were recorded (such as acute respiratory failure, chronic obstructive pulmonary disease, cirrhosis, colon cancer, etc.). In addition, measurements to assess physiological status were recorded throughout the hospital stay, and the present analysis makes use of these physiological measurements on the first day of the hospital stay. Demographic information was obtained from the patient or surrogate interviews. Patients were coded as having had RHC if it was performed within the first 24 hours of the study. As mentioned previously, the outcome of interest is the length of hospital stay; measured in days, it is the difference between the entrance to and exit from the hospital. The outcome variable, length of hospital stay, is right-skewed, so our analysis will work with a log-transformed version of the outcome variable.

4.2.3 Data preparation and preliminary analysis

Before estimating a treatment effect and applying our sensitivity analysis methods, we use variable selection with the RHC data to purposely omit some variables. After backward stepwise variable selection, 19 variables remain in the model and approximately 30 other variables can be considered “omitted” in future analyses. Stepwise variable selection often has the consequence of inflating standard errors (Faraway, 1992). To account for this potential inflation, we utilize a nonparametric bootstrap in which regressions are computed for bootstrapped samples to obtain a distribution of t-statistics of the estimated regression coefficients for the treatment. From this distribution, bootstrapped quantiles of the t distribution can be obtained. These quantiles will be used in future confidence intervals, though they are similar to the quantiles of the relevant t -distribution. From the regression model with the 19 covariates, our preliminary estimate of the effect of RHC on the length of hospital stay is about 0.11, which indicates that RHC lengthens hospital stays by about 12%. Confidence intervals on both the log scale and percent increase scale are presented in Table 4.1.

From this initial data analysis, the need to consider adjustments to an entire interval can be seen with the case of the variable DNR, which indicates the presence of a do-not-resuscitate order on file on the date of entry. The interval for treatment effect with DNR in the model is $(0.06, 0.16)$, but if DNR is excluded from the model, the interval changes to (0.09, 0.20). A single variable has the potential to shift both the point estimate and interval width, indicating the need to consider both in a

Table 4.1: Effect of RHC on length of stay: 95% confidence intervals

	CI on log-scale	CI for percent increase
Effect of RHC on length of hospital stay	(0.06, 0.16)	(6%, 18%)

method of sensitivity analysis.

4.3 Effect estimates accounting for omitted variables

To understand how an omitted variable, W, would affect the coefficient on the treatment variable, Z, in a hypothetical regression of a given outcome on these and other variables,

$$Y = \alpha \mathbf{X}^T + \beta Z + \delta W + e\,, \tag{4.1}$$

we can begin by examining how included variables affect the treatment coefficient in the regression that was actually performed,

$$Y = a\mathbf{X}^T + bZ + e. \tag{4.2}$$

The current section demonstrates this process and its uses in sensitivity analyses.

4.3.1 Sensitivity Zone

To relate interval estimates for β to those for b, we quantitatively describe the omitted variable, W, in terms of its relationship with both the treatment variable and outcome variable conditional on the other included covariates. The extent to which a variable is confounded with the treatment variable is of central importance to assessing omitted variable bias. If a regression of the treatment assignment Z on all of the confounders $\mathbf{X}$ is fit, each of the t-statistics associated with the regression coefficients provide a measure of association between Z and a given confounder. To obtain a measure of the relationship between Z and W, a similar process can be used. Rather than regressing Z on $\mathbf{X}$ alone, Z is regressed on $\mathbf{X}$ and W, and the t-statistic associated with W, denoted t_W, is taken to be a measure of what we will call treatment-confoundedness. In the case of a non-continuous Z – a binary Z, for example– a nonlinear model may have to be used to allow the t-statistics to serve as inferential statistics, however, in the sensitivity analysis, the t-statistics from a linear model are used for purely descriptive purposes.

To assess a confounder's impact on a regression model, it is typical to compare values of R^2 for models with and without the specific confounder. Comparing R^2 values allows an estimation of the association between the outcome variable and the particular confounder of interest, conditional on all other confounders in the model. Similarly, as R^2 captures the proportion of variation explained, one could assess the impact of a particular variable by comparing the proportion of variation not explained, or $1 - R^2$. To capture the association between an omitted variable and an outcome variable, define $\rho^2_{y\cdot w|z\mathbf{x}}$ as $[(1 - R^2_{\text{no}\,W}) - (1 - R^2_{\text{with}\,W})]/(1 - R^2_{\text{no}\,W})$, or the proportionate reduction in unexplained variance when W is added as a regressor (Christensen, 1996, Ch. 6). For a sensitivity analysis, we require values of both t_W and $\rho^2_{y\cdot w|z\mathbf{x}}$, and the collection of permissible values for $(t_W, \rho^2_{y\cdot w|z\mathbf{x}})$ will be referred to as the sensitivity zone, following Small (2007).

If each of the 19 regressors in the model are placed in the role of W one at a time and "omitted" from the model, values of t_W and $\rho^2_{y\cdot w|z\mathbf{x}}$ can be obtained for each of these variables. Table 4.2 presents these results, indicating the variability in the possible associations of W with the treatment variable and of W with the response variable, conditional on the other included covariates. Some variables, such as the PaO2/FIO2 ratio, are highly confounded with treatment ($t_W = 15.2$), but have a small relationship with the response ($\rho^2_{y\cdot w|z\mathbf{x}} = 0.1\%$). Other variables, such as the DNR order, have a substantially stronger association with the outcome ($\rho^2_{y \cdot w|z\mathbf{x}} = 3.3\%$), but slightly weaker, though still relatively strong, relationship with the treatment variable ($t_W = 6.5$). Many variables – such as upper GI bleed, white blood cell count, body temperature, and heart rate – have little relationship to either treatment or response in comparison to the other variables present in the model. Proposition IV.1 and Proposition IV.2 will illuminate the relative importance of these variations among the confounders and their corresponding impact on the bias and standard error of the treatment effect.

Table 4.2: Selected included covariates' relation to treatment and response variables, given remaining included variables

	confounding with RHC ($\lvert t_W \rvert$) (rounded)	% decrease in unexplained variation by adding W ($100\rho^2_{y \cdot w\|z\mathbf{x}}$)
income	6.8	0.3
initial disease cat. 1	48.1	3.4
initial disease cat. 2	20.2	0.8
Comorbidities illness:		
renal disease	2.1	0.2
Upper GI bleed	0.7	0.1
Day 1 Measurements:		
APACHE score	5.1	0.1
white blood cell ct	0.5	0.0
heart rate	2.5	0.0
temperature	2.3	0.1
PaO2/FIO2	15.4	0.1
albumin	2.3	0.7
hematocrit	3.3	0.9
bilirubin	2.2	0.1
sodium	3.1	0.1
PaCo2	6.8	0.2
DNR	6.5	3.3
PH	3.7	0.3
Admit Diagnosis Categories:		
neurology	5.9	0.2
hematology	3.6	0.1

4.3.2 Benchmarking the sensitivity zone

If W is an omitted variable, how can t_W and $\rho^2_{y \cdot w|z\mathbf{x}}$ be computed, as they seem to rely on knowledge about W? An analyst must use information from the measured variables to guide intuition about potential bounds for the sensitivity zone; this process will be referred to as benchmarking the sensitivity zone. In the present study of the RHC data, we have access to a collection of variables omitted from the stepwise procedure, a procedure which eliminates variables from the model on the basis of their conditional relationship with the outcome. To calibrate intuitions about the treatment-confoundedness t_W, these deliberately omitted variables can be used, while variables present in the model can guide ideas about the value of $\rho^2_{y \cdot w|z\mathbf{x}}$.

To determine a bound for t_W, we use the variables omitted by the stepwise procedure, a process that should have little to do with the treatment-confoundedness of the variables. By placing these variables in the role of W, t_W values can be obtained for each of these variables as they were obtained for the included variables in Table 4.2. For example, for the average blood pressure on the day of entry, $t_W = 8.6$, while for an indicator variable of a patient's immunosuppresion, $t_W = 0.4$. The range of values of t_W obtained by these regressors offers an idea of a typical range of such values for variables in this application. If other data were available, values for t_W could be obtained by performing similar calculations for variables of interest in an external data set.

An analyst could select bounds for $\rho^2_{y \cdot w|z\mathbf{x}}$ with substantive knowledge of how a certain W may plausibly affect R^2 if such information were available. In its absence, bounds for $\rho^2_{y \cdot w|z\mathbf{x}}$ can be benchmarked using variables in the model. When the covariates included in our model are placed in the role of W one at a time (see Table 4.2), 17 of the 19 regressors yield values of $\rho^2_{y \cdot w|z\mathbf{x}}$ of less than 1%. Only two variables have corresponding $\rho^2_{y\cdot w|z\mathbf{x}}$ values of approximately 3%, and only when placing all of the included covariates collectively in the role of W – fitting an intercept only model – does $\rho^2_{y\cdot w|z\mathbf{x}}$ come close to 10%. Based on this information, for the purposes of our sensitivity analysis, we consider $\rho^2_{y\cdot w|z\mathbf{x}} \leq .01$ to be a reasonable bound and $\rho^2_{y\cdot w|z\mathbf{x}} \leq .10$ to be a conservative upper bound.

4.3.3 Relating the sensitivity zone to bias and standard error

If b is the coefficient of treatment in the absence of W and β is the coefficient of treatment if W is included in the model, then the bias in the effect estimate due to W is given by Proposition IV.1.

Proposition IV.1. *If* $R^2_{y\cdot z\mathbf{x}} < 1$ *and* t_W *is finite, then*

$$b - \beta = \mathrm{SE}(b) t_W \rho_{y\cdot w|z\mathbf{x}}. \tag{4.3}$$

A proof is given in Section 4.4.2.

Bias due to the omission of W is the product of the standard error of the estimated Z-coefficient, the treatment-confoundedness of W, and the value of $\rho_{y\cdot w|z\mathbf{x}}$. As $|\rho_{y\cdot w|z\mathbf{x}}| \leq 1$, the value of t_W can potentially have a much larger impact on the bias; however, the effect of even a very large t_W will negligible if $\rho_{y\cdot w|z\mathbf{x}}$ is very small. We can illustrate this notion by considering some of the variables previously mentioned. Recall that PaO2/FIO2 ratio is highly confounded with treatment but has a small association with the outcome, DNR order has a moderately large value of the treatment-confoundedness but is – relatively speaking – highly associated with the outcome variable, and the presence of an upper GI bleed has a very small association with both treatment and outcome. Consequently, the biases from omitting PaO2/FIO2, DNR order, and upper GI bleed are approximately 0.012, 0.030, and 0, respectively (or, about half a standard error, one standard error, and zero standard errors, respectively), indicating the relative contributions of the values of t_W and $\rho_{y\cdot w|z\mathbf{x}}$ to the bias calculation. While large values of treatment-confounding lead to bias, the magnitude of the bias is both a function of the treatment confoundedness and the proportionate reduction in unexplained variation produced by adding W.

If sensitivity analysis only focuses on omitted variable bias, part of the effect of an omitted variable can be overlooked. Table 4.3 presents the effects of excluding several of the variables included in the length of stay model of the RHC data in the second column ($df = 5700$). With the $t_W = 0.7$ and $\rho^2_{y\cdot w|z\mathbf{x}} = 0.1\%$, the variable upper GI bleed had little effect on bias, and Table 4.3 illustrates that omitting or including upper GI bleed has no effect on the standard error of the treatment effect. In contrast, excluding the variable DNR order increases the standard error (from 0.0260 to 0.0264), while excluding PaO2/FIO2 decreases standard error (from 0.0260 to 0.0255). The changes to standard error are more noticeable for non-negligible values of t_W and $\rho^2_{y\cdot w|z\mathbf{x}}$. In the column corresponding to the RHC study, in which $df = 5700$, the changes in the standard errors due to variable exclusion are moderated by the large sample size. If, instead, a smaller sample size is considered and the corresponding $df = 50$, the effects on the standard errors are much more substantial. Adjusting for PaO2/FIO2 in the small sample size scenario more than doubles SE(β), and this change could greatly alter subsequent inferences. It appears that assessing the sensitivity of the standard error estimate to the omission of a confounder is crucial with moderate or small samples.

The findings illustrated in Table 4.3 are reflected in Proposition IV.2, which formally demonstrates how to adjust the standard error estimates of b to align them with those of β.

Table 4.3: RHC-coefficient and its standard error after removing specified variables.

Excluded		Standard Error	
Variable	Estimate	df=5700	df = 50
no exclusion	0.112	0.0260	0.278
DNR order	0.143	0.0264	0.206
GI bleed	0.112	0.0260	0.274
PaO2/FIO2	0.122	0.0255	0.115

Proposition IV.2. *If $R^2_{y\cdot z\mathbf{x}} < 1$ and t_W is finite, then*

$$\mathrm{SE}(\beta) = \mathrm{SE}(b)\sqrt{\frac{df + {t_W}^2}{df - 1}}\sqrt{1 - {\rho_{y\cdot w|z\mathbf{x}}}^2}. \tag{4.4}$$

Here $df = n - \mathrm{rank}(X) - 1$, the residual degrees of freedom after Y is regressed on X and Z.

Proposition IV.2 will be strengthened in Proposition IV.6, which is proved in the Appendix.

Similar to the bias of β, the standard error of β is both a function of $|\rho_{y\cdot w|z\mathbf{x}}|$ and $|t_W|$, but the two values of the sensitivity zone act in opposite directions on $\mathrm{SE}(\beta)$. Proposition IV.2 uses the term $\sqrt{1 - {\rho_{y\cdot w|z\mathbf{x}}}^2}$ as opposed to $\rho_{y\cdot w|z\mathbf{x}}$. This explains why a variable like PaO2/FIO2 could decrease $\mathrm{SE}(\beta)$, while another variable like DNR could increase the standard error. The *df* term in Proposition IV.2 leads to the sample size effects shown in Table 4.3. With a large sample size, a large value of *df* is an obvious consequence, and the effects of t_W and $\rho_{y\cdot w|z\mathbf{x}}$ on the standard error are overpowered by the sample size effects.

4.4 Sensitivity Intervals

Propositions IV.1 and IV.2 can be combined to provide a closed-form expression for the adjusted interval estimate of the treatment effect of Z on an outcome Y for a potential omitted variable W, controlling for all measured confounders $\mathbf{X}$. The desired interval estimates are of the form $\beta \pm q\mathrm{SE}(\beta)$, and revolve around the estimated quantities b, SE(b), and our sensitivity zone $(t_W, \rho_{y\cdot w|z\mathbf{x}})$. Bounds are placed on the values in the sensitivity zone such that we can obtain a union of intervals for $t_W \leq T$ and $\rho^2_{y\cdot w|z\mathbf{x}} \leq D$, for non-negative T and D. This union of intervals results in an expression for a single adjusted interval we call the sensitivity interval following Rosenbaum (2002). Proposition IV.3 describes how sensitivity zones map to sensitivity intervals.

For $\rho^2 \in [0, 1]$ and $df_w > 0$ define f by

$$(4.5) \qquad f(t, \rho^2, q, df_w) = \begin{cases} t|\rho| + q\sqrt{(1 + \frac{1}{df_w}(1+t^2))(1-\rho^2)}, & \rho^2 \leq g(t, q, df_w) \\ \sqrt{t^2 + q^2 \cdot (1 + \frac{1}{df_w}(1+t^2))}, & \rho^2 > g(t, q, df_w); \end{cases}$$

where $g(t, q, df_w) = \sin^2\{\arctan[tq^{-1}(1 + {df_w}^{-1}(1+t^2))^{-1/2}]\}$.

Proposition IV.3. *Let Y, $\mathbf{X}$, Z and W be as in (4.1) and (4.2), with both regressions fit either by ordinary least squares or by weighted least squares with common weights. Let $\rho_{y\cdot w|z\mathbf{x}}$ and t_W be as defined in § 4.3.1; suppose $R^2_{y\cdot z\mathbf{x}} < 1$, $|t_W| < \infty$. Fix $q > 0$. Then*

$$(4.6) \qquad \beta \pm qSE(\beta) \subset b \pm f(|t_W|, \rho^2_{y\cdot w|z\mathbf{x}}, df_w, q)SE(b)$$

where f is given by (4.5) and $df_w = df - 1$ is the residual degrees of freedom in the regression of Y on $\mathbf{X}$, Z, and W.

The proof of Proposition IV.3 follows later in this section.

4.4.1 Applying Proposition IV.3

As detailed in Section 4.3.2, a process called benchmarking can help an analyst select values of $\rho^2_{y\cdot w|z\mathbf{x}}$ and t_W that are reasonable for a given application. For studying the effects of RHC on length of stay, bounds for $\rho^2_{y\cdot w|z\mathbf{x}}$ of 1% and 10% are selected on the basis of values in Table 4.2. It is also instructive to consider situations in which no bound can be put on $\rho^2_{y\cdot w|z\mathbf{x}}$, but these would be highly pessimistic sensitivity analyses. Values of t_W are chosen to reflect the magnitude of t_W values that can be computed for variables omitted from the stepwise elimination. Table 4.4 presents sensitivity intervals for hypothetical omitted variables in which the hypothetical variables have t_W values of the confounders listed in the table, and the selected bounds for $\rho^2_{y\cdot w|z\mathbf{x}}$.

Table 4.4: 95% sensitivity intervals for the treatment coefficient, with the putative unobserved variable's treatment-confounding ($|t_W|$) hypothesized to be no greater than the treatment-confounding of 6 deliberately omitted variables. The decrease it would bring to the variance of response residuals is hypothesized to be no greater than either of 2 index values, 1% and 10%, or is not restricted.

	Treatment		% decrease in unexplained variation		
	confounding			$(100\rho^2_{y\cdot w|z\mathbf{x}})$	
Variable	benchmark		1%	10%	Unrestricted
Insurance class	12.2	most	(0.03, 0.20)	(-0.04, 0.26)	(-0.21, 0.43)
Respiratory eval.	8.9	some	(0.04, 0.19)	(-0.01, 0.23)	(-0.12, 0.35)
Mean blood press.	8.6	some	(0.04, 0.19)	(-0.01, 0.23)	(-0.12, 0.34)
Cardiovascular eval.	8.5	some	(0.04, 0.19)	(-0.01, 0.23)	(-0.11, 0.34)
Weight (kg)	6.1	some	(0.04, 0.18)	(0.01 , 0.21)	(-0.05, 0.28)
Immunosuppression	0.4	least	(0.06, 0.16)	(0.06 , 0.16)	(0.06 , 0.16)

In Table 4.4, Proposition IV.3 is applied, allowing the contributions of t_W and $\rho^2_{y\cdot w|z\mathbf{x}}$ values to the sensitivity interval to be jointly examined. The highly pessimistic case, in which no bound is put on $\rho^2_{y\cdot w|z\mathbf{x}}$, leads to interval estimates that are highly sensitive to omitted confounders except in the case of very low levels of treatment confounding. Putting a bound of 1% or 10% on $\rho^2_{y\cdot w|z\mathbf{x}}$, the interval estimates are resistant to the omission of a confounder in cases where treatment-confoundedness is similar to that of most of the variables included in the model. For variables with larger levels of treatment confounding, those with t_W values similar to insurance class or mean blood pressure, omitting such a variable noticeably changes the interval estimate of treatment effect.

4.4.2 Theoretical results underlying sensitivity formulas

Propositions IV.1, IV.2, and IV.3 extend better-known descriptions of bias in regression coefficients' point estimates due to variable omission (*e.g.*, Seber, 1977, p.66) to interval estimates. Much of the earlier literature on variable omission considers only adjustment to the point estimate but not the standard error. Developments from the pre-computer era offer formulas for the numerical adjustment of multiple regression results for the addition or removal of a covariate in place of refiguring the entire regression (Cochran, 1938). In this section, the proof of Proposition IV.1, which relates to these earlier developments, and the proof of Proposition IV.3 are presented.

Consider $\mathbf{X}$ to be a matrix containing a column of 1's (or columns from which a column of 1's can be recovered as a linear combination) and let Y, Z, and W be column vectors of common length, equal to the number of rows of $\mathbf{X}$. An inner product is defined as $(A, B) := \sum w_i a_i b_i / \sum w_i$, where w_i is a quadratic weight for the ith observation (in the case of unweighted least squares regression, $w_i \equiv 1$). Write $\mathbf{1}$ for the column vector of 1s. For vectors A, B, and C, let $\mathrm{Pj}(A|B, C)$ represent the projection of A into the subspace spanned by B and C. Variances and covariances are defined as follows: $\sigma_{ab\cdot c} := (A - \mathrm{Pj}(A|C), B - \mathrm{Pj}(B|C))$, $\sigma^2_{a\cdot c} = \sigma_{aa\cdot c}$;$\sigma_{ab} = \sigma_{ab\cdot\mathbf{1}}$, $\sigma^2_a = \sigma^2_{a\cdot\mathbf{1}}$. Partial correlations are then given as: $\rho_{ab} := \sigma_{ab}/(\sigma_a\sigma_b)$; $\rho_{ab\cdot c} := \sigma_{ab\cdot c}/(\sigma_{a\cdot c}\sigma_{b\cdot c})$. Denote the degrees of freedom available for estimating b as $df = n - m - 1$, where $m = \text{column.rank}(X)$. The nominal standard error estimates for $\hat{b}$ and $\hat{\beta}$ (*cf.* (4.2) and (4.1)) are then

$$SE(b) = \frac{df^{-1/2}\sigma_{y\cdot z\mathbf{x}}}{\sigma_{z\cdot\mathbf{x}}}, \qquad SE(\beta) = \frac{(df - 1)^{-1/2}\sigma_{y\cdot z\mathbf{xw}}}{\sigma_{z\cdot\mathbf{xw}}} \tag{4.7}$$

Proof of Proposition IV.1. To show $b - \beta = \mathrm{SE}(b) t_W \rho_{y\cdot w|z\mathbf{x}}$, write

$$\mathrm{Pj}(W|Z, \mathbf{X}) =: B^* Z + C^* \mathbf{X}^t. \tag{4.8}$$

Using (4.2), (4.1), and (4.8), the bias in the treatment coefficient can be written as $\hat{b} - \hat{\beta} = B^* \hat{\delta}$, a well-known result (Seber, 1977, p. 66).

Write $W^{\perp_{\mathbf{x}}}$ for $W - \mathrm{Pj}(W|\mathbf{1}, \mathbf{X})$, $Z^{\perp_{\mathbf{x}}}$ for $Z - \mathrm{Pj}(Z|\mathbf{1}, \mathbf{X})$, $Y^{\perp_{z\mathbf{x}}}$ for $Y - \mathrm{Pj}(Y|\mathbf{1}, Z, \mathbf{X})$, and $W^{\perp_{z\mathbf{x}}}$ for $W - \mathrm{Pj}(W|\mathbf{1}, Z, \mathbf{X})$. Then $\mathrm{Pj}(W^{\perp_{\mathbf{x}}}|Z^{\perp_{\mathbf{x}}}) = B^* Z^{\perp_{\mathbf{x}}}$, and $\mathrm{Pj}(Y^{\perp_{z\mathbf{x}}}|W^{\perp_{z\mathbf{x}}}) = \hat{\delta} W^{\perp_{z\mathbf{x}}}$. These formulas imply $B^* = \sigma_{wz\cdot\mathbf{x}}/\sigma^2{}_{z\cdot\mathbf{x}}$ and $\hat{\delta} = \sigma_{yw\cdot z\mathbf{x}}/\sigma^2{}_{w\cdot z\mathbf{x}} = \rho_{yw\cdot z\mathbf{x}}\sigma_{y\cdot z\mathbf{x}}/\sigma_{w\cdot z\mathbf{x}}$, so that $b - \beta = B^*\hat{\delta}$ can be written as the product of $\sigma_{y\cdot z\mathbf{x}}/\sigma_{z\cdot\mathbf{x}}$, $\sigma_{wz\cdot\mathbf{x}}/(\sigma_{z\cdot\mathbf{x}}\sigma_{w\cdot z\mathbf{x}})$ and $\rho_{yw\cdot z\mathbf{x}}$. Introducing mutually cancelling factors of $(df)^{\pm 1/2}$ to the first and second of these and applying (4.7) turns this into the product of $\mathrm{SE}(b)$, $(df)^{1/2}\sigma_{wz\cdot\mathbf{x}}/(\sigma_{z\cdot\mathbf{x}}\sigma_{w\cdot z\mathbf{x}})$ and $\rho_{yw\cdot z\mathbf{x}}$. But t_W is just the ratio of $\sigma_{wz\cdot\mathbf{x}}/\sigma^2{}_{w\cdot\mathbf{x}}$ to $\sigma_{z\cdot w\mathbf{x}}/[(df)^{1/2}\sigma_{w\cdot\mathbf{x}}]$, which simplifies to the second of these terms. The result follows. □

Proof of Proposition IV.3. The proof of Proposition IV.3 can be divided into two cases depending on inequality comparing $\rho^2_{y\cdot w|z\mathbf{x}}$ and $g(t, q, c)$ as defined in the proposition.

If $\rho^2_{y\cdot w|z\mathbf{x}} > g(t_w, q, c)$:

For a fixed q and t_w, by Proposition IV.1 and Proposition IV.2, the desired confidence interval $\beta \pm q\mathrm{SE}(\beta)$ can be represented as follows:

$$\beta - q\mathrm{SE}(\beta) = \hat{b} + l(\arcsin\rho_{y\cdot w|z\mathbf{x}})$$

$$\beta + q\mathrm{SE}(\beta) = \hat{b} + u(\arcsin\rho_{y\cdot w|z\mathbf{x}})$$

where $\text{arcsin}\rho_{y\cdot w|z\mathbf{x}} \in (-\pi/2, \pi/2)$ and

$$l(\theta) := a\text{sin}\theta - d\text{cos}\theta \tag{4.9}$$

$$u(\theta) := a\text{sin}\theta + d\text{cos}\theta, \tag{4.10}$$

$a := -t_W\text{SE}(b)$, and $d := \sqrt{(t_W^2 + df)/(df-1)\text{SE}(b)}$. By calculus, $u(\cdot)$ can be seen to have its maximum over $(-\pi/2, \pi/2)$ at arctan $a/d \in (-\pi/2, \pi/2)$. By algebra and trigonometric identities, that maximum is $\sqrt{a^2+d^2}$. Similarly, over $(-\pi/2, \pi/2)$, $l(\cdot)$ takes its minimum value of $-\sqrt{a^2+d^2}$ at arctan $-a/d$. The part of Proposition IV.3 for $\rho^2_{y\cdot w|z\mathbf{x}} > g(t_w, q, c)$ follows.

If $\rho^2_{y\cdot w|z\mathbf{x}} \leq g(t_w, q, c)$:

Note that as θ varies in $(-\pi/2, \pi/2)$, the function $u(\theta)$ (4.10) takes its maximum at arctana/d, or arctan$\frac{-t_W}{\sqrt{(t_W^2+df)/(df-1)}}$ and $l(\theta)$ takes its minimum value at arctan$\frac{t_W}{\sqrt{(t_W^2+df)/(df-1)}}$.

For some ρ such that $0 \leq \rho \leq \text{sinarctan}\frac{|t_W|}{\sqrt{(t_W^2+df)/(df-1)}}$. Let $\rho_{y\cdot w|z\mathbf{x}} \in [-\rho, \rho]$, then

$$\arg\max_{\rho_{y\cdot w|z\mathbf{x}}} u(\text{arcsin}\rho_{y\cdot w|z\mathbf{x}}) = \begin{cases} \rho, & t_W < 0 \\ -\rho, & t_W > 0 \end{cases} \tag{4.11}$$

$$\arg\min_{\rho_{y\cdot w|z\mathbf{x}}} l(\text{arcsin}\rho_{y\cdot w|z\mathbf{x}}) = \begin{cases} -\rho, & t_W < 0 \\ \rho, & t_W > 0 \end{cases} \tag{4.12}$$

To verify (4.11), note $u(\text{arcsin}(\cdot))$ takes only one maximum as its argument ranges over $[-1, 1]$ at

$$\sin\ \arctan \frac{|t_W|}{\sqrt{(t_W^2 + df)/(df - 1)}} \geq \rho$$

Thus, $u(\arcsin \rho_{y\cdot w|z\mathbf{x}})$ must take its maximum as $\rho_{y\cdot w|z\mathbf{x}}$ ranges over $[-\rho, \rho]$ at one of the endpoints of that interval. (4.12) is established by similar reasoning. □

4.5 Extensions

In Sections 4.4 and 4.3, our method of sensitivity analysis is presented for a single omitted variable with the only confounder controls made by including such confounders as predictors in the linear model. While this method is applicable to many modeling scenarios, it has limitations; extensions presented in this section allow our sensitivity analysis method to be adapted to a broader range of contexts. The method can be used in the case of multiple variable omissions (or similarly, when a single omitted confounder is a factor variable with more than two levels). In addition, confounder controls can be made not only by including more predictors in the model; our method can be used when an analyst incorporates propensity score matched sets to control for background information.

4.5.1 Several variables omitted at once

In the event that W is a factor variable with more than two levels or a collection of omitted variables (any case in which $rank(W) > 1$ in the design matrix), our method applies with some adjustments. The manner for describing the relationship of W with the outcome variable, $\rho_{y\cdot w|z\mathbf{x}}$, can be conceptualized and benchmarked as before. Rather than describing treatment-confoundedness with t_W, we use F_W instead, where F_W is the F-statistic from an ANOVA model comparison of models with and without W. In the case of a univariate W, $F_W = t_W^2$, so the propositions presented in this section are generalizations of earlier formulas. A correction for degrees of freedom must be made to earlier propositions. For W such that $rank(W) > 1$, define t^2_W to be $[k * df/(df + 1 + k)]F_W$. Proposition IV.1 can be generalized to put a bound on the bias when W is not univariate.

Corollary IV.4. *Suppose* $R^2_{y\cdot z\mathbf{x}} < 1$*,* ${t^2}_w$ *is finite, and* $\mathrm{rank}(W) = k > 1$*. Then*

$$\begin{aligned} (b-\beta)^2 &\leq \hat{V}(b)\frac{k(df)}{df+1-k}F_w\rho^2_{y\cdot w|z\mathbf{x}}; \text{ or equivalently} \\ |b-\beta| &\leq \mathrm{SE}(b)t_W|\rho_{y\cdot w|z\mathbf{x}}|. \end{aligned}$$

Proof of Corollary IV.4. Without loss of generality, W is uncorrelated with Z and X: if not, replacing W with $W - \mathrm{Pj}(W|X,Z)$ leaves Z-coefficients and their standard errors unchanged. Define $\tilde{W} = \mathrm{Pj}(Y^{\perp \mathbf{x},z}|W)$, where $Y^{\perp \mathbf{x},z} = Y - \mathrm{Pj}(Y|X,Z)$. Again without loss of generality, $W = (\tilde{W}, W_2, \ldots W_k)$, where $\tilde{W} \perp (W_2, \ldots W_k)$. Writing

$$\mathrm{Pj}(Y|Z,\mathbf{X},W) =: \alpha + \beta Z + \gamma\mathbf{X}^T + \delta_1\tilde{W} + \delta_2 W_2 + \cdots + \hat{\delta}_k W_k\,, \tag{4.13}$$

it is immediate that $\delta_2, \ldots, \delta_k = 0$, since $W_2, \ldots, W_k$ are orthogonal to $\mathrm{Pj}(Y^{\perp \mathbf{x},z}|W)$, and hence orthogonal to $Y^{\perp \mathbf{x},z}$. Projecting (4.13) onto the span of $Z, \mathbf{X}$, and then equating the Z-coefficient in what results with the Z-coefficient in (4.2) yields

$$\beta + \delta_1 B_1^* = b, \tag{4.14}$$

where B_1^* is defined by $\mathrm{Pj}(\tilde{W}|Z,X) = B_1^* Z + C^* X$. In other words, b and β are related just as they would have been had W been of rank 1, rather than k, consisting only of $\tilde{W}$.

Corollary IV.4 requires that relationships between quantities for W and $\tilde{W}$ be established. Lemma IV.5 establishes these relationships and is proved in the appendix.

Lemma IV.5. *Suppose* $R^2_{y\cdot z\mathbf{x}} < 1$, ${t^2}_w$ *is finite, and* $\operatorname{rank}(W) = k$. *Then*

1. $\rho^2_{y\cdot w|z\mathbf{x}} = \rho^2_{y\cdot \tilde{w}|z\mathbf{x}}$
2. $t^2_{\tilde{W}} \leq k\frac{df}{df+1-k}F_W$

The desired result now follows from (4.14), Proposition IV.1, and Lemma IV.5. □

When $rank(W) > 1$, Proposition IV.2 can be generalized to Proposition IV.6, which is proved in the appendix.

Proposition IV.6. *Suppose* $R^2_{y\cdot z\mathbf{x}} < 1$, t^2_W *is finite, and* $\operatorname{rank}(W) = k$, $k > 1$. *Then*

$$\hat{V}(\beta) = \hat{V}(b)\left[\frac{df + t^2_W}{df - k}\right](1 - \rho^2_{y\cdot w|z\mathbf{x}}). \tag{4.15}$$

The sensitivity intervals in Proposition IV.3 follow algebraically from the bias and standard error representations (4.3) and (4.4), they work for W of arbitrary rank. The proofs of Proposition IV.3 and the following are essentially the same.

Proposition IV.7. *In the setting of Proposition IV.3 except with* $\operatorname{rank}(W) > 1$, *(4.6) holds with* $df_w = df - \operatorname{rank}(W)$ *and* $t_W = \{[k(df)/(df + 1 - k)]F_W\}^{1/2}$.

The ideas of Proposition IV.7 have already been applied earlier in Section 4.4 in the case of the variable "Insurance class", which is a factor variable with 6 levels. The value of t_W presented in Table 4.4 is an adjusted version of $\sqrt{F_W}$.

4.5.2 Propensity-adjusted estimates of treatment effects

Observational studies, in the absence of random assignment to treatment groups, are subject to the criticism that covariate distributions differ across treatment groups. Stratification on the propensity score is a method of covariate adjustment commonly used with observational studies in an attempt to balance covariate distributions across treatment groups. The propensity score is an individual's probability of receiving treatment, in this case RHC, conditioned on the observed covariates (Rosenbaum and Rubin, 1983). The standard recommendation dictates that stratification into five subclasses on the basis of the propensity score can remove 90% of the bias due to imbalances in observed covariates across treatment groups (Rosenbaum and Rubin, 1984). After creating propensity subclasses, a straightforward way of obtaining treatment effect estimates is to use the coefficient of the treatment variable in a regression of the outcome on an indicator of propensity subclass and the treatment variable. In this way, interval estimates of the treatment effect should be comparable to those seen in Table 4.4, but with the additional adjustment to balance observed covariates across treatment groups in acknowledgement of the lack of randomization.

The propensity matched set regression models could also be viewed as an analogue to "controlling for" variables by including them in an OLS model. Typically, when OLS regression models are used, variable selection, as was performed in Section 2, is advisable; in the case of computing propensity scores, the established advice is to put all variables in the propensity score without variable selection, regardless of the additional impact on the propensity score (Rubin and Thomas, 1996). The methods used in both Section 4.1 and 4.2 could be viewed as different ways of adjusting for the approximately fifty variables for which one could account with this data. Of course, propensity matched set adjusted regressions are typically more advisable in the analysis of observational data.

To illustrate the sensitivity analysis results with various propensity score adjusted regressions, propensity subclasses must be created. The model used to estimate the propensity score is a logistic regression model in which the predictors are all variables that could be reasonably assumed to be measured prior to a patient's treatment with

RHC and the response is the indicator of receiving RHC. A randomization-based test is used to test for covariate balance across the treatment groups: this test aims to see if the covariate distributions observed across treatment groups could have reasonably resulted from a randomized experiment (for motivation and specifics of this test, refer to Hansen and Bowers (2008)). The test indicates that with the standard five subclasses it cannot be reasonably concluded that the covariate distributions across treatment groups could have resulted from a randomized experiment. With six subclasses, however, there is not a significant imbalance in the observed covariates at a level of $\alpha = 0.10$

In the case of propensity-adjusted regressions, benchmarking follows a similar process with adjustments to account for the propensity score strata. To compute values of t_W, one at at time, variables are withheld from the propensity score, the subclassification into sextiles is made, and the variable of interest is then added into the regression model with an indicator for propensity matched set and the treatment assignment variable. The t statistic corresponding to the variable under study is the taken to be the t_W value for that variable. These values are presented for several covariates in the right side of Table 4.5. Values of $\rho^2_{y\cdot w|z\mathbf{x}}$ can be determined as before, and in this section, we consider the same bounds considered in Table 4.4.

Table 4.5 presents the results of the sensitivity analysis methods in the setting of propensity adjusted regressions for a subset variables (the entire table is presented in the Appendix in Table 4.7). The columns on the left side of Table 4.5 are columns of Table 4.4 presented here for comparison, and the right side of Table 4.5 displays similar results for propensity score matched set adjusted regressions. From Table 4.5, one can observe that the more associated an omitted variable is with the treatment variable, the propensity matched set adjusted regressions are more stable. For the variables that have high associations with treatment, such as respiratory evaluation and mean blood pressure, the confidence intervals obtained from the propensity score matched set regressions fall well within those obtained by applying the sensitivity analysis methods to the OLS regressions. For variables with less association with treatment, such as immunosuppression and weight, the intervals obtained from sensitivity analyses on the OLS regressions fall slightly (though not substantially) within those obtained for the propensity matched set regressions. As has been seen in earlier results of Table 4.4, the association of an omitted variable with the treatment variable largely controls the outer limits of the confidence intervals, so it logically follows that for variables with moderate or large associations with treatment, the propensity score, which is formulated based on relationships to the treatment variable, should better control these outer limits.

Table 4.5: Sensitivity intervals for the treatment effect after ordinary covariance and propensity-score adjustment, illustrating that propensity adjustment better limits sensitivity to the omission of adjustment variables. For covariance adjustment, t_W is limited by the confoundedness with the treatment of 6 variables that had been eliminated by a preliminary variable-selection procedure, as in Table 4.4; for propensity adjustment, limits on treatment confounding are set by separately removing each of these and calculating their t_W's after propensity adjustment for remaining variables.

	OLS regression			Propensity adjusted regression		
	$\lvert t_w \rvert$	$\rho^2_{y\cdot w\lvert z\mathbf{x}} \leq .01$	$\rho^2_{y\cdot w\lvert z\mathbf{x}} \leq .1$	$\lvert t_w \rvert$	$\rho^2_{y\cdot w\lvert z\mathbf{x}} \leq .01$	$\rho^2_{y\cdot w\lvert z\mathbf{x}} \leq .1$
Insurance class	12.2	(0.03, 0.20)	(-0.04 , 0.26)	8.6	(0.02, 0.18)	(-0.03, 0.23)
Respiratory eval	8.9	(0.04 , 0.19)	(-0.01 , 0.23)	3.1	(0.04, 0.17)	(0.02, 0.19)
Mean blood press.	8.6	(0.04 , 0.19)	(-0.01 , 0.23)	6.8	(0.03, 0.18)	(-0.01, 0.22)
Cardiovascular eval	8.5	(0.04 , 0.19)	(-0.01 , 0.23)	5.4	(0.03, 0.17)	(0, 0.20)
Weight (kg)	6.1	(0.04, 0.18)	(0.01 , 0.21)	5.1	(0.03, 0.18)	(0, 0.21)
Immunosuppression	0.4	(0.06 , 0.16)	(0.06 , 0.16)	0.5	(0.04, 0.16)	(0.04, 0.16)

4.6 Summary

With a multiple regression analysis in which the estimated coefficient of the treatment variable is taken as an estimate of treatment effect, there is potential for an unobserved variable to impact the the resulting effect estimate. Methods of sensitivity analysis presented here can be used to ascertain the extent to which an unmeasured confounder could alter an interval estimate of treatment effect rather than only assessing its impact on the point estimate or p-value. Our method makes use of two key quantities: the association of an omitted variable with the response variable as measured by $\rho_{y\cdot w|z\mathbf{x}}$ and its association with the treatment variable, or treatment-confounding, represented as t_W. A sensitivity analysis in which a reasonable bound for $\rho_{y\cdot w|z\mathbf{x}}$ is used yields narrower confidence intervals than a sensitivity analysis without such bounds, and the present analysis illustrates how such a bound could be chosen from examination of the data. The value of t_W may need to be more carefully chosen by an analyst and tends to have a greater impact on the resulting interval estimates of treatment effect, and estimating its value could make use of external data sources, if available. With adjusted interval estimates, an analyst can present a more realistic picture of the effect of some treatment, acknowledging that some useful information may not have been measured. While some may skeptically toss aside conclusions reached from a study in which an omitted variable could affect the results, with careful consideration of the potential omitted variables, an analyst can bolster resulting inferences by determining the extent to which such a variable could alter the treatment effect estimate.

4.7 Appendix

The appendix contains two pieces: proofs of theoretical results and tables of sensitivity intervals analogous to earlier tables.

4.7.1 Proofs of theoretical results

Proof of Lemma IV.5. Under the conditions of the lemma, (1) and (2) can be established:

1. In a regression of Z and $\mathbf{X}$ on Y, adding $\tilde{W}$ has the same effect on the Z-coefficient and model R^2 as adding W. Thus, (1) holds.

2. Furthermore, $\tilde{W} \in \text{span}(W)$, so $\tilde{W}$ explains no more variation in Z than does W. Thus, $R^2_{z\cdot w|x} \geq R^2_{z\cdot \tilde{w}|x}$, which implies $\frac{\rho^2{}_{z\cdot w|x}}{1-\rho^2{}_{z\cdot w|x}} \geq \frac{\rho^2{}_{z\cdot \tilde{w}|x}}{1-\rho^2{}_{z\cdot \tilde{w}|x}}$. As the ANOVA F-statistic is defined as

$$F_W = \frac{(\sigma^2_{z\cdot\mathbf{x}} - \sigma^2_{z\cdot\mathbf{xw}})/(k)}{\sigma^2_{z\cdot\mathbf{xw}}/(df+1-k)} = \frac{df+1-k}{k}\frac{\rho^2_{z\cdot w|\mathbf{x}}}{1-\rho^2_{z\cdot w|\mathbf{x}}}$$

and $\text{rank}(\tilde{W}) = 1$, so $t^2_{\tilde{W}} = df \frac{\rho^2{}_{z\cdot\tilde{w}|x}}{1-\rho^2{}_{z\cdot\tilde{w}|x}}$. The result follows.

□

Proof of Proposition IV.6. To relate SE(b) and SE(β), begin by rewriting some of the variance terms:

$$\begin{aligned} \sigma^2{}_{y\cdot z\mathbf{x}w} &= \text{Var}(Y^{\perp z\mathbf{x}w}) \\ &= \text{Var}(Y^{\perp z\mathbf{x}} - \text{Pj}(Y^{\perp z\mathbf{x}}|W^{\perp z\mathbf{x}})) \\ &= \text{Var}(Y^{\perp z\mathbf{x}} - \tfrac{\sigma_{y\cdot z\mathbf{x}}}{\sigma_{w\cdot z\mathbf{x}}}W^{\perp z\mathbf{x}})) \\ &= \sigma^2{}_{y\cdot z\mathbf{x}} - \sigma^2{}_{y\cdot z\mathbf{x}}\rho^2{}_{yw\cdot z\mathbf{x}}, \end{aligned} \tag{4.16}$$

so that

$$\sigma_{y\cdot z\mathbf{x}w} = \sigma_{y\cdot z\mathbf{x}}\sqrt{1-\rho^2{}_{yw\cdot z\mathbf{x}}}.$$

Since $\rho^2_{z\cdot w|\mathbf{x}}$ is the proportionate reduction in residual variance when W is added to the regression of Z on X,

$$\sigma^2_{z\cdot\mathbf{x}w} = \sigma^2_{z\cdot\mathbf{x}}(1-\rho^2_{z\cdot w|\mathbf{x}}). \tag{4.17}$$

Thus

$$\begin{aligned} \mathrm{SE}(\beta) &= \tfrac{(df-k)^{-1/2}\sigma_{y\cdot z\mathbf{x}w}}{\sigma_{z\cdot\mathbf{x}w}}\sqrt{\tfrac{1-{\rho^2}_{yw\cdot z\mathbf{x}}}{1-{\rho^2}_{zw\cdot\mathbf{x}}}} \\ &= \mathrm{SE}(b)\sqrt{\tfrac{df}{df-k}}\sqrt{\tfrac{1-{\rho^2}_{yw\cdot z\mathbf{x}}}{1-{\rho^2}_{zw\cdot\mathbf{x}}}}. \end{aligned} \tag{4.18}$$

Now from the definition of the ANOVA F-statistic and (4.17),

$$F_W = \frac{(\sigma^2_{z\cdot\mathbf{x}} - \sigma^2_{z\cdot\mathbf{xw}})/(k)}{\sigma^2_{z\cdot\mathbf{xw}}/(df+1-k)} = \frac{df+1-k}{k}\frac{\rho^2_{z\cdot w|\mathbf{x}}}{1-\rho^2_{z\cdot w|\mathbf{x}}},$$

whence

$$\frac{1}{1-\rho^2_{z\cdot w|\mathbf{x}}} = 1 + \frac{k}{df+1-k}F_W.$$

In Section 4.5.1, t_W is defined for multivariate W in such a way that $t_W^2 = [k(df)/(df+1-k)]F_W$. Since $k = \mathrm{rank}(W)$, in case of univariate W to say that $t_W^2 = [k(df)/(df+1-k)]F_W$ is the same as to assert $t_W^2 = F_W$ — a well-known relationship between regression coefficients' t-statistics and ANOVA F statistics. It follows that

$$\frac{1}{1-\rho^2_{z\cdot w|\mathbf{x}}} = \frac{df+t_W^2}{df},$$

which combines with (4.18) to give (4.15). □

4.7.2 Result tables for largest models

Table 5 and Table 6 present sensitivity results — analogues to those shown in Table 4.5— for all variables in the OLS and propensity score models, respectively.

Table 4.6: Sensitivity adjustments to 95% t-intervals in a OLS regression. The first column represents the confidence interval that results if the row-variable, V, is added to the regression with the 18 covariates in Table 4.2. (For comparison, the baseline interval adjusting only for these 18 covariates is $(0.06, 0.16)$. Columns 2–4 give the union of intervals resulting from adding covariates W that are no more confounded with the treatment than V, $t^2_W \leq t^2_V$, but which better predict the response (decreasing unexplained variation by 1%, 10%, or by any amount).

		Interval containing $\hat{\beta} \pm q\mathrm{SE}(\hat{\beta})$, if $t^2_W \leq t^2_V$ and		
V	t^2_W	$\rho^2_{y \cdot w \vert z\mathbf{x}} \leq .01$	$\rho^2_{y \cdot w \vert z\mathbf{x}} \leq .1$	$\rho^2_{y \cdot w \vert z\mathbf{x}} \in [0, 1]$
Background characteristics:				
initial disease category 1	32.4	(−0.02,0.27)	(−0.21,0.46)	(−0.76, 1.01)
initial disease category 2	17.6	(0.01 ,0.22)	(−0.09,0.32)	(−0.36, 0.60)
insurance class	9.4	(0.02 ,0.18)	(−0.03,0.23)	(−0.16, 0.36)
education	2.3	(0.04 ,0.16)	(0.03 ,0.17)	(0.02 , 0.18)
income	1.4	(0.04 ,0.16)	(0.04 ,0.16)	(0.03 , 0.17)
race	1.2	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
sex	0.7	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
age	0.5	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
Comorbidities illness:				
transfer from another hosp.	4.4	(0.04 ,0.17)	(0.01 ,0.19)	(−0.03, 0.23)
dementia	3.5	(0.04 ,0.16)	(0.02 ,0.18)	(−0.01, 0.21)
psychological disorder	3.3	(0.04 ,0.16)	(0.02 ,0.18)	(0.00 , 0.21)
renal disease	2.1	(0.04 ,0.16)	(0.03 ,0.17)	(0.02 , 0.18)
existing cancer	1.1	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
GI bleed	0.8	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
myocardial infarction	0.6	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
cardiovascular disease	0.5	(0.05 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
congestive heart fail.	0.5	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
immunosuppression	0.5	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
liver disease	0.4	(0.05 ,0.16)	(0.05 ,0.16)	(0.05 , 0.16)
malignant tumor	0.2	(0.05 ,0.15)	(0.05 ,0.15)	(0.05 , 0.15)
pulmonary disease	0.1	(0.05 ,0.15)	(0.05 ,0.15)	(0.05 , 0.15)
Day 1 Measurements:				
PaO2/FIO2	15.1	(0.02 ,0.21)	(−0.07,0.29)	(−0.29, 0.52)
respiration rate	9.0	(0.03 ,0.18)	(−0.02,0.23)	(−0.14, 0.35)
mean blood press.	6.8	(0.03 ,0.17)	(−0.01,0.21)	(−0.09, 0.29)
PaCo2	6.7	(0.02 ,0.16)	(−0.02,0.20)	(−0.10, 0.28)
heart rate	5.6	(0.03 ,0.17)	(0.00 ,0.20)	(−0.06, 0.26)
weight (kg)	4.8	(0.03 ,0.17)	(0.01 ,0.19)	(−0.04, 0.24)
DNR order	4.7	(0.05 ,0.19)	(0.03 ,0.22)	(−0.02, 0.26)
potassium	4.7	(0.03 ,0.17)	(0.01 ,0.19)	(−0.04, 0.24)
PH	4.0	(0.03 ,0.16)	(0.01 ,0.18)	(−0.03, 0.22)
APACHE score	3.9	(0.03 ,0.16)	(0.01 ,0.18)	(−0.02, 0.22)
sodium	3.3	(0.03 ,0.16)	(0.02 ,0.18)	(−0.01, 0.20)
hematocrit	2.3	(0.04 ,0.17)	(0.03 ,0.18)	(0.02 , 0.19)
creatinine	2.0	(0.04 ,0.16)	(0.03 ,0.17)	(0.02 , 0.18)
temperature	1.8	(0.04 ,0.16)	(0.03 ,0.17)	(0.03 , 0.17)
albumin	1.8	(0.04 ,0.16)	(0.04 ,0.17)	(0.03 , 0.18)
bilirubin	1.4	(0.04 ,0.16)	(0.04 ,0.16)	(0.03 , 0.16)
urine output	1.2	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
white blood cell ct	0.3	(0.05 ,0.15)	(0.05 ,0.15)	(0.05 , 0.15)
coma score	0.2	(0.05 ,0.15)	(0.05 ,0.15)	(0.05 , 0.15)
Admit Diagnosis Categories:				
cardiovascular	6.4	(0.03 ,0.17)	(−0.01,0.20)	(−0.08, 0.28)
trauma	3.8	(0.04 ,0.17)	(0.02 ,0.19)	(−0.02, 0.22)
respiratory	3.8	(0.04 ,0.16)	(0.02 ,0.18)	(−0.02, 0.22)
neurology	3.5	(0.03 ,0.16)	(0.01 ,0.18)	(−0.01, 0.21)
hematologic	2.9	(0.04 ,0.16)	(0.03 ,0.18)	(0.01 , 0.20)
gastrology	2.9	(0.04 ,0.16)	(0.02 ,0.18)	(0.01 , 0.20)
renal	1.8	(0.04 ,0.16)	(0.03 ,0.17)	(0.03 , 0.17)
orthopedic	1.3	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
metabolic	1.3	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.17)
sepsis	1.0	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)

Table 4.7: Sensitivity adjustments in a regression incorporating propensity subclasses. The first column represents the confidence interval that results if the row-variable, V, is *excluded* from the propensity adjustment but then *added* to the outcome model as a regressor, alongside propensity subclasses. (For comparison, the baseline interval adjusting for all 50 variables by propensity subclassification is $(0.04, 0.16)$.) Columns 2–4 give the union of intervals resulting from adding regressors W to the outcome model which (given the 50-variable propensity adjustment) are confounded with the treatment variable no more than V was (given the 49-variable propensity adjustment that excluded V itself), but which better predict the response (decreasing unexplained variation by 1%, 10%, or by any amount).

		Interval containing $\hat{\beta} \pm q\mathrm{SE}(\hat{\beta})$, if $t_W^2 \leq t_V^2$ and		
V	$\lvert t_W \rvert$	$\rho^2_{y\cdot w\lvert z\mathbf{x}} \leq .01$	$\rho^2_{y\cdot w\lvert z\mathbf{x}} \leq .1$	$\rho^2_{y\cdot w\lvert z\mathbf{x}} \in [0,1]$
Background characteristics:				
initial disease category 1	27.6	(−0.01,0.26)	(−0.18,0.43)	(−0.66, 0.91)
initial disease category 2	14.9	(0.02 ,0.22)	(−0.07,0.31)	(−0.31, 0.55)
insurance class	8.6	(0.02 ,0.18)	(−0.03,0.23)	(−0.15, 0.36)
education	1.8	(0.04 ,0.17)	(0.03 ,0.18)	(0.03 , 0.18)
income	1.4	(0.04 ,0.16)	(0.03 ,0.17)	(0.03 , 0.17)
race	1.0	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.17)
sex	0.7	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
age	0.3	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
Comorbidities illness:				
transfer from another hosp.	4.3	(0.03 ,0.17)	(0.01 ,0.20)	(−0.03, 0.24)
dementia	3.4	(0.04 ,0.17)	(0.02 ,0.19)	(−0.01, 0.22)
psychological disorder	3.0	(0.03 ,0.17)	(0.02 ,0.18)	(0.00 , 0.20)
renal disease	1.7	(0.04 ,0.16)	(0.03 ,0.17)	(0.03 , 0.18)
GI bleed	0.9	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.17)
existing cancer	0.8	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
congestive heart fail.	0.7	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
cardiovascular disease	0.6	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
myocardial infarction	0.6	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
immunosuppression	0.5	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
malignant tumor	0.4	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
pulmonary disease	0.1	(0.04 ,0.16)	(0.05 ,0.16)	(0.04 , 0.16)
liver disease	0.1	(0.04 ,0.16)	(0.05 ,0.16)	(0.04 , 0.16)
Day 1 Measurements:				
PaO2/FIO2	13.3	(0.02 ,0.21)	(−0.06,0.28)	(−0.27, 0.49)
respiration rate	7.4	(0.03 ,0.18)	(−0.02,0.23)	(−0.12, 0.32)
mean blood press.	6.9	(0.03 ,0.18)	(−0.01,0.22)	(−0.10, 0.31)
weight (kg)	5.1	(0.03 ,0.18)	(0.01 ,0.21)	(−0.05, 0.26)
potassium	4.2	(0.03 ,0.17)	(0.01 ,0.19)	(−0.03, 0.24)
DNR order	4.9	(0.06 ,0.20)	(0.03 ,0.23)	(−0.02, 0.28)
heart rate	4.7	(0.03 ,0.18)	(0.01 ,0.20)	(−0.04, 0.25)
PaCo2	4.6	(0.03 ,0.17)	(0.00 ,0.20)	(−0.04, 0.24)
PH	3.4	(0.03 ,0.16)	(0.01 ,0.18)	(−0.02, 0.21)
sodium	3.2	(0.04 ,0.17)	(0.02 ,0.19)	(−0.00, 0.21)
APACHE score	3.2	(0.03 ,0.17)	(0.02 ,0.18)	(−0.01, 0.21)
hematocrit	2.6	(0.04 ,0.17)	(0.03 ,0.19)	(0.01 , 0.20)
albumin	1.9	(0.04 ,0.17)	(0.03 ,0.18)	(0.03 , 0.18)
creatinine	1.9	(0.04 ,0.16)	(0.03 ,0.17)	(0.02 , 0.18)
temperature	1.8	(0.04 ,0.16)	(0.03 ,0.17)	(0.02 , 0.18)
bilirubin	1.5	(0.04 ,0.16)	(0.03 ,0.17)	(0.03 , 0.17)
urine output	1.1	(0.04 ,0.16)	(0.03 ,0.16)	(0.03 , 0.16)
coma score	0.5	(0.04 ,0.16)	(0.04 ,0.16)	(0.04 , 0.16)
white blood cell ct.	0.0	(0.04 ,0.16)	(0.05 ,0.16)	(0.04 , 0.16)
Admit Diagnosis Categories:				
cardiovascular	5.4	(0.03 ,0.17)	(0.00 ,0.20)	(−0.07, 0.26)
trauma	3.8	(0.03 ,0.17)	(0.01 ,0.19)	(−0.02, 0.22)
respiratory	3.1	(0.04 ,0.17)	(0.02 ,0.19)	(0.00 , 0.21)
neurology	2.9	(0.03 ,0.17)	(0.02 ,0.18)	(0.00 , 0.20)
hematologic	2.4	(0.04 ,0.17)	(0.03 ,0.18)	(0.02 , 0.20)
gastrology	2.3	(0.04 ,0.17)	(0.03 ,0.18)	(0.02 , 0.19)
renal	1.8	(0.04 ,0.16)	(0.03 ,0.17)	(0.02 , 0.18)
sepsis	1.2	(0.04 ,0.16)	(0.04 ,0.17)	(0.03 , 0.17)
orthopedic	1.2	(0.04 ,0.16)	(0.03 ,0.17)	(0.03 , 0.17)
metabolic	1.1	(0.04 ,0.16)	(0.04 ,0.17)	(0.04 , 0.17)

BIBLIOGRAPHY

Bibliography

Angrist, J. and Imbens, G. (2002). Comment. *Statist. Sci.*, **17** 304–07.

Austin, P. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, **28** 3083–3107.

Belson, W. (1956). A technique for studying the effects of a television broadcast. *Applied Statistics* 195–202.

Brookhart, M., Schneeweiss, S., Rothman, K., Glynn, R., Avorn, J. and Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, **163** 1149.

Christensen, R. (1996). *Plane Answers to Complex Questions: The Theory of Linear Models.* Springer.

Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *Supplement to the Journal of the Royal Statistical Society*, **5** 171–176.

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, **128** 234–266.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, **24** 295–313.

Connors, A. J., Speroff, T., Dawson, N., Thomas, C., Harrell, F. E. J., Wagner, D., Desbiens, N., Goldman, L., Wu, A., Califf, R., Fulkerson, W. J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J. and Knaus, W. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. support investigators. *J. Amer. Med. Assoc.*, **276** 889–97.

Copas, J. B. and Li, H. G. (1997). Inference for non-random samples. *J. Roy. Statist. Soc. Ser. B*, **59** 55–95. With discussion and a reply by the authors.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. and Wydner, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, **22** 173–203.

Derigs, U. (1988). Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research*, **13** 225–261.

DiNardo, J. (2002). Propensity score reweighting and changes in wage distributions. *University of Michigan, mimeograph.*

Faraway, J. J. (1992). On the cost of data analysis. *Journal of Computational and Graphical Statistics*, **1** 213–229.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, **98** 39–82.

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological methods and research*, **29** 147–194.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, **86** 77–90.

Gastwirth, J. and Greenhouse, S. (1995). Biostatistical concepts and methods in the legal setting. *Statistics in Medicine*, **14** 1641–1653.

Gelman, A. (2004). *Bayesian data analysis.* CRC press.

Gore, J., Goldberg, R., Spodick, D. and et al (1987). A community-wide assessment of the use of pulmonary artery catheters in patients with acute myocardial infarction. *Chest*, **92** 721–727.

Hansen, B. (2008). The prognostic analogue of the propensity score. *Biometrika*, **95** 481.

Hansen, B. and Bowers, J. (2009). Attributing Effects to a Cluster-Randomized Get-Out-the-Vote Campaign. *Journal of the American Statistical Association*, **104** 873–885.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Am. Statist. Assoc.*, **99** 609–618.

Hansen, B. B. and Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, **23** 219–236.

Harvey, S., Harrison, D. and et al (2005). Assessment of the clinical effectiveness of pulmonary artery catheters in management of patients in intensive care (pac-man): a randomised controlled trial. *Lancet*, **366** 472–7.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources*, **32** 441–462.

Ho, D., Imai, K., King, G. and Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*.

Holland, P. W. (1986). Statistics and causal inference. *J. Am. Statist. Assoc.*, **81** 945–960.

Hong, G. and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: a case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.*, **101** 901–910.

Imai, K., King, G. and Stuart, E. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, **171** 481–502.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, **86** 4–29.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* 126–132.

Lee, W.-S. (2008). Propensity score matching and variations on the balancing test. *Manuscript.*

Lin, D., Psaty, B. and Kronmal, R. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, **54** 948–963.

Lu, B., Zanutto, E., Hornik, R. and Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *J. Amer. Statist. Assoc.*, **96** 1245–1253.

Marcus, S. M. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *J. Educ. and Behav. Stat.*, **22** 193–201.

Ñopo, H. (2008). Matching as a tool to decompose wage gaps. *The Review of Economics and Statistics*, **90** 290–299.

Peters, C. (1941). A method of matching groups for experiment with no loss of population. *The Journal of Educational Research* 606–612.

Powers, D. and Rock, D. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement*, **36** 93–118.

Rhodes, A., Cusack, R., Newman, M. and et al (2002). A randomized, controlled trial of the pulmonary artery catheter in critically ill patients. *Intensive Care Medicine*, **348** 5–14.

Richard, C. and et al (2003). Early use of the pulmonary artery catheter and outcomes in patients with shock and acute respiratory distress syndrome: a randomized controlled trial. *J. Amer. Med. Assoc.*, **290** 2732–4.

Robins, J. M., Scharfstein, D. and Rotnitzky, A. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment and Clinical Trials* (M. Halloran and D. Berry, eds.). Springer, New York, 1–94.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *J. Am. Statist. Assoc.*, **79** 41–48.

Rosenbaum, P. R. (1988). Sensitivity analysis for matching with multiple controls. *Biometrika*, **75** 577–581.

Rosenbaum, P. R. (1991). Sensitivity analysis for matched case-control studies. *Biometrics*, **47** 87–100.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc. B*, **45** 212–218.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Statist. Assoc.*, **79** 516–524.

Rubin, D. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, **127** 757.

Rubin, D. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in medicine*, **26** 20.

Rubin, D. (2008). For objective causal inference, design trumps analysis. *Annals*, **2** 808–840.

Rubin, D. and Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, **52** 249–264.

Sampson, R. J., Raudenbush, S. W. and Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, **277** 918–924.

Sandhan, J. D., Hull, R. D., Brant, R. F., Knox, L., Pineo, G. F., Doig, C. J., Laporta, D. P., Viner, S., Passerini, L., Devitt, H., Kirby, A., Jacka, M. and the Canadian Critical Care Clinical Trials Group (2003). A Randomized, Controlled Trial of the Use of Pulmonary-Artery Catheters in High-Risk Surgical Patients. *New England Journal of Medicine*, **348** 5–14.

Seber, G. (1977). *Linear Regression Analysis.* John R. Wiley and Sons.

Shah, M. and Stevenson, L. (2004). Evaluation study of congestive heart failure and pulmonary artery catheterization effectiveness: the escape trial. *American Heart Association Scientific Sessions, New Orleans.*

Sherman, L., Schmidt, J., Rogan, D. and Smith, D. (1992). Variable effects of arrest on criminal careers: The Milwaukee domestic violence experiment, the. *J. Crim. L. & Criminology*, **83** 137.

Small, D. S. (2007). Sensitivity Analysis for Instrumental Variables Regression With Overidentifying Restrictions. *J. Am. Statist. Assoc.*, **102** 1049.

Stuart, E. (2010). Matching Methods for Causal Inference: A review. *Statistical Science.*

Stuart, E., Cole, S., Bradshaw, C. and Leaf, P. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174** 369–386.

Zion, M., Balkin, J., Rosenmann, D. and et al. (1990). Use of pulmonary artery catheters in patients with acute myocaridal infarction. *Chest*, **98** 1331–1335.