TESTING FOR COVARIATE BALANCE IN COMPARATIVE STUDIES

by

Yevgeniya N. Kleyman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2009

Doctoral Committee:

Assistant Professor Bendek B. Hansen, Chair
Professor Edward D. Rothman
Professor Jeffrey A. Smith
Professor Yu Xie

To my family

ACKNOWLEDGEMENTS

This dissertation is a reflection of the effort of many people besides myself and I am very grateful for all the support that culminated in this work.

I am heartily thankful to my advisor, Ben Hansen, for his mentorship, encouragement, patience, supervision and support which started before my graduate studies began and continue to this day.

I would also like to express my deep appreciation to Edward Rothman for his profound influence on my interest in Statistics and specifically in causal inference, and on my career path at the University of Michigan. It is a pleasure to thank Jeffrey Smith and Yu Xie for providing their unique and valuable perspectives that helped develop my background in causal inference. I am thankful to all of my committee members for their enthusiasm about my work.

I am thankful to my fellow Statistics students for their help and support. I am also grateful to several faculty members at the Statistics Department whose mentorship over the years has been invaluable.

Finally, I owe a debt of gratitude to my family. My grandparents Alfred and Lyudmila Leybman, my parents, Naum and Yuliya, my brother Ilya, and my husband Lev constitute an incredibly loving, patient and supportive family, without which this surely would not be possible.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

TESTING FOR COVARIATE BALANCE IN COMPARATIVE STUDIES

by

Yevgeniya N. Kleyman

Chair: Bendek B. Hansen

In comparative studies, causal inference necessitates effective adjustments for important covariates. This becomes particularly relevant in observational studies, where covariates are rarely jointly balanced between treatment and control groups, and this lack of covariate balance can generate misleading results. Such adjustments as propensity score matching and stratification are frequently used to align dissimilar groups. The credibility of the analysis may be bolstered by demonstrating that such an adjustment has improved balance on observed covariates. It is not automatic that these measures achieve this objective. To appraise whether they have, practitioners use a variety of techniques, many of them common in hypothesis testing. However, some of these "balance tests" lack formal motivation, may give contradictory messages about balance, and have in some cases been shown to have undesirable statistical properties.

We begin by identifying goals of balance testing in comparative studies and evaluating arguments used in the literature to support and oppose using significance

tests to appraise covariate balance. We survey the literature for existing measures and appraisals of balance, with an interest in their advantages and shortcomings in relation to the goals we identify. We study the performance of some existing ways to assess covariate balance through an examination of contemporary research and identify that a permutation version of the balance-testing procedure originally suggested in Dehejia and Wahba (2002) has been shown to outperform some of the other approaches. We supplement our findings from the existing literature with a thorough simulation study based on real observational data. We use this simulation study to evaluate the impact of using the various balance diagnostics on the validity of statistical inferences about treatment effects.

Our literature review and simulation results suggest an important role for a new formal balance diagnostic. We develop several ideas aimed at this end and test them in various simulation settings that resemble authentic analysis conditions. Using the results of the literature survey and our simulation study, we are able to recommend new and existing techniques for testing covariate balance using randomization-based inference. We also propose dependable combinations of procedures for inference in comparative studies and provide some examples for application of our recommended techniques.

CHAPTER I

Literature Review

1.1 Introduction

There is a rich statistical literature addressing methodology for causal inference in observational studies. Strands of this literature concern adjustments for observed covariates, including methods like matching and stratification which are designed to split control and treated subjects into subclasses based on some measure of comparability. Other parts of the literature have been dedicated to researching the ways to properly adjust for the stratification in estimation of the causal treatment effects. Under randomization, we expect observed and unobserved covariates to be balanced between the treatment and control groups. One goal of matching and stratification in an observational study is to produce subclasses of subjects which result in balance on observed covariates similar to that of a block-randomized experiment. Then, under ceratin assumptions, and armed with appropriate modeling techniques, the researcher can proceed to estimate treatment effects having reduced or even eliminated imbalance on observed variables due to lack of randomization. The persuasiveness and adequacy of the treatment effect estimate then depends in part on the ability of the subclassification technique to recover the covariate balance that would approximate that of a block-randomized experiment. Existing literature on causal inference,

although it offers suggestions, does not provide a clear answer as to how to evaluate the success of a stratification procedure in achieving this goal. The purpose of this dissertation is to explore the utility of balance testing, review the current techniques, explore their statistical properties, and, finally propose an original way to test covariate balance. We start with a brief review of existing research on causal inference in observational studies, issues that have already been confronted, and their relationship to this line of work.

1.2 Overview of Statistical Literature on Causal Inference in Observational Studies

Cochran (1965) elucidated the benefits of learning from reliably planned, measured and analyzed observational studies. This was followed by an expansion of the involvement of statisticians and other scientists in the foundation, sustainment, and advancement of methodology to support causal inference in settings that only approximate carefully designed randomized experiments. Socially important observational studies concern, among other issues, healthcare, education, and economic reforms and have a great potential to shape the quality of our lives. At the same time, they present profound statistical quandaries. One of the early motivating applications for the study of causal inference in observational studies was the relationship between smoking and lung cancer. Claims about this causal relationship were being investigated by Cornfield et al. (1959), Cochran (1968), and others. These causal questions motivated the establishment and development of research focused on observational studies. William Cochran was a major contributor to the literature in this field. In 1965, wrote an influential comprehensive article about observational studies, providing a complete and clear infrastructure for their planning and analysis. The studies of interest for this dissertation have the following two characteristics, as identified in

Cochran (1965):

> (i) The objective is to elucidate cause-and-effect relationships, or at least to investigate the relationships between one set of specified variables x_i and a second set y_i in a way that suggests or appraises hypotheses about causation.
>
> (ii) 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. Some randomization may be employed, however, e.g. in selecting for measurement a random sample from a population that seems suitable for the enquiry at hand.

Cochran (1965) clearly identified the main difficulties in causal inference in observational studies, and I will follow some of his outline in my discussion.

1.2.1 Setting Up the Comparisons and Measurement

When a researcher is organizing a randomized experiment, she can usually plan the treatment and its assignment carefully and deliberately design the experiment to vary important factors in the way that will facilitate estimation of treatment effects at the analysis stage. This power to plan ahead, to design, and to randomize, has the potential to result in obvious and simple comparisons to estimate causal effects. In an observational study, in contrast, a researcher's challenge has often to do with anticipation and solution of design problems as well as with conscientious planning. Existence of an answer to a research question that requires a particular experimental design now may depend on the availability of a proper study environment, which fulfills the necessary conditions and that makes available the necessary comparisons.

Even if that is the case, additional assumptions and adjustments are necessary before such comparisons are made. How to tell whether the study has provided some ground for informative comparisons is precisely what I investigate in the rest of this dissertation.

Cochran (1965) articulates that while there are many potential reasons for trepidation arising from observational data, those observational studies that are substantial and valuable necessitate the development of proper analysis to usefully extract otherwise expensive information. His main concerns have to do, on the one hand, with measurement and data quality, and on the other hand, with appropriate statistical analysis techniques; his apprehension still has resonance today. Many disciplines set out to measure complex and important concepts like "quality of life" and "perceived attitude towards drugs". However, researchers might use different tools to measure the same concepts, and their findings may not necessarily be comparable. Another concern with measurement techniques is measurement error. Non-response and response biases might enter into the study. Variables important to the outcome may be measured incorrectly or not at all, decreasing the effectiveness of later adjustments in outcome analysis. However, the ability of well-designed and accurately measured observational studies still rests on appropriate analysis techniques. Their development will be discussed shortly, following this motivating example.

1.2.2 A Case Study

A fruitful debate resulted from the investigation, in which LaLonde (1986) compares the effects of a randomly assigned job training program on earnings based on a dataset from a field experiment to the estimates one would obtain using techniques developed in econometrics for analyzing observational data. The conclusion of the study was that econometric tools for observational data are not fit to closely ap-

proximate experimental results. What began as a debate about the properties of statistical techniques metamorphosed into a deliberation about data quality; both issues exemplified Cochran's main concerns about analysis of observational data.

The Debate of Lalonde's Conclusions

The program that LaLonde (1986) discusses is the National Supported Work Demonstration (NSW) designed to provide disadvantaged workers with basic job skills and work experience required for them to re-enter the labor market. Qualified applicants were randomly assigned to treatment in this program. Program participants trained at dedicated sites for the duration of the program and then attempted to find regular employment. Information on their earnings was collected as the outcome data, along with prior earnings and demographic variables, which were collected at baseline. After estimating the effect of the program using the experimental NSW dataset, Lalonde attempts to reconstruct the estimate using two other data sources, this time, observational. He uses the treated subjects from the NSW dataset and obtains control subjects from the Panel Study of Income Dynamics (PSID) and the Current Population Survey - Social Security Administration File (CPS-SSA) as his populations for sampling. Lalonde examines the performance of regression, fixed effects analysis and latent variable selection models. He complains that the econometric estimators that he uses to evaluate the effect of the training program from observational data are sensitive to both the composition of the comparison group and to the specification of the econometric model. He also points out that the nonexperimental data at hand do not allow the researcher to check whether the obtained estimates have been properly adjusted for pre-treatment variables (it is a variant of this particular problem that we will discuss and seek to resolve in the following chapters). He generally cautions against usage of observational data

to estimate treatment effects. In fact, his pessimistic conclusions about the quality of estimation of treatment effects from non-experimental data started a very copious debate about the conditions necessary for various causal estimators to produce quality results. For example, Heckman and Hotz (1989), in response to LaLonde (1986) discuss methods that help choose reliable econometric estimators and replicate the experimental results. Following that paper and incorporating innovations in the statistical literature by Rosenbaum and Rubin (1983), Dehejia and Wahba (1999) reanalyze Lalonde's dataset, and claim to have approximated the conclusions from the experimental design. However, their methodology was later questioned in Smith and Todd (2001), who explain that Dehejia and Wahba's result is very dependent on the subsample of subjects used for analysis and the set of covariates used for matching. However, they attribute this poor performance of econometric estimators, to inadequate data quality with inconsistent measurement definitions for earnings between the treatment and control groups and omitted important covariates needed for matching, thus shifting the focus of the debate from the quality of the statistical procedures to data quality. More details of the debate over the conclusions to be drawn from this dataset will be introduced in section 1.3, once we have established some statistical prerequisites. It suffices to say here, however, that the controversy sparked an expansive discussion of observational estimators, their quality, and appropriate circumstances for their use.

1.2.3 The Handling of Disturbing Variables

In the section dedicated to the handling of disturbing variables, Cochran (1965) discusses adjustment for covariates which affect the response other than those under investigation, or the confounders important to the outcome. He mentions the following three "types of weapon" provided to the researcher by experimental design:

1. The researcher is usually able to directly and precisely control the experimental environment, for example, in a lab or in a field. Thus, the experimenter has an opportunity to exclude disturbing variables from design and eliminate or at least lessen their effects.
2. If the researcher has an idea of a covariate that might be important to the outcome, she might choose to block (particularly if the variable is categorical) or otherwise adjust (e.g. regression if the variable is continuous) for this variable at the analysis stage. This should, in principle, help increase precision of the treatment effect estimate.
3. If the researcher has an opportunity to randomize, then, on average, the effect of the remaining disturbing variables, whether they are observed or unobserved, should diminish to the point of little interference in treatment effect estimation.

In an observational study, these "weapons" are not always available, and sometimes have different use. Much work has been done since Cochran's paper to improve the methodology of analysis for causal inference in observational settings. This includes development of methodology to, first, implement adjustments in order to reduce differences on observed covariates due to lack of randomization, second, evaluate the effectiveness of such adjustments to check whether an observational study can approximate a randomized experiment, and, finally, recover this masked experiment and estimate the treatment effects. Cochran's three points as they apply to observational studies today, might be stated as the following:

1. The researcher is sometimes able to control the study environment. For example, a program called Reading First, which was a part of No Child Left Behind Act, is a carefully designed observational study with many financial resources allocated to make sure that the design is implemented correctly. In this study, certain

underachieving schools are assigned to treatment, which is a specialized reading program. The teachers who work at the treated schools all had to undergo rigorous training, and teaching sessions often get supervised to make sure that variability due to teachers is decreased (Carlisle et al. 2006). If the researcher cannot control the study environment, then carefully planning the study may help identify a favorable environment for the observational study.

2. Especially in medical studies, but also in others, much covariate information may be available that is relevant both to the outcome and selection into treatment. Because selection into treatment is usually non-random, the researcher is interested in learning about the selection mechanism in addition to modeling the outcome (Rubin 1991). Rubin (1977) discusses a simple case where assignment to treatment is based on one covariate, and articulates the importance of accounting for the assignment process. He then suggests estimating the treatment effect through learning about the conditional distribution of Y given X in each treatment group and averaging the difference between those conditional expectations. Rubin also discusses several ways to estimate treatment effects after adjustment for the selection into treatment.

   Blocking and matching on pre-treatment variables as well as using them in covariance adjustment have become typical analysis practices. In fact, Rosenbaum and Rubin (1983) came up with a method for combining a multi-dimensional covariate into a uni-dimensional measure which can be used for blocking. However, as will be discussed in Section 1.2.4, these adjustments have a different purpose in an observational study than they do in an experiment. In an experiment, blocking and other adjustments are usually done to increase the precision of the estimates. In an observational study, lack of randomization often results

in different distributions of covariates between groups. Blocking, matching, and covariance adjustment are employed to account for these differences and reduce or eliminate bias in the response variable due to these differences. One of my contributions in this dissertation will be an assessment to help identify if differences on observed covariates have, in fact been reduced to a negligible enough level to proceed with treatment effect estimation.

3. Since the researcher rarely has an opportunity to randomize, the effect of remaining disturbing variables may be bias in treatment effect estimation. The first order of business in this case might have to do with being able to adjust for differences in observed covariates. Measuring the success of these adjustments is the subject of study in this dissertation. If the researcher has successfully adjusted for observed covariates, she has also adjusted for the unobserved covariates to the extent that they are correlated to the observables. Before she proceeds to estimate the treatment effect, she has to assume that there is no unobserved covariates which might bias her conclusions. It is then possible to test the sensitivity of her final conclusions to violations of this assumption; a small discussion and some references on sensitivity analysis are provided in section 1.3.2.

1.2.4 Adjusting for X - Survey of Early Ideas from Cochran and Rubin

As a necessary part of estimating the effect of treatment, we seek to adjust for confounders that matter to the outcome. Under random assignment we expect the treatment and control groups to be similar, on average, on those confounders, whether or not they are observed and measured, at least in samples of reasonable size. In the observational framework, since such similarity is not guaranteed, and often is absent, we seek to carefully adjust for observed covariates. Competing approaches for such

adjustments have emerged in the causal inference literature over time; some of them are presented below. In connection to an adjustment mechanism, there emerges a need for a technique that can be used to appraise that mechanism's success. Defining what a successful adjustment means and how to appraise such an accomplishment is one of the main goals of this dissertation; this information will be spanned in Chapters 2 and 3. As we discuss the evolution in the mechanisms to adjust for observed covariates, we also focus on the progression of measures to evaluate the mechanisms' performance.

Stratification

With the goal of investigating the causal effect of smoking on death rates, Cochran (1968) presents data on death rates for different classes of smokers, including the main confounding variable - age. In this paper Cochran demonstrates adjustment by subclassification on a discretized version of a continuous covariate, age, in this case, and its effectiveness. One of the main contributions of the paper is Cochran's suggestion that if a continuous confounder is normally distributed and has the same variance in the compared populations, then stratification into five or six strata on that covariate can remove at least 90% of the difference in means of that covariate for the two groups (which Cochran refers to as bias on that covariate). This paper also explores the idea of "the percent reduction in the bias of X due to matched sampling" as a way to appraise the success of a stratification technique.

Matching for Robustness

In the event that a large pool of control subjects is available and if collecting outcome information on controls is expensive, one might consider only following up those control subjects that closely resemble treated subjects in important respects.

Matched sampling from the control reservoir can be useful for this purpose. An appealing goal of matching and stratification is that outcomes are examined for *comparable* subjects, at least in terms of observed covariates. In their discussion of matched sampling, Althauser and Rubin (1971) differentiate between good and bad matches. Like Cochran (1968), they use proximity of matched subjects on the matching measure and percent reduction in bias on a covariate as a way to evaluate the quality of the match. As matched sampling became a common method to choose control subjects, the question of how best to match treated and control subjects became increasingly relevant. Cochran and Rubin (1973) discuss matched sampling and regression adjustments for observed confounders, as well as their combinations. With the goal of improving subjects' comparability they evaluate matching within calipers on a covariate $\mathbf{X}$ and the nearest available metric matching algorithm, extending these to a multivariate setting. They propose matching on the best linear discriminant or a Mahalanobis distance metric to facilitate adjustment for several covariates simultaneously. The authors emphasize that the main focus of matched sampling in observational studies is reduction of bias in the response due to observed covariates, as opposed to randomized experiments, where blocking might be used to increase the precision of the treatment effect estimate. They conclude that a combination of matching and regression adjustment appears to be superior to either of the two techniques when used by themselves, in terms of reducing bias. In a paper entitled "Matching to Remove Bias in Observational Studies", Rubin (1973b) evaluates several matching methods on their ability to remove bias of the matching variable. To evaluate the differences between matching methods, he uses Cochran's formulation "the percent reduction in the bias of X due to matched sampling". Rubin concludes that compared to mean-matching, the nearest available pair-matching method based

on a random ordering results in closer matches and is thus superior, but also points out that with the help of a computer, one might perform several matchings and pick the best one. To define 'best', Rubin suggests recording for all matched samples $\bar{d} = \bar{x}_{1.} - \bar{x}_{2.}$ and $\bar{d}^2 = 1/N \sum (x_{1j} - x_{2j})^2$ (1 and 2 here are the two groups and j is the subject index) and minimizing first the first quantity and then the second. In the conclusion, the author posits that a more complex matching algorithm might result in better matches than the ones he discusses. In the extension of this paper, Rubin (1973a) concludes from a Monte Carlo study that a combination of matched sampling with regression adjustment tends to result in the least biased estimates as compared to either approach on its own. Rubin (1979) uses another Monte Carlo study to conclude that the effectiveness of the combination of multivariate matched sampling and regression adjustment holds for reducing imbalance on a bivariate $\mathbf{X}$, even when the response may not be linear in $\mathbf{X}$. His measure of appraisal is percentage reduction in bias in the response due to covariate. "Bias reduction using Mahalanobis-Metric Matching" (Rubin 1980) presents a Monte Carlo study with a focus on the ability of a nearest-available Mahalanobis-metric matching algorithm to bring closer together the means of matching variables within matched samples. It is noteworthy that here Rubin introduces another idea of measurement of the success of a matched adjustment - comparison of the means of variables used on the match.

Observed Covariates and the Assignment Mechanism

Rubin (1977) explores the possibility that in observational studies the non-random selection into treatment can be based on one or more covariates. In this paper Rubin talks about the assignment to treatment as it pertains to observational studies. Although there isn't a literate assignment to treatment in observational studies, the author gestures towards a probability model in which a random variable is to be

used to model assignment to treatment. He examines a simplified situation in which selection into treatment is based on a single covariate X. Rubin's solution is for the researcher to focus on estimating the conditional expectation of Y given X in the treatment and control groups. He suggests that this can be done either by regression adjustment (if the assumptions are met) or by blocking on X, noting that the coarseness of blocking affects the quality of the estimate. He also points to the importance of the overlap in the distribution of X. This would turn out to be one of the foundational papers in what is now known in Statistics as the Rubin Causal Model.

1.2.5 Potential Outcomes and the Rubin Causal Model

Though other approaches for treatment effect estimation have been suggested in, for example, Pearl (1995), Pearl (2000) and Dawid (2000), the potential outcomes approach to causal inference is the fundamental approach of interest in this dissertation. In the paper "Estimating the Causal Effects of Treatments in Randomized and Nonrandomized Studies", Rubin (1974) discusses the idea of potential outcomes (Neyman (1990), Cox (1958)) with application to models of treatment effects, which later became the heart of an important model of causation, labeled by Holland (1986) as the Rubin Causal Model (RCM). Consider a binary treatment assignment indicator Z, such that each subject can only get assigned to treatment ($Z = 1$) or control ($Z = 0$). Suppose also that the assignment to treatment happens at time t_1 and the outcome is observed at a later time t_2. If a subject is assigned to treatment ($Z = 1$) at time t_1, define Y_t to be her response measured at time t_2. Similarly, define Y_c to be the observed response for our subject at time t_2 if she had been assigned to control ($Z = 0$) at time t_1. Then, Y_t and Y_c are the subject's potential outcomes, and the causal treatment effect for this subject is simply $Y_t - Y_c$. Here, Rubin also

states what would later be known as the Fundamental Problem of Causal Inference (Holland 1986): Y_t and Y_c can never be observed for the same subject, since each subject only gets one treatment assignment at time t_1. In a population of such subjects, we might be interested in estimating the Average Treatment Effect (ATE) or Effect of Treatment on the Treated (ETT):

$$\text{ATE} = E[Y_t] - E[Y_c] \tag{1.1}$$

$$\text{ETT} = E[Y_t - Y_c] \mid Z = 1] \tag{1.2}$$

In the case of random assignment to treatment, estimating ATE or ETT is very straightforward, in fact, even the simple difference in sample means is an acceptable unbiased estimator. Under any treatment assignment mechanism, if the important characteristics are balanced or similar between the treatment and control groups, then any observed difference in outcome has the potential, with proper analysis, to be credited to either chance or the difference in treatment regimes, and the treatment effect can be estimated in straightforward ways without bias. In any particular realization of the treatment assignment, we are likely to observe imbalances, by chance or otherwise. Rubin discusses at length the importance of matching and / or adjusting for variables that might have an impact on the response, in both randomized experiments and observational studies. He suggests that a non-randomized study that has been carefully designed and controlled can give way to similar conclusions that might be reached in an analogous experiment. As an example, Rubin (1978) gives a manufacturing setting, in which one is likely to know and be able to adjust for all important confounders.

The Fundamental Problem of Causal Inference

As defined in Holland (1986), the fundamental problem of causal inference is that once treatment assignment has occurred, each subject is assigned either to treatment or control, so only one of the two potential outcomes is observed. For the treated subjects, we observe their Y_t, and for control subjects, we only see their Y_c. In a simple observational study with four subjects, we might observe data like that summarized in Table 1.2.5. All four subjects i have potential outcomes Y_{it} and Y_{ic}. However, after treatment Z is assigned, subjects 1 and 2 are chosen to be in the treated group, leaving subjects 3 and 4 in the control group. We thus observe Y_{1t} and Y_{2t}, as well as Y_{3c} and Y_{4c} rather than all 8 potential outcomes. Then the main ingredients to estimate the treatment effect $Y_{it} - Y_{ic}$ are not available for any subject i, although we might have enough information to estimate the average treatment effect under the assumptions below.

Subject	Potential Y_t	Potential Y_c	Z	Observed Y_t	Observed Y_c	Causal Effect $Y_{it} - Y_{ic}$
1	Y_{1t}	Y_{1c}	1	Y_{1t}	?	?
2	Y_{2t}	Y_{2c}	1	Y_{2t}	?	?
3	Y_{3t}	Y_{3c}	0	?	Y_{3c}	?
4	Y_{4t}	Y_{4c}	0	?	Y_{3c}	?
Mean				$\frac{Y_{1t}+Y_{2t}}{2}$	$\frac{Y_{3c}+Y_{4c}}{2}$	$\frac{Y_{1t}+Y_{2t}}{2} - \frac{Y_{3c}+Y_{4c}}{2}$

Table 1.1: Fundamental Problem of Causal Inference in Terms of Potential Outcomes

1.2.6 Assumptions

Stable Unit Treatment Value Assumption

The Stable Unit Treatment Value Assumption (SUTVA) is usually a part of matching and stratification frameworks and simplifies treatment effect estimation. SUTVA requires that the potential outcomes for any particular unit be independent of treatment assignment for all other units, that is to say that Y_t and Y_c for a particular unit remain the same regardless of the treatment allocation vector. So if subject

i is assigned to treatment (or control), we will observe her same Y_{it} (or Y_{ic}) independently of everyone else's treatment status and of the assignment mechanism (Rubin 1986). SUTVA is usually assumed, but its validity is rarely tested. In practice, interference between units can result in SUTVA being violated (Rubin 1990).

Strong Ignorability of Treatment Assignment

The Conditional Independence Assumption (CIA), introduced as Strong Ignorability of Treatment Assignment by Rosenbaum and Rubin (1983), assumes selection into treatment based on observed covariates. This assumption is critical in matching, stratification, and covariance adjustment. It asserts that among subjects with the same pre-treatment characteristics $\mathbf{X}$, selection into treatment groups is random, and is not based on the benefits of different treatments. If we imagine subclassifying subjects into strata based on the discrete covariate $\mathbf{X}$, this assumption would imply that within the stratum in which $\mathbf{X} = \mathbf{x}$, treatment assignment Z is determined as in a randomized experiment within stratum $\mathbf{X} = \mathbf{x}$, as a Bernoulli random variable with probability of success $P(Z = 1 \mid \mathbf{X} = \mathbf{x})$. This creates a link between the observational study and a block-randomized experiment. More formally, the Strong Ignorability of Treatment Assignment is stated as (Rosenbaum and Rubin 1983)

$$(Y_t, Y_c \perp\!\!\!\perp Z \mid \mathbf{X}) \tag{1.3}$$

Along with the Strong Ignorability assumption, (Rosenbaum and Rubin 1983) also present the Common Support Condition:

$$0 < P(Z \mid X) < 1 \tag{1.4}$$

If these assumptions hold, then the following identity holds for estimating the Average Treatment Effect and the Effect of Treatment on the Treated using the observed

covariate $\mathbf{X}$:

$$\begin{aligned}\text{ATE} = E[Y_t] - E[Y_c] &= E\left[E[Y_t \mid \mathbf{X}] - E[Y_c \mid \mathbf{X}]\right] \\ &= E[E[Y \mid Z = 1, \mathbf{X}] - E[Y \mid Z = 0, \mathbf{X}]];\end{aligned}$$

and

$$\begin{aligned}\text{ETT} = E[Y_t - Y_c \mid Z = 1] &= E[E[Y_t \mid Z = 1, \mathbf{X}] - E[Y_c \mid Z = 0, \mathbf{X}] \mid Z = 1] \\ &= E[E[Y \mid Z = 1, \mathbf{X}] - E[Y \mid Z = 0, \mathbf{X}] \mid Z = 1] \\ &= E[Y \mid Z = 1] - E[E[Y \mid Z = 0, \mathbf{X}] \mid Z = 1]\end{aligned}$$

1.2.7 Developments in X-Adjustment Techniques

Cochran (1965) expresses concern about having to adjust for many potentially causal or "important disturbing variables". He draws attention to the importance of exploring new multivariate techniques. Since then, several dimension-reduction techniques have appeared in the literature; of them the propensity score (Rosenbaum and Rubin 1983) will be discussed and utilized in this work. As they discuss the advantages of reducing the dimension of the covariates, Rosenbaum and Rubin (1985) point out that although it is most intuitive to subclassify on categories of individual covariates, as the dimension of X increases, that kind of an approach quickly becomes infeasible. They give the example of 20 covariates each with just 2 categories. This results in $2^{20} \doteq 1$ million categories. So exact matches are hard to find, and it is difficult to categorize what an approximate match means in this context. This problem became known as the curse of dimensionality. It necessitates that the dimension of discrepancies between subjects be reduced before subclassification or matching.

X-Metrics

It is desirable that closeness of two observations be measured on multiple covariates simultaneously. For this purpose, the Mahalanobis metric can be utilized. For a given set of covariates $x = (x_1, x_2, \ldots, x_p)^T$ with mean $\mu = (\mu_1, \mu_2, \ldots, \mu_p)^T$ and a covariance matrix Σ, the Mahalanobis distance is defined as

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}.$$

This distance measure takes into account the covariance structure among the variables and is scale-invariant. For each pair of treated and control subjects, a Mahalanobis distance on observed covariates is computed. The resulting distance matrix summarizes the proximity of each treated to each control subject and facilitates comparison of units similar to each other on variables included in the Mahalanobis distance computation. Rosenbaum and Rubin (1985) show that under the assumptions presented above, matching on a Mahalanobis distance on covariates within calipers of the estimated propensity score (discussed below) reduces standardized differences on covariates, their squares and crossproducts more than any alternative matching distances.

Balancing Scores

The goal of observed covariate adjustment can be interpreted as to homogenize the distribution of X between the treated and control groups. This is the experimental benchmark which non-experimental studies strive to accomplish. Once the covariate distribution is similar for the two groups, suitable adjustments should result in reliable estimates of treatment effects. Adjustments based on balancing scores are made with this experimental benchmark in mind. A balancing score $b(X)$ is defined as a function of the observed covariate X, such that the distribution of X is the same

for the treated ($z = 1$) and control ($z = 0$) group given the $b(X)$ (Rosenbaum and Rubin 1983). Formally,

$$X \perp\!\!\!\perp Z \mid b(X) \tag{1.5}$$

The simplest balancing score is X itself.

Reducing the Dimension of X - the Propensity Score

The *propensity score* is the conditional probability of receiving treatment rather than control given observed covariates (Rosenbaum and Rubin 1983). The propensity score is the coarsest balancing score. It can be specified as:

$$e(X) = P(z = 1 \mid X) \tag{1.6}$$

Propensity-based analyses model treatment as stochastic, and assume that a researcher, armed with data and modeling techniques, can build a correct enough model of treatment assignment - model of $Z \mid \mathbf{X}$ which will reduce bias in estimation of treatment effects to substantively negligible levels. In a randomized experiment with 2 groups of equal size, the propensity score would be $\frac{1}{2}$ for all subjects regardless of covariates. Similarly, in a matched pairs randomized experiment (thinking of pairs as blocks), within each block the propensity score would be $\frac{1}{2}$ for each subject. In an observational study, the propensity score tends to vary with covariate information and be higher for treated subjects as compared with control. Overlap of the distributions of the estimated propensity scores for the treated and control groups is often used as a criterion for selection into the matched sample as an assessment of comparability of treated and control subjects.

True propensity scores are usually unavailable. The propensity score is frequently estimated using logistic regression of the treatment variable on the observed covariates. If an adjustment for all observed covariates is sufficient for unbiased estimation

of treatment effect, then so is adjustment for the propensity score alone (Rosenbaum and Rubin 1983). Estimated propensity scores behave like true propensity scores, in the sense that matching treated and control subjects on the estimated scores still tends to balance covariate distributions between the treatment and control groups. In fact, matching on the estimated propensity score rather than the true propensity score can be useful to adjust for imbalances that occur in a sample by chance, due to sampling variability (Rosenbaum and Rubin 1983). Hansen (2008b) and Hansen (2009) engage the discussion of matching on the estimated rather than true propensity scores. Both papers posit that since matching exactly on the true propensity score should balance covariates that went into its estimation, then assessments of post-matching covariate balance can be viewed of goodness-of-fit tests of this model to the data. Hansen (2009) substitutes the issue of trying to characterize the uncertainty due to the estimated propensity score for the issue of inference under the assumptions that the subjects within matched sets are not too far off on the true propensity score and that covariate balance is maintained, offering approximations to certain sampling distributions under these conditions.

Often, the logit (the linear predictor) of the propensity score is used for matching or stratification, rather than the raw probability of being in the treatment group because those probabilities are compressed (Rosenbaum and Rubin 1985). We will follow that convention in this dissertation when we talk about matching and stratification on the estimated propensity score. Rosenbaum and Rubin (1983) and Rosenbaum and Rubin (1984) show that subclassification on the propensity score should balance the observed covariates that went into its estimation, or given the propensity

score e the covariate x is independent of the treatment assignment z:

$$P(x, z \mid e) = P(x \mid e)P(z \mid e)$$

Within subclasses that are homogeneous on the propensity score (which can only strictly occur with discrete $\mathbf{X}$), the joint distribution of observed covariates should be the same between treated and control subjects. This gave rise to various types of usage of propensity scores, including stratification and matching, which will be the main techniques of interest in subsequent discussions. In particular, measuring success of subclassification on the propensity score with respect to covariate balance is the main question addressed in the following chapters.

Some limitations of the propensity score are discussed in Rubin (1997). In summary, they are:

- They only help adjust for observed covariates, and unobserved to the extent that they are correlated with observed.
- They work better in large samples. The distributional balance of observed covariates that is supposed to result from stratification on the propensity score is an expected balance. So in small samples, large chance imbalances are possible despite stratification. Also, since propensity scores are usually modeled using regression techniques, limitations of regression analysis in small samples (e.g. problems of overfitting) apply to propensity-based methods.
- Covariates related to treatment assignment and not the outcome are treated the same as the ones strongly related to the outcome and not treatment. Inclusion of irrelevant variables might undermine the efficiency of controlling for relevant variables, but Rubin and Thomas (1996) discuss this and show that in moderately large and large studies the bias from leaving out weakly-predictive

covariates dominates the efficiency gains from their omission.

Another possible limitation to using the propensity score is that the parametric propensity model can easily be misspecified. Misspecification of the propensity model might be due to, among other causes, omission of important terms, or incorrect assumptions about the form of the relationship between covariates and the response. For example, most users assume that they are related by a logit function, which does not need to be the case. Such misspecification is difficult to diagnose, and the consequences of it are elusive.

Post-stratification and Matching Using the Propensity Score

Cochran (1968) discussed stratification on one continuous covariate, leaving an open question about how to apply such a technique when there are multiple confounders. The dimension-reducing property of the propensity score can be utilized to help alleviate several complications that can arise at the post-stratification stage of analysis of a study with a high-dimensional covariate space. The literature points to usage of the propensity score for assessment of subjects' comparability, subclassification, and as tool for manipulation of covariate balance. As adjustment methods incorporating the propensity score evolved, so did the procedures for appraising their efficacy. Following Cochran's ideas on stratification on a continuous covariate, Rosenbaum and Rubin (1984) discuss subclassification into five strata based on quintiles on the estimated propensity score and provide an empirical example of such an approach. The authors present boxplots of the estimated propensity scores for the treatment and control groups, illustrating overlap on the estimated propensity score as a way to demonstrate the comparability of treated and control subjects on their observed covariates. To measure the usefulness and 'success' of their subclassification technique, the authors look to the improvement in covariate balance. Covariate

balance is evaluated using F-tests, by conducting a two-way ANOVA for each covariate using as factors treatment and subclassification. They refer to the F-test as an "approximate test of the adequacy of the model for the propensity score", "approximate primarily because the subclasses are not exactly homogeneous in the fitted propensity score". In case of a failure of this balance assessment, the authors respecify the propensity model to include squares of unbalanced variables and their cross-products with other important variables, illustrating one way to iterate fitting of the propensity model and testing for balance to make more convincing inferences.

Rosenbaum and Rubin (1985) present the propensity score as part of a methodology for multivariate matched sampling. They investigate several matching methods, including those which utilize the estimated propensity score as a distinct matching variable. The methods investigated include nearest available matching on the estimated propensity score, Mahalanobis metric matching including the propensity score, and nearest available Mahalanobis metric matching within propensity calipers. The authors conclude that matching on this metric within propensity calipers was successful in balancing covariates, their squares and cross-products. To measure covariate balance in this paper, the authors use a measure of standardized bias on individual covariates, giving criteria for its adequacy. The definition of standardized bias and details of this approach to measuring balance will be discussed in the following chapter. For a theoretical justification and simulation-based conclusions about the performance of propensity score matching its reliability as an adjustment method in a wide range of conditions, we point to Rubin and Thomas (1996). For the rest of this chapter, when we talk about stratification and matching, we will narrow our scope on the kind of methods presented by Cochran and Rosenbaum, which split up treated and control subjects into subclasses, and where the modeling portion of the

estimator is limited just to estimating the propensity score. In contrast, techniques like kernel matching, which are common in parallel fields, combine matching with regression modeling in the estimation process.

More and Less Important Covariates

Rosenbaum (1986) gives an example of using matching on a propensity score-like measure to estimate the effect of dropping out on cognitive achievement test scores. This analogue of the propensity score is modeled using a selection of observed covariates, focusing the reader's attention on the idea that not all covariates are equally important for either outcomes or treatment assignment. In this study, the subjects are sequentially matched by a multivariate procedure within schools. Matched pair differences are further adjusted using regression to estimate the treatment effect. This is done to remedy some imbalance on covariates remaining after matching, according to Rosenbaum, due to large initial differences between groups and to the limitation that subjects had to be within schools.

An important contribution of the paper is the distinction between more and less important confounders. It can also be found in Rubin and Thomas (2000), a discussion of nearest available matching, in which authors conclude that in addition to matching on the estimated propensity score, it is also useful to match on a Mahalanobis metric of covariates important to the outcome. They emphasize that the actual estimated propensity score should be balanced between the two groups, rather than focusing just on balancing important prognostic covariates, even if those covariates account for majority of variation in the outcome. This concept is then further developed in Hansen and Bowers (2008) and Hansen (2008a), who suggest that when it comes to balancing observed covariates, balancing those strongly related to the outcome is more critical than others.

Advances in Matching

Several matching algorithms have evolved in order to optimize the quality of a matching procedure, a dimension of which can be summarized by a measure of proximity of the matched subjects on important variables, like the propensity score. One way to minimize the distance between matched comparison units on the propensity score is through matching with replacement. According to this algorithm, each treatment unit can be matched to the nearest control unit, even if that comparison unit is matched more than once. Because this approach can provide closer matches on the propensity score than nearest-available matching without replacement, it can be beneficial for reducing bias, but, often, at the price of decreasing precision. Another way to minimize a distance between treated and control units within matched sets but matching without replacement is optimal matching. It can be opposed to "greedy" matching, where the closest control match for each treated unit is considered one at a time, without minimizing a global distance measure. The optimal matching algorithm would instead seek to match subjects to minimize a global discrepancy measure, like the sum of distances within matched sets. Rosenbaum (1991) develops the idea to improve matching algorithms with the goal of optimizing the overall similarity of matched subjects. "Full" matching is contrasted with pair or 1:k matching, because it allows for matched set structures to include one treated subject with one or more controls or for one control to be matched with at least one treated subject. The flexibility of this matching algorithm can result in using more of the data at hand and yield more effective comparisons (in terms of effective sample size) and closest-possible matches on any given distance (Rosenbaum 1991). As shown in Hansen and Klopfer (2006), full matching results in improvements in efficiency and bias over greedy matching algorithms. Several matching methods, including full

matching are evaluated in a simulation study by Gu and Rosenbaum (1993). The authors conclude that full matching on the propensity score improves balance better than other methods in datasets with many covariates and large systematic biases. Also, they conclude that generally, full matching results in closer matches than the greedy algorithms, like nearest-neighbor matching. Although the two types of algorithms tend to select more or less the same controls, they assign them to different treated units, affecting distance within pairs but not balance overall. As in Rosenbaum and Rubin (1985), Gu and Rosenbaum (1993) conclude that matching on a Mahalanobis distance metric within propensity calipers is the method of choice for controlling chance covariate imbalances.

As an example of an application of full matching, we look at the paper by Hansen (2004). In this comparison of SAT scores for coached and uncoached students, the author illustrates the bias reduction in observed covariates that results from using optimal full matching with restrictions to subclassify students into comparable matches. The matching strategy employed in the study involves first stratifying on covariates highly predictive of treatment status, and then, within the strata, full matching on the propensity score, while also incorporating information on missing data. The author uses the standardized bias approach from Rosenbaum and Rubin (1985) to appraise balance on observed covariates. Diligent accounting for variables related to treatment assignment and the incorporation of all the available control subjects in the full matching procedure results in a 99% reduction of bias on the estimated propensity score as well as a considerable decrease in bias on measured confounders to statistically insignificant levels.

1.2.8 Balance Testing

Matching and stratification are often used in observational studies with the goal of creating matched sets which result in covariate balance between the treatment and control groups, that is to create similar distributions of observed covariates between groups. With proper adjustments for the matching or stratification structure, one can reduce overt biases due to imbalances on observed covariates and proceed with estimation of treatment effects. Appraisal of the extent to which matching or stratification have subclassified subjects into homogeneous groups is one of the roles of a balance test. As seen in the previous section, the literature suggests that although differences on covariates after stratification or matching have been used since Cochran (1968) to appraise the success of a stratification technique, defining precisely what 'success' means has proven to be a difficult task.

Some suggest using significance-based hypothesis tests (see, for example Rosenbaum and Rubin (1984) or Hansen and Bowers (2008)) to distinguish between balance and imbalance, but this approach also has opponents. Senn (1994), for example, argues that testing balance in this way is misleading because any realization of assignment, randomized or not, is imbalanced to some extent. He further argues that even covariates that appear 'balanced' should still be included in subsequent outcome analysis, whereas obtaining 'balance' on a covariate can lead researchers to not include this information in the outcome model. Sekhon (2007) and Imai et al. (2008) criticize methods that many researchers use to appraise balance and also suggest that significance-based balance testing is delusory and that whatever metric a researcher chooses for measuring balance, that metric should be maximized without limit. Lee (2008), Hansen and Bowers (2008) and Hansen (2008b) oppose this view and argue in favor of balance testing as a way to evaluate the success of a stratification technique,

as they evaluate existing techniques and suggest improvements. This discussion is carried out in full in Chapter 2.

1.3 Related Contributions

Several parallel research fields have uses for analogous techniques in application to related but different data issues. A brief survey of contributions from those fields is in order. The nature of economic reforms and the necessity to evaluate governmental economic programs led to the development of a parallel literature and competing approaches in economics. Similarly, in epidemiology and the medical sciences, observational studies address questions, answers to which cannot ethically be obtained from randomized experiments. The realization of the capacity of what could be learned from these data if they are properly analyzed, inspired discussions and publications which contribute significantly to the causal inference literature. For example, as early as the 1970s, Heckman (1978) and Heckman (1979) discussed a two-stage regression-based estimator for behavioral relationships in non-random samples, calling it the analysis of 'sample-selection bias'. Some techniques common to statistics and econometrics, received mixed reviews in economics and other areas. LaLonde (1986) compares the effects of a randomly assigned job training program on earnings based on a dataset from a field experiment to the estimates one would obtain using techniques commonly used in econometrics for analyzing observational data only to conclude that econometric observational tools are not able to approximate experimental results.

The Debate on Lalonde's Conclusions, Continued

LaLonde (1986) points out that while econometric estimators fail to reproduce experimental results even when the models pass specification tests, the estimator

from Heckman (1978) outperforms one-stage estimators. The validity of his estimates would later be questioned by Smith and Todd (2005a). Heckman and Hotz (1989), in response to LaLonde (1986) discuss specification tests that are meant to help choose a reliable econometric estimator. Heckman and Hotz argue that the use of these specification tests helps eliminate estimators that are unreliable and misleading, and instead results in estimators able to closely replicate the experimental results.

Following that paper and integrating the propensity score (Rosenbaum and Rubin 1983), Dehejia and Wahba (1999) reanalyze Lalonde's dataset, using stratification on the estimated propensity score. They discard controls that they deem not comparable to their treated subjects, based on the overlap on the estimated propensity score and make an argument that Lalonde failed to replicate experimental results because he used many controls which are unsuitable based on their approach to subject comparability. The authors utilize an original way of testing covariate balance. They conduct t-tests for each covariate within each stratum, and if any of them are significant, they respecify the propensity model by including higher order terms for the covariate in question, and stratify again. The process is repeated until none of the covariates in any of the strata test to be significantly different. At that point, the specification of the propensity score is accepted and the authors proceed to matching with replacement before estimating the treatment effect, with the resulting estimator approximately replicating experimental results. Heckman et al. (1997) and Heckman et al. (1998) evaluate the performance of matching estimators in another job training dataset and define data conditions which they consider to be necessary for reliable treatment effect estimation, such as same data and measurement sources for control and treated group, same geographic location for both groups (this condition is offered as specific to active labor market context), and a sufficiently rich set of co-

variates. Smith and Todd (2001) tend to side with these conditions and go on to say that a broader econometric analysis of the NSW dataset results in biased estimators. In response, Dehejia and Wahba (2002) confirm their earlier findings and maintain that using propensity score as way to discard irrelevant controls in combination with matching with replacement (with replacement because there are few valuable controls) tends to reproduce experimental results. Smith and Todd (2005a) insist that Dehejia and Wahba's results are contingent on their choice of sample and dispute the advantages of propensity score matching in this application, indicating a preference for difference-in-differences estimation combined with matching (Heckman et al. (1997), Heckman et al. (1998)). Dehejia (2005) insists on the rationale for choosing their samples and discusses the importance of selecting a proper propensity model specification and enforcing the common support condition. Despite their differences, Dehejia and Wahba and Smith and Todd agreed on the fact that it would be useful to have a balance test that would help to evaluate the suitability of a propensity score specification for a given study. In 'Rejoinder' Smith and Todd (2005b) introduce another variant of the balancing test, which will be discussed in the next chapter. They discuss three techniques for evaluating balance and note their poor performance as well as restate the need for research in this general area. Smith and Todd also point out the small size of the dataset, the lack of usable controls and the fact that the Conditional Independence Assumption might not hold because the set of measured covariates is likely lacking important confounders. They insist that propensity score matching does not solve the problem of selection bias in the NSW study.

Some methods introduced in economics as variants of propensity score matching involve kernel-based matching methods, and general framework as matching as an application of non-parametric regression methods. Heckman et al. (1998) show these

estimators to be as good and sometimes better than nearest-neighbor matching estimators in terms of reducing the variance or resulting estimator. This is an example of a difference in the development of these methodologies; whereas statisticians have mostly focused on reducing bias on observed covariates (Cochran (1968), Althauser and Rubin (1971), Rosenbaum and Rubin (1983), and others), in economics and epidemiology, researchers have been more actively focused on reducing the variance of the resulting estimator, for instance for weighting estimators (Hirano et al. 2003) and kernel matching (Heckman et al. 1998). Frölich (2004), for example, investigates several estimators in finite and especially small samples, an investigation that concludes that for pair matching,, while ridge and kernel matching performed the best and nearest-neighbor matching performed the worst, in terms of MSE, again, exhibiting a particular interest in precision. Economists Abadie and Imbens (2002) provide an estimator for the conditional variance of a matching estimator. Other contributions involve an extensive literature on weighting and evaluation of its performance, such as in Hirano and Imbens (2001), a paper that discusses the usage of weights to analyze the effects of Right Heart Catheterization, a response to an earlier paper by Connors et al. (1996a). An important result is the poor performance of inverse-probability weighting using the propensity score in small samples, discussed in detail Kang and Schafer (2007).

Very recent work by Busso et al. (2009) investigates the finite sample properties of propensity score matching and reweighting estimators. They consider with-replacement pair matching and regular estimators which are root-n consistent. The authors establish a semiparametric efficiency bound for a particular class of estimators, especially important to samples with a small-dimensional covariate. They conclude that pair matching is not efficient in relationship to this bound and tend

to favor reweighting estimators.

A comparison of stratification and weighting on the propensity score is presented in Lunceford and Davidian (2004). The authors conclude that stratification into a fixed number of strata (the authors consider five) may not result in sufficient bias reduction, whereas a combination of weighting on the propensity score with a semi-parametric regression estimator from Robins et al. (1994) resulted in efficient and precise estimation. Kang and Schafer (2007) show that these doubly-robust techniques for inverse-probability weighting are quite sensitive to misspecification of the propensity model when some estimated propensities are small, implying that use of this technique is not desirable in small samples. Another contribution from the area of medicine is by Austin et al. (2007), who address the question of which variables should be included in the estimation of the propensity score, and conclude that it is most advantageous for precision to include true confounders (variables associated with the treatment assignment and outcome) since this approach appears to result in more matched pairs than others. But to simultaneously maximize balance, and minimize bias and MSE, the authors recommend including all potential confounders.

1.3.1 Other Developments

Design

The literature on the quality of observational studies continued to evolve since Cochran's contributions; one of the more recent papers on successfully designing observational studies to be able to distinguish between treatment effects and biases is due to Rosenbaum (1999). He emphasizes the importance of carefully choosing a narrow and focused hypothesis, the necessity and benefit of a control group, the proper way to define treated and control groups, and other important aspects of planning an observational study.

Causal Inference Using Regression

Cochran (1965) mentions problems with using regression to estimate causal effects. His main concern is that one might confuse correlation with causation because regression does not actually speak to the direction of the causal path. In 1969, Cochran addresses a concern he posed in his 1968 paper, the concern of having to adjust for more than one covariate at once. He presents a study presented in Belson (1956), and discusses the circumstances optimal for Belson's approach to adjust the difference in mean outcomes by regression on disturbing variables in the control group, deeming regression an acceptable adjustment method for multiple confounders given certain data conditions.

Since then, much literature has been dedicated to discussion and critique of the usage of regression for causal inference. Berk (2004) dedicated a large part of his book to discussing controversy over causal inferences drawn from regression. He views the goal of a causal analysis as learning about the distribution of the state of the unit with and without the intervention. This is very similar to the potential outcomes framework, in which causal effects cannot be observed and must therefore be inferred. Berk's particular concerns are checking both the plausibility of a regression model in relation to the data in general and, once that has been established, checking regression assumptions. Although he points out the advantages of regression as a descriptive tool that can be used to characterize the conditional distribution of response using a set of covariates, many questions cannot be answered using a combination of observational data and regression analysis. It is factors outside of the data which determine plausibility of causal inference. Still, if it can be established that there is, in fact, a causal effect to be estimated, he points out that regression is more of a descriptive tool and that "regression analysis of $y \mid x$ is absolutely silent on

whether any observed patterns are causal." Berk (2004) does mention that "simple linear regression can sometimes be used to estimate causal effects once a definition and supporting rationale have been provided. This "supporting rationale refers to having information about the data-generating process and to the plausibility of a regression model, as well as to checking regression assumptions. Berk mentions that useful diagnostic procedures are often omitted from published papers, but also insists that diagnostics are generally oversold since they assume that the model is almost correct, allowing for a minor fixable departure. Neither a good overall fit, specification tests, nor regression diagnostics can demonstrate causal effects. Berk emphasizes that if the researcher is not armed with a model that is nearly correct, then even more complicated regression models may not be able to solve potential problems with causal estimators. Freedman (1991) points out that we should not count on regression to carry much of the burden in a causal argument. His position is similar to Berk's in that, especially in social sciences, the researcher is unlikely to start with a correct model, and thus making causal arguments based on statistical significance of the coefficients from that model, is erroneous in most cases. He points out that even given a model with good fit and diagnostics, it is possible that this model did not approximate the actual data-generating process, since significance of coefficients relies on specification of the model, especially the error structure. He suggests that the model needs to either be empirically tested or it needs to rely on very strong theoretical arguments to be valid. Freedman's fear appears to be that "regression makes it all too easy to substitute technique for work. He adds that little is done in current work to test underlying model assumptions. Finally, he also emphasizes that sophisticated regression techniques are not a substitute for data collection and good study design.

Incorporating Information About Outcomes

The causal inference literature has produced many more rich ideas that go in various directions, including ways to supplement matching on the propensity score. Zhao (2004) investigates different matching strategies and concludes, among other things, that in small samples, the matching estimators based on the propensity score do not perform as well as some other estimators, according to Zhao, because the variance of the estimator starts dominating the bias. He also notes that propensity matching does perform well in sufficiently large samples and when selection on observables is strong. Zhao (2004) makes general favorable conclusions about Mahalanobis metric matching and its robustness and MSE. He also suggests a possible benefit of incorporating outcome information into the matching metric, a concept related to the prognostic score developed by Hansen (2008a). Hansen (2008a) suggests usage of prognostic scores in complement to propensity score adjustments. The prognostic score is analogous to the propensity score in that it is a unidimensional summary of observed covariates. In contrast to the propensity score, the prognostic score measures the relationship between observed variables and potential outcomes. First, the outcome is modeled just in the control group. Then, the obtained model is used to predict the response in the treated group, using the coefficients from the control group. The fitted values from the model are the prognostic score - the predicted potential response to control. In a sense, the prognostic score is a very important covariate, as for causal inference, we would like to match treated and control individuals who would respond in a similar fashion if given the control. Hansen (2008a) presents the prognostic score as a useful accompaniment to the propensity score, suggesting that subclassification be done on a combination of the two scores.

1.3.2 Subsequent Analysis Steps

Modes of Inference Compatible with Matching

Assuming that a method of adjustment like stratification or matching has resulted in balance on observed covariates, the next step in the analysis is estimation of the treatment effect. For a binary outcome, a version of the Mantel-Haenszel test is often used, to estimate a confidence interval for the odds ratio conditional on the stratification. Other recommended modes of inference for categorical outcomes include randomization inference (Rosenbaum (2002c), Agresti (2002)) and conditional inference with fixed effects (Agresti 2002). Randomization-based inference methods for continuous outcomes is discussed in detail in Rosenbaum (2002a) and Rosenbaum (2002b), and conditional inference using regression methods is in there too. Agresti (2002) and Raudenbush and Bryk (2002) also using random effects for the stratification in the regression model predicting the outcome. For other modes of inference based on the ideas from Bayesian and repeated-sampling, see Rubin (1991). Rubin (1997) discusses using propensity scores in large datasets to help assess through overlap general overall comparability of the subjects in the treatment and control groups. There is also a rich set of literature in econometrics and epidemiology (discussed previously) with a focus on the properties of estimators compatible with matching.

Inferences from Sample to Population

Though this is not a topic I plan to address in detail in my dissertation, it is worth mentioning the issue of external validity of a study. Often, observational studies consist of samples of volunteers, in which case it is likely that they do not accurately represent the population to which we wish to generalize the study's conclusions (Cochran 1965). In fact, the only situation in which inference from sample to population is straightforwardly credible, is that in which a random sample was

taken. Then, even if the sample is different from the population of interest in observable ways, sampling techniques can help adjust the treatment effect estimates to reflect population information. If the sample is not taken at random, which is often the case, one way to improve the study's generalizability is to replicate the study on several other samples, each with different idiosyncratic descriptives. Otherwise, if the sample is not random and does not represent the population of interest well, a researcher might settle for only making causal conclusions in the sample.

Sensitivity Analysis

Rosenbaum (1999) makes the important distinction between overt bias - bias that is observed and measured, and hidden bias, which, although unknown, can have an impact on the study's conclusion. At this stage, he discusses the importance of sensitivity analysis of the study's conclusions to presence of hidden bias and introduces a methodology for such evaluations. Other methods of sensitivity analysis, ones with specific references to modeling treatment effects using regression-based methods, are discussed in Lin et al. (1998) and Hosman et al. (2009). Rosenbaum (2002a) emphasizes the importance of thorough sensitivity analysis for the persuasiveness and usability of conclusions from observational studies.

CHAPTER II

Utility of Balance Assessments

2.1 Introduction: Role of Covariate Balance in Causal Inference

In comparative studies, causal inference necessitates effective adjustments for important covariates. In expectation, under experimental assignment to treatment, comparability of groups in insured and the need for such adjustments is diminished. In absence of random assignment, or in experiments of a modest size, however, one expects to find some differences on observed covariates between the treatment and control groups. Depending on the discipline, there are several types of popular adjustments used to address differences on observed covariates, including covariance adjustment and matching or stratification on the propensity score. The persuasiveness of causal conclusions might then depend on a compelling demonstration that the observed covariates have been properly accounted for by these methods[1].

When discussing covariance adjustment in observational studies, Cochran (1965) reminds us that we want to protect against bias entering into the estimate of the difference between the two group means for the dependent variable by verifying that the means of the independent disturbing variables do not differ by more than sampling error for the simple case of linear regression of y on x. We can take this

[1]The treatment of differences on unobserved covariates is also important for a credible analysis, but will not be discussed in detail in this dissertation.

to mean that if there is balance on the disturbing variable x between the treatment and control groups, then omitting this variable from the linear regression is unlikely to cause a large bias in the treatment effect estimate. Absence of covariate balance in this case forces the researcher to rely on having at least a nearly correct model, which can be difficult to test.

The discussion in Cochran (1965) can be extended to say that in expectation, in a randomized experiment, no adjustments for covariates need to be made for treatment effect estimation. Freedman (2008) applies this logic to covariance adjustment and implies that if the randomization were successful (meaning that prognostically important covariates are balanced between the two groups), we would expect that the covariate-adjusted results of a randomized controlled trial should concur with unadjusted estimates. Meanwhile, Freedman also emphasizes that randomization in itself does not justify regression models and their assumptions such as independence and linearity of the response. He also points out that regression results generally are biased and may not result in a clear improvement in precision of treatment effect estimation. Freedman's arguments that randomization does not itself provide ground for covariance adjustment and that in a truly randomized experiment such adjustment should not be necessary, suggest that balancing covariates can be advantageous for inference over covariate adjustment even in randomized experiments (Freedman 2008). In comparative studies, balance tests should help identify imbalances on covariates which can cause adjusted and unadjusted estimates of the treatment effect disagree.

Matching and stratification on the propensity score can be used to 'recover the latent experiment from an observational study (Hansen 2009), in the following sense. The strong ignorability assumption provides us with selection into treatment on ob-

served covariates, and matching closely on the propensity score allows us to compare subjects with similar probabilities of assignment to treatment. If matching or stratification properly addresses the non-random selection into treatment, then we would expect observed covariates to be balanced between the treatment and control groups. Then, assuming strong ignorability of treatment assignment, a combination of close proximity of subjects within each stratum on the propensity score with the similarity of distributions of observed covariates for two groups (conditional on the stratification) brings the observational dataset closer to one that could have resulted from a block-randomized experiment (Hansen (2009), Rosenbaum and Rubin (1983)). In this dissertation, we consider primarily the importance of balance testing in analyses that utilize matching and stratification on the propensity score to adjust for differences on observed covariates between the two groups.

2.2 Balance Test: Preliminaries

2.2.1 Objectives

Although, as will be seen in this chapter, the exact formulation of the purpose of a balance test can be debatable, it can generally be described as to determine whether the observed covariates are distributed similarly between treatment and (matched) control groups. If there exist sizable differences on observed covariates, a balance test should detect them and specify to the user that the given data structure may result in misestimation of the treatment effects. Under this framework, lack of bias in the treatment effect estimate can be used in evaluating the quality of a balance test. Following the discussion in Cochran (1965) which designates randomized experiments as a guideline for planning observational studies, one might understand "balance" in an observational study as a statement that the joint distributions of covariates are similar enough between the two groups to have resulted from random assignment to

treatment.

Testing exactly that statement about joint distributions of covariates is not straightforward in practice. Nor, as we will discuss, is it always clear that the similarity of joint distributions is precisely what a balance test should set out to evaluate. Many currently use balance tests to evaluate the quality of propensity score models and the success of the stratification techniques. These tests are easier to conduct, and we expect that a combination of specification of the propensity model with an adjustment technique that passes a balance test should result in unbiased treatment effect estimates, in context where strong ignorability of treatment assignment holds. We start by making the distinction between two errors that a balance test might be expected to detect: errors of specification and errors of aggregation.

2.2.2 Pre-Stratification Testing: Errors of Specification

In debates over conclusions from the NSW dataset, first analyzed by LaLonde (1986), Smith and Todd (2005b) and Dehejia and Wahba (1999) focus on balance tests as techniques for evaluation of the specification of the propensity score. Indeed, if the propensity score has been specified correctly, then properly adjusting for it should balance observed covariates, whether it is to be used for weighting, stratification, or in another form. As will be discussed later, this thinking involves a particular definition for the null hypothesis of balance.

Evaluating the Propensity Model

The purpose of model fit diagnostics is to assess how well the given model fits the data. In contrast, a balance test is a test of whether the propensity score is an adequate balancing score (Lee 2008). When assessing the quality of the propensity specification, modeling random processes is of less interest than is a comparison

of two groups on their baseline characteristics. The specification of the propensity model would then be assessed by balance tests rather than by model fit diagnostics. This approach is taken in Dehejia and Wahba (1999). The authors take "failure" of their balance test to mean that the propensity model is not specified correctly. The details of the test will be discussed in a later section.

2.2.3 Post-Stratification Testing: Errors of Aggregation

Another way to think of balance assessment is as evaluation of a particular stratification, often a match or subclassification on the propensity score. If the observational study in question actually conceals a block-randomized experiment, a balance test can serve as an assessment of whether stratification on the propensity score (or perhaps other features of the dataset) resulted in balance of observed covariates so that the underlying block-randomized experiment is "recovered". In contrast, Imai et al. (2008), formulate the purpose of matching and stratification as reducing imbalance on observed covariates. Counterintuitively, the authors argue against testing whether such a purpose has indeed been accomplished, despite the result of such a test being of utmost importance to a researcher before evaluating an effect of treatment. If matching or stratification was successful in creating blocks of comparable subjects, and the stratified observational study now resembles a block-randomized experiment, at least on observed covariates, then, under strong ignorability, a simple difference in means on the response is an acceptable estimator for the average treatment effect. If even after accounting for subclasses, observed covariates are still distributed differently between the two groups, then, it is possible that the stratification is not fine enough, and the researcher has aggregated subjects into subclasses which could not have plausibly resulted from random assignment to treatment within those subclasses. In this case, a combination of a different stratification and post-matching

adjustments might necessary before advancing to treatment effect estimation.

Evaluating the Stratification

One goal of propensity score matching and stratification is to make the distributions of observed covariates similar between the treated and control groups, as similar as one might expect for data which came from a block-randomized experiment. If adjusting for observed covariates is sufficient to reduce to negligible levels the bias due to lack of randomization, a successful matching or stratification should enable treatment effect evaluations similar to the experimental benchmark. Propensity score analyses model treatment assignment as a random variable which is a function of known covariates. Then, subjects are matched as closely as possible on the estimated propensity score, and the analysis is done under the assumption that matched subjects' true propensity scores are equal. A balance test can be regarded as a test of the goodness of fit of this model to the data (Hansen 2008b). A finding of statistical or substantive imbalance suggests that the stratification should be redone, either by revising the specification of a propensity score on which it is based, by more stratifying in such a way as to match more closely on the propensity score itself, or perhaps both. Balance tests can help with either choosing the specification of the propensity score (Dehejia and Wahba 2002) or they can suggest that an alternate stratification method should be used. The final purpose of a balance test is the assessment of whether the pre-existing differences on covariates have been decreased or eliminated using the adjustment of choice to the point where the differences could have resulted by chance from a randomized experiment of a finite sample size. It is possible that a particular dataset might remain unbalanced regardless of the choice of propensity score specification or stratification method. This would indicate that methods other than, or in addition to, propensity score stratification should be used

to estimate treatment effects.

2.2.4 Null Hypotheses Tested

If balance tests are to be statistical tests in the ordinary sense, they must evaluate null hypotheses. Based on existing literature, there are several plausible null hypothesis for a test of covariate balance. This explains, in part, why different researchers have chosen various tests and why the results of these tests may not be consistent with each other. As mentioned previously, some balance tests can be conducted before any adjustments are attempted. These are designed to check for the necessity of such adjustments and also to test the specification of the propensity score. Lee (2008) defines balance as

$$P(X \mid Z = 1) = P(X \mid Z = 0). \tag{2.1}$$

This definition of balance does not take stratification into account (perhaps it is just a pre-stratification check) and assumes that $\mathbf{X}$ is random and treatment assignment is not random, or is random but is part of the conditioning statement. This is one type of a null hypothesis that a balance appraisal can set out to test - the joint distribution of covariates is similar between the treated and control groups. An extension of this hypothesis is to include strata S (modeled as random), and model treatment assignment as random:

$$P(X \mid Z = 1, S) = P(X \mid Z = 0, S). \tag{2.2}$$

Another way to think of pre-stratification balance is that covariates should have no predictive power for treatment assignment. More formally this null hypothesis of balance has the following form:

$$P(Z \mid X) = P(Z). \tag{2.3}$$

This test taking stratification into account would have as its null hypothesis the statement that given the stratification, the covariates should not carry additional information for treatment assignment. Others are used post-stratification to assess the effectiveness of an adjustment on the balance on observed covariates before evaluating treatment effects. This null hypothesis of balance would read:

$$P(Z \mid X, S) = P(Z \mid S). \tag{2.4}$$

Some researchers also use balance assessments to test if the propensity score is specified correctly. The hypotheses they are testing are:

$$H_0 : P(Z = 1 \mid X = x_i) = \hat{e_i} = f_0(x), \forall i$$

where $\hat{e_i}$ is the propensity score for person i, and f_0 might be defined as

$$f_0 = \frac{e^{X\beta}}{1 + e^{X\beta}} \text{ for some } \beta$$

for testing based on logistic regression, but could also be defined otherwise if a different model for treatment assignment is preferred. A similar null hypothesis might be:

$$(X \perp\!\!\!\perp Z \mid \hat{e}) \tag{2.5}$$

In equation 2.2, the hypothesis tested is that covariate distributions are similar between the treated and control groups once stratification is taken into account. In equation 2.4, the hypothesis of balance is interpreted to imply that the covariates X should have no predictive power for treatment assignment after an adjustment for stratification.

2.2.5 Utility of Balance Tests

The utility of assessing balance on observed covariates and also the necessity for a reliable metric to do so, has emerged in many recent causal inference papers. In

his 2008 paper, Lee calls for us to distinguish between pre- and post-stratification balance assessments as well as discusses existing techniques and their properties. As explained above, the pre-stratification balance check can be used to assess the specification of the propensity score and to evaluate the necessity of a stratification adjustment. A post-stratification check can be used to assess similarity between the stratified observational study and a block-randomized experiment. In the debate between Smith and Todd and Dehejia and Wahba, both parties point out the importance of assessing the specification of propensity scores through a pre-stratification balance check. Smith and Todd add that existing procedures give inconsistent results and are lacking formal investigation. Sekhon (2007) points out the importance of measuring success of the matching procedure before estimation. Most recently, Hansen and Bowers (2008) formalize a need to evaluate how close the distributions of pre-treatment variables are between the treatment and control groups in both comparative experiments and observational studies. They go on to introduce a randomization-based technique for such an evaluation. Balance tests have a role in experiments as well; particularly in small experiments, randomization may still give way to some imbalance.

2.3 Existing Methodology

Many researchers rely on procedures such as t-tests (Smith and Todd 2005b), (Dehejia and Wahba 1999) and F-tests (Rosenbaum 2002a) for significance-based balance assessments. Some use the standardized bias approach suggested in Rosenbaum and Rubin (1984). Other, less widely used ways to assess balance include a regression-based approach from Smith and Todd (2005b) and a Kolmogorov-Smirnov test approach discussed in Sekhon (2007). (It should be mentioned that the author

no longer recommends the Kolmogorov-Smirnov test and is investigating another method). Finally, a randomization-based test was recently suggested in Hansen and Bowers (2008).

2.3.1 Review of Techniques

Standardized Bias

Rosenbaum and Rubin (1985) define the standardized bias on a covariate as the mean difference as a percentage of the average standard deviation:

$$\frac{100\,(\bar{x}_t - \bar{x}_c)}{\sqrt{\frac{s_t^2 + s_c^2}{2}}}, \tag{2.6}$$

where $\bar{x}_t$ and $\bar{x}_c$ are the sample means for a particular covariate in the treated and control groups, respectively, and s_t^2 and s_c^2 are sample variances. This balance assessment is used relatively frequently in the literature; some examples include Hansen (2004) and Hill et al. (2004). A standardized difference is usually computed for each covariate included in the matching. Smith and Todd (2005) suggest also other moments of these variables as well as interactions between them, which results in higher power. A problem with using this statistic as a measure of balance is that there is no formal criterion of when it is too big; though it is suggested that 20 is to be considered large (Rosenbaum and Rubin 1985). The standardized difference approach also has the advantage of a nice graphical representation. An example of usage of the method from Hansen (2004) is a table reporting standardized differences on important observed covariates and a graphical representation, follows in Figure 2.3.1. The right panel of Figure 2.3.1 depicts pre- and post- stratification standardized differences on covariates important to the outcome in Hansen (2004). The empty dots represent the unadjusted standardized difference, and the black dots represent standardized differences on the covariates after full matching on the propensity score has been

Table 1. Selected Pretreatment Variables

Variable	Range of values	Standardized bias	Percentage of sample
Math section of PSAT	20–43	−.1	18
	45–51	.1	17
	52–57	−.1	16
	58–80	.1	15
	Not taken	.1	34
Mean SAT at respondent's first-choice college	787–987	−.3	16
	988–1,060	−.2	16
	1,061–1,123	.1	16
	1,124–1,336	.3	16
	No response	.0	36
Father's education	High school	−.4	40
	A.A. or B.A.	−.1	26
	Graduate	.4	25
	No response	.2	9
Average math grade	"Excellent"	.1	35
	"Good"–"fair"	−.1	59
	No response	.1	6
Foreign language years taken	0–2	−.3	64
	3–4	.3	27
	No response	.1	9

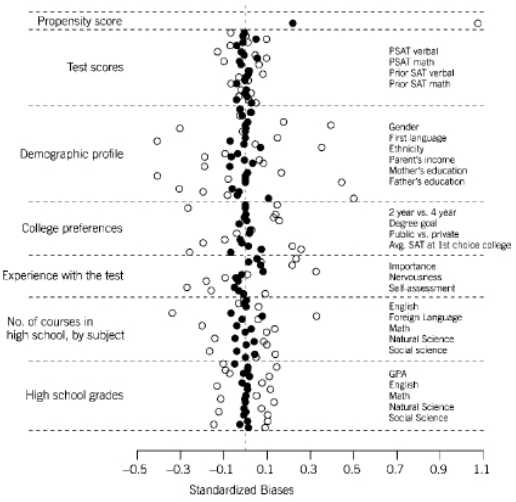


Figure 3. Standardized Biases Without Stratification or Matching, Open Circles, and Under the Optimal [.5, 2] Full Match, Shaded Circles.

performed. It is easily seen from the picture that post-stratification standardized differences (filled) are much closer to zero than pre-stratification differences (empty). This would indicate that the propensity score was likely specified correctly and full matching has been able to adjust for much of the overt bias due to these covariates.

T-test

In the case of pair matching, it is common to use the paired t-test as a way to assess balance. The null hypothesis being tested for each covariate is that the population mean difference on that covariate μ_D is equal to zero. The test statistic in this case is

$$T = \frac{\bar{x}}{\frac{s_d}{\sqrt{n}}}$$

$$t \sim t(n-1)$$

In case of stratification or full matching, however, sets of different sizes are usually produced. In this case, an independent samples t-test is frequently used to appraise balance. The null hypothesis for each covariate tested is that there is no difference in population means on that covariate between the treated and comparison groups:

$H_0 : \mu_1 - \mu_2 = 0$. The test statistic for independent samples t-test is:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$T \sim t\left(\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}\right)$$

Hotelling's T^2 Test

The conceptual advantage of the Hotelling T^2 test is that it tests equality of means on all covariates simultaneously. Still, the concern remains, that the assumption is on similar covariate distributions, while the test concerns the means only. The test statistic for Hotelling's two-sample T^2 test for p covariates is

$$t^2 = \frac{n_t n_c}{n_t + n_c}(\bar{x}_t - \bar{x}_c)'\mathbf{W}^{-1}(\bar{x}_t - \bar{x}_c) \sim T^2\left(p, n_t + n_c - 2\right) \tag{2.7}$$

where

$$\mathbf{W} = \frac{\sum_{i=1}^{n_t}\left(x_{it} - \bar{x_t}\right)\left(x_{it} - \bar{x_t}\right)' + \sum_{i=1}^{n_c}\left(x_{ic} - \bar{x_c}\right)\left(x_{it} - \bar{x_t}\right)'}{n_t + n_c - 2} \tag{2.8}$$

and

$$\frac{n_t + n_c - p - 1}{\left(n_t + n_c - 2\right)p} t^2 \sim F\left(p, n_t + n_c - 1 - p\right) \tag{2.9}$$

Logistic Regression Test

Rosenbaum and Rubin (1983) show in their Theorem 1 that under the assumptions listed in their paper, for a balancing score $b(X)$ as defined in the previous chapter,

$$X \perp\!\!\!\perp Z \mid b(x)$$

which implies

$$E(Z \mid X, e(X)) = E(Z \mid e(X))$$

Accordingly, if an observational study is balanced on observed covariates to begin with, those covariates should not be predictive of treatment assignment (Imai 2005). Extending this logic, once a researcher has conditioned on the propensity score $e(X)$, adjustment for the covariate X should not provide any additional information about treatment assignment. Imai (2005) suggests the following pre-stratification logistic regression test. First, using logistic regression, the treatment assignment is regressed on covariates and a constant, then on a constant alone. Then, he conducts a deviance test to check if all the covariates jointly predict treatment assignment. This pre-stratification logistic regression test has the advantage of being very intuitive and giving one p-value for balance on all covariates instead of one p-value per covariate. The analog of this test for post-stratification assessment would model treatment assignment as a function of covariates and subclassification, then on subclassification alone and see which model is preferred. In both of these approaches, we would conclude balance if there is no extra information to be gained from the covariates about the treatment assignment.

Dehejia and Wahba Method

Dehejia and Wahba (2002) use an adaptation of the t-test in their analysis of data originally presented by LaLonde (1986). They stratify the dataset, with strata chosen so that the mean estimated propensity scores within each stratum are not statistically significantly different for the treated and control groups. Once the strata are constructed, the authors conduct t-tests for difference in means within each stratum for each covariate. If covariates important to the authors are found unbalanced between the treated and matched control observations within a stratum, the propensity score model is modified to include higher order and interaction terms. The authors also discuss increasing the number of strata to increase the proximity of the subjects

within each stratum on the estimated propensity score. The process is iterated until all important covariates are balanced.

Smith and Todd Method

A test from Smith and Todd (2005) is based on ordinary least squares regression. They suggest estimating the following model for each variable X_k included in the matching:

$$X_k = \beta_0 + \beta_1\hat{P}(X) + \beta_2\hat{P}(X)^2 + \beta_3\hat{P}(X)^3$$
$$+\beta_4\hat{P}(X)^4 + \beta_5 Z + \beta_6 Z\hat{P}(X) + \beta_7 Z\hat{P}(X)^2$$
$$+\beta_8 Z\hat{P}(X)^3 + \beta_9 Z\hat{P}(X)^4 + \eta$$

where Z is the treatment indicator and $\hat{P}(X)$ is the estimated propensity score based on the entire covariate X. The null hypothesis being tested is $H_0 : \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9$, and assuming it to be true, treatment assignment Z should not provide any information about the variable X_k given the quartic in the propensity score. Polynomials of degree other than four might also have been considered, but the authors employ this particular order in the paper. Statistical properties of this test have not been investigated.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) test has the conceptual advantage that it actually is a goodness of fit test for distributions, as opposed to a test about population means. For a given cumulative distribution function $F(x)$ and empirical distribution function $F_n(x)$, the Kolmogorov-Smirnov statistic is:

$$D_n = sup \mid F_n(x) - F(x) \mid \tag{2.10}$$

Under the null hypothesis, $\sqrt{(n)}D_n$ converges to the Kolmogorov distribution.

Hansen-Bowers Method

If the sample were divided into treatment and control groups randomly, then chance variability would produce some differences between the groups in their observed characteristics. The magnitude of these differences would, however, be limited, following probability distributions that could be derived from the randomized design and baseline characteristics of subjects enrolled in the study. Hypothesizing a block-randomized study conducted with the subjects of this observational study by controlled random assignment, rather than uncontrolled if perhaps haphazard assignment, of this study's treatment conditions within each of its matched sets, Hansen and Bowers (2008) suggest using the probability distribution that would govern the hypothetical study as a yardstick for appraising balance in the actual study. Randomization inference is used to assess how close the matched sample is to a block-randomized design in terms of covariate distributions (Hansen and Bowers 2008). To calculate balance on the covariate distributions between the treatment and control groups, they use precision-weighted averages of differences on the individual covariates within matched sets (Hansen and Bowers 2008). These comparisons produce permutation tests, which are then combined into a χ^2 statistic which compares this study to a randomized experiment. Under the null model, the differences on the distribution of the covariates between the two groups can be attributed to chance, that is to say, the data is comparable to a blocked randomized experiment.

2.3.2 Statistical Properties of Balance Tests

The statistical properties of most existing balance tests, which were not evaluated until recently, leave much to be desired. Lee (2008) and Hansen and Bowers (2008) show that a t-test for differences in means results in unacceptably high rejection rates,

even when adjusted for multiple testing which occurs both due to many covariates being tested and because the tests have to be done in each stratum individually. Though the F-test offers an improvement in that it automatically accounts for the many strata being tested, in Lee's (2008) examples, the F-test still has high rejection rates. That is, the null hypothesis of balance is rejected too frequently even under true random assignment to treatment. The approach suggested by Smith and Todd (2005b) is analyzed by Lee as well and apparently shares the problems exhibited by other tests. Finally, the logistic regression test is shown to have high rejection rates in Hansen and Bowers (2008). The permutation-based test suggested in Hansen and Bowers (2008) yields more promising results, as indicated by the paper itself and as shown by Kleyman and Hansen (2008)

2.3.3 Critiques

While Lee (2008) and Hansen and Bowers (2008) direct their papers at evaluating current techniques for balance testing and their statistical properties, Imai et al. (2008) and Sekhon (2007), criticize other aspects of balance testing, namely the validity of its current usage. Although Imai et al. focus their discussion primarily on the t-test, they express several concerns that target balance testing in general. They mention the following major issues, and in his 2007 paper, Sekhon appears to agree with the last two claims:

1. The t-statistic can be decreased and made insignificant simply by randomly throwing away observations. A researcher could thus discard observations in order for the data to pass a balance test like the t-test.

2. The power of the test decreases as n decreases. This concern is similar to the one above. The power of the t-test decreases with sample size. If a t-test is

conducted within each stratum, then making the strata smaller in size could falsely indicate better balance on covariates simply due to a reduction in power of the test.

3. The t-test along with several other balance assessments assume that there exists a background population and are designed to use sample statistics to make inference about parameters of that population. In fact, we are usually interested in whether balance on covariates has been accomplished in the particular sample, and techniques assuming a background population may be ill-suited for testing such a hypothesis.

4. Balance should be maximized without limits, and balance tests should not be used as stopping rules in attempts to maximize balance.

2.4 Rejoinder

Imai et al. dedicate a section of their paper to discussing conceptual and statistical problems with balance testing. They examine the t-test for difference in means and point out its disadvantages. Though they point out some disadvantages of the t-test (detailed above), there is a bigger problem at hand. The problem is, that even though t-test is convenient for balance checking, it is in no way appropriate for assessing covariate balance. It is a test that is designed to see if a disparity between sample means gives evidence of difference in population means. This hypothesis has little in common with the null hypothesis of balance, however that is formulated. When the test is conducted within strata, its power is greatly affected. Furthermore, researchers rarely account for multiple testing which occurs when t-tests are conducted for each covariate within each stratum. Finally, Lee (2008) and Hansen and Bowers (2008) show that the t-test has poor properties for balance assessment. Further, sample size

and power issues, brought up by Imai et al., and discussed below, have little to do with balance testing alone; those issues are common to all hypothesis tests.

Reduction in Sample Size and Power

Imai et al. express dissatisfaction that the power of the t-test decreases with sample size. It should be noted, that many other hypothesis tests, ones that have nothing to do with balance, share this property. Despite this, many researchers continue to proceed with hypothesis testing. Decrease in power due to reduction in sample size is a weak argument against balance testing. Imai et al. express concern that a researcher might be tempted to randomly throw away observations in order to accomplish 'balance' by a measure with similar properties. But balance testing would not be the first area of statistics or science in general, which would rely on a researcher's integrity to not throw data away. In fact, to extend this logic, to maximize balance, then, we could throw away vast majority of the data. Also, throwing away data reduces power for treatment effect estimation, which we would expect to discourage researchers from such practices.

The Genmatch algorithm (Sekhon 2007) requires the sample to be matched to be specified ahead of time, so the issue of decreasing sample size is avoided. However, this technique still is not immune to a researcher specifying a favorable subsample. A similar situation arises in the debate between Smith and Todd and. Dehejia and Wahba. Dehejia and Wahba (2002) choose a subset of the data for their analysis which produces results confirming their hypothesis. Smith and Todd (2005a), however, claim that the results that Dehejia and Wahba produce are actually specific to that particular subsample Smith and Todd (2005b) and, in fact, do not generalize to different equally reasonable subsamples of the same dataset. Genmatch, along with other balance-testing techniques would still be vulnerable to such choices

by the researcher. However, it is clear that small subsamples of the data result in variable estimates sensitive to the particular subsetting used and analyses involving such techniques are usually appropriately discounted, as noted by Smith and Todd (2005b).

Another issue in reduction of power of balance tests with sample size has to do with the fact that some matching or stratification methods may result in excluding subjects from analysis. For example, matching within propensity calipers might leave subjects with extreme propensity scores unmatched. Or, stratification into five strata on the propensity score might result in a stratum containing only treated or only control subjects. A concern voiced by Imai et al. (2008) is that due to such sample size reductions, a researcher might conclude that balance on observed covariates has improved from the unstratified to the stratified sample, when, in fact, all that has occurred is that the power of the test to detect imbalance has decreased. An important point of discussion here is whether or not we should, at all, be interested in *improvement* in balance from the unmatched to the matched sample. One convention in hypothesis testing is that without quantifying the parameter of interest, we might simply be interested in extracting a result about it (Fisher 1956). For example, a researcher might want to know just whether balance on observed covariates is *good enough* for causal inference without bias due to these variables. According to this convention, the actual *change* in a test statistic or p-value reflecting balance is not of interest, so the fact that sample sizes are different should not matter. For researchers who disagree with this approach, there is still a workaround. Suppose a particular matching algorithm, indeed, does result in some subjects being unmatched and thus unused in treatment effect estimation. A researcher could then evaluate pre- and post-matching balance on the *matched* sample only. This way, the reduction

in sample size does not affect the power of a balance test, and conclusions about improvement in balance due to matching can be made.

A final concern with reduction of power of a balance test with sample size has to do with the fact that initially, the t-test for balance is evaluated in a large, unstratified sample. Its results are then compared with a series of t-tests within much smaller strata of the same dataset, in which the standard errors get inflated and the power of the test necessarily shrinks. Smith and Todd (2005b) mention that the standardized differences approach to assessing balance can also run into sample size issues, this time by adding observations to the sample and increasing variance. The authors also rightly criticize a t-test-based approach to testing balance implemented in Dehejia and Wahba (2002) in which the starting number of strata is not specified, and the power of the tests conducted within strata decreases as the strata get narrower. Taking into account that the overall final purpose of stratification and balance tests is evaluating the effect of treatment, and, as hypothesized in Hansen (2008b), inflation of standard errors of baseline imbalances through matching helps insure coverage for treatment effect estimates somewhat addresses the concern of Smith and Todd (2005b). Though their criticism is valid, there are balance tests which do not explicitly conduct evaluations within strata. One of those is discussed in Hansen and Bowers (2008), and we will present others in the next chapter. In each of these procedures, the particular reduction in sample size that Smith and Todd (2005b) describe is not a concern.

Absence of Superpopulation

Imai et al. (2008) point out an important flaw in most current balance-testing approaches: the absence of a superpopulation. Usually, the researcher is interested in evaluating balance on observed covariates *only* in the current sample, without

reference to the population from which the sample was extracted. In other words, balance that we evaluate is strictly a property of the sample. Hypothesis tests such as t-test, F-test, and others, presuppose an underlying population that gave rise to the current sample, and are designed with the purpose of using sample information to make an inference about a population parameter. In balance testing, however, such a population is difficult to define. Thus, even if a particular superpopulation-based test is useful for detecting imbalance on observed covariates, we can detect a problem with its underlying assumptions. Thought the complaint of Imai et al. (2008) is valid, it still does not provide evidence against significance-based testing for covariate balance overall. Rather, it is a hint towards a different approach to balance testing, one based on the sample alone. Significance testing can be done without assumptions about an underlying population using permutation-based inference and bootstrap methods. Focusing on such sample-based procedures allows the researcher to assess balance using methods that do not rely on inference about a population parameter, and helps to avoid the contradiction specified in (Imai et al. 2008).

Procedures like the t-test and F-test, although somewhat robust, still rely on the assumption that the data tested are normally distributed, at least in small samples. This assumption is rarely actually tested for each covariate, is unlikely to be true for all of them, and is certainly violated for binary variables. With regard to cases where it is desirable to assume a background population, Lehmann (1990) talks about permutation testing as a robust alternative to t-tests when the assumption of normality is violated. Then the switch from such procedures to ones based on permutation inference is not undesirable even setting aside the superpopulation issue.

Sekhon (2007) suggests using bootstrap Kolmogorov-Smirnov tests for baseline covariates, which rely only on the sample distribution instead of assuming a super-

population. Hansen and Bowers (2008) suggest a permutation-based test for balance that utilizes the standardized bias approach from Rosenbaum and Rubin (1984). In a later section, I will introduce a balance test which relies on a permutation distribution of a likelihood ratio test statistic in logistic regression.

Maximizing Balance without Limits

The notion that balance on observed covariates should be somehow evaluated, and that such evaluation might speak to the quality of an adjustment procedure, such as stratification on the propensity score, seems to be accepted by most methodologists of matching. Researchers have also addressed the value of looking at more than means for observed covariates, in particular linear combinations of covariates are important to balance as they are important features of the joint covariate distributions. Most current disagreements about balance assessment can generally be reduced to one comprehensive issue, and that is whether covariate balance should be "maximized without limits" (Imai et al. (2008) and Sekhon (2007)). This philosophy asserts that using balance tests as a stopping rule, rather than explicitly maximizing balance, leads to misleading inferences. An opposing viewpoint is that instead, balance should be assessed on the property of being good enough to decrease Type I error rates in treatment effect estimation due to observed covariates to acceptable levels (Hansen 2008b).

What might it mean to "maximize balance without limit"? Consider an observational dataset, in which the researcher has estimated the propensity score and is ready to perform a matching procedure. Suppose also that the researcher has several matching procedures in mind for this dataset, perhaps involving matching with restrictions of some kind, like matching within propensity calipers. How should she choose among them? She can start by evaluating the success of each match at

balancing observed covariates using a balance test, which gives her some aggregate indicator of balance corresponding to each procedure. This could be some measure of a maximum discrepancy, like the maximum standardized difference on the covariates, or a maximum distance on the propensity score within a matched set, or it could be a p-value. What should the next step be? Should she pick the procedure with the highest balance measurement or are there other considerations? The position that balance should be maximized without limit tales that whatever other considerations may come up must be secondary to the consideration of maximizing the aggregate measure of balance.

A contrasting position holds that the purpose of a comparative study is to evaluate the effect of treatment with as little bias and variance as possible. All components of an analysis should work towards that purpose. It is possible to have a study that is so confounded (even just on observed covariates) that even the post-stratification inducing maximum balance based on matching on a propensity score, say, still does not yield conditional distributions $P(X|S)$ that well approximate what they would be in a block randomized experiment. Maximizing balance without limit does not necessarily align with that purpose. One problem with maximizing balance without limits and not testing its statistical significance, is the lack of assurance that even the maximum accomplishable balance on observed covariates in a particular study is actually good enough for unbiased inference. A dataset could simply be too confounded, and even the maximum balance on observed covariates, might not account for selection bias on those covariates. On the other hand, it is also possible for data to produce a good estimate of the treatment effect without maximizing balance to the extreme.

Another problem to maximizing balance without limit is that such maximization

often comes at a cost. Thus, doing so does not make sense if the marginal benefit in reduction of bias in the treatment effect estimate is small compared to the cost. There is a need for a tool to assess how likely a sample imbalance is to be due to chance alone and whether a particular sample imbalance will have an effect on the Type I error rate in estimating the treatment effect. Maximizing balance on observed covariates may be tied to discarding subjects from the study. If unnecessarily maximizing balance to the extreme implies reducing sample size and increasing variance in the treatment effect estimate, then this is also not aligned with the original purpose. In the extreme case of maximizing balance without limit one might end up with very few or no subjects in the study. Reductions in sample size that can result from maximizing balance is just another example of a common trade-off in statistics - that between bias and variance. Significance-based balance tests are a useful tool for finding an equilibrium in this trade-off. An example of this trade-off can be seen in a later chapter in Table 4.10. This table is a summary of various matching procedures performed as part of the analysis of the Right Heart Catheterization dataset from (Connors et al. 1996a), which is discussed in detail and analyzed in the last chapter. For now, consider looking at the second and third rows of the table. These are the results from matching on a variation of the propensity score which is modeled using prognostic scores (Hansen 2008a). In the second row of the table, matching is done without restrictions, and all subjects are retained in the analysis. The p-value for a test of balance based on the procedure in Hansen and Bowers (2008) is zero, indicating significant imbalance. Estimating treatment effects in this setting may be affected by bias on observed covariates. In the next line of the table, a caliper of 0.2 pooled standard deviations has been imposed on the match, resulting in excluding 129 subjects from the study but also achieving a balance p-value of 0.442

and suggesting a substantively important reduction in overt bias. Now, looking at the sixth line of the table, we can see that if on top of the propensity caliper, we let the software optimally eliminate 20% of controls, the balance p-value goes up even higher to 0.71, and 713 subjects now have been excluded from analysis. A researcher could continue along the path of excluding more subjects to maximize balance, but based on p-values associated with these results, she might also consider stopping at the two stages described above and retain study subjects knowing that there is no longer strong evidence against balance.

Imai et al. (2008) argue that using balance tests as a stopping rule for matching results in biased treatment effect estimates. However, that depends on exactly which stopping rule is used. For example, a researcher does not have to stop maximizing balance after a nominal pre-decided p-value like 0.05. Indeed, there is nothing sacred about 0.05 in other hypothesis tests as well. Even with balance test p-values above 0.05, a researcher may continue to maximize balance, but she is able to figure out how many subjects are being discarded and stop when the cost is too high. In addition, balance tests and their p-values can be used in comparisons of stratification configurations. In cases where a stratification results in strata with too large or too small treatment-control ratios, a researcher might want to restructure the stratification to be closer to pairs, for example, balance usually gets worse, and this cost can easily be quantified.

CHAPTER III

Proposal for a New Balance Test

3.1 Balance Testing with Logistic Regression

3.1.1 Explanation

In his analysis of the effectiveness of a "Get Out the Vote" campaign, Imai (2005) expresses a concern that observed covariates may not be balanced between the experimental treatment and control groups in his data. He uses a balance test to try to decide whether a study that claimed to be randomized, in fact, had been. He proposes to evaluate covariate balance in the following way:

- First, he uses logistic regression to predict treatment assignment using all covariates and their first order interactions as predictors
- Then, he conducts the residual deviance test to examine whether these covariates jointly significantly predict the treatment assignment

The idea behind this test is that the ability of covariates to predict treatment assignment well is evidence against covariate balance that would result from random assignment. If treatment had been truly randomized, we would expect it to be independent from covariates and their functions. This test is consistent with the definition of balance given in 2.3. For a matched or stratified study, this approach could be extended to definition 2.4 in the following way:

- First, we can use logistic regression to predict treatment assignment using all covariates and the strata indicators
- Then, we could compare this model with one that uses only the strata indicator to predict treatment assignment, using a model selection tool.

If the smaller model is preferred, then the stratification has accomplished its purpose: in the presence of the strata indicator, observed covariates have no more predictive value for treatment assignment and can thus be considered balanced according to definition 2.4. In an observational study, this could be interpreted to mean that the block-randomization model fits the data reasonably well. A brief discussion of a related approach that uses probit regression can be found in Sianesi (2002).

3.1.2 Advantages

Aside from its simplicity, one of the main advantages to this approach to balance testing as opposed to many others, like the standardized bias approach, is the ability of the researcher to obtain a global p-value for an overall test of balance, while still retaining access to individual p-values for each of the covariates of interest. Also important is the opportunity to easily include in the model interactions, squared terms or other functions of covariates and their linear combinations that might be of interest or concern. There is, however, a limitation to this that has to do with sample size, which will be discussed below. Another benefit of evaluating balance using logistic regression is that in case of a lack of balance, including interactions with the matched set indicators can facilitate checking which covariate and matched set combinations are a threat to balance. Finally, in research that concerns heterogeneous treatment effects or for another situation in which balance on certain covariates is quantifiably more important than on others, the researcher could use weights in the

regression equations to account for her needs.

3.1.3 Disadvantages

Hansen and Bowers (2008) point out that even though logistic regression appears to be well-suited to test for covariate balance in the manner described above, in fact, stringent sample size requirements for logistic regression present a problem big enough to render the method for testing balance unsuitable in small samples. In particular, the logistic regression approach to balance testing has very high Type I error rates (rejecting the null hypothesis of covariate balance in a data setting where treatment assignment was randomized), in the authors' examples reaching more than three times the nominal level in some cases. It is likely that such high rejection rates are the result of overcrowding the logistic regression model, which is remarkably easy to do. Harrell (2001) and Peduzzi et al. (1996), for example, suggest that in order to not overfit the logistic model at least 10 observations per confounder are necessary. To address overfitting, we propose modifications of the logistic regression model using penalized likelihood and Bayesian modeling techniques.

3.2 A Small-Sample Approach to Logistic Regression

3.2.1 Bias Reduced Logistic Regression

For simplicity, consider a collection of p independent variables denoted by $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$, a vector of parameters of interest β, and a binary response variable Y (a more complex approach might include squared and interaction terms). Logistic regression is used to model the conditional probability of the outcome being present given the vector of covariates: $P(Y = 1 \mid \mathbf{x}) = \pi(x)$. The logistic regression model is defined as

$$\text{Prob}(y_i = 1|x_i, \beta) = \pi_i = \left[1 + \exp\left(-\sum_{r=1}^{p} x_{ir}\beta_r\right)\right]^{-1}$$

Its likelihood function, which is maximized in order to estimated the parameter vector β is calculated as follows:

$$\begin{aligned} L(\beta, \sigma^2 \mid Y, X) &= \prod_{i=1}^{n} \left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \right)^{y_i} \left(1 - \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \right)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \right)^{y_i} \left(\frac{1}{1+e^{x_i'\beta}} \right)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(e^{x_i'\beta} \right)^{y_i} \left(\frac{1}{1+e^{x_i'\beta}} \right) \\ &= \prod_{i=1}^{n} \frac{\left(e^{x_i'\beta} \right)^{y_i}}{1+e^{x_i'\beta}} \end{aligned}$$

Taking the above L as the likelihood function and β_r, $(r = 1, ..., p)$ to be regression parameters, the maximum likelihood estimates of β_r are the solutions to the score equations $U(\beta_r) = 0$ where $U(\beta_r) = \frac{\partial \log L}{\partial \beta_r}$ The log-likelihood is:

$$\text{logL} = \sum_{i=1}^{n} y_i x_i \beta - \log(1 + e^{x_i \beta}),$$

so the score equation for each parameter of interest β_r comes out to be:

$$U(\beta_r) = \sum_{i=1}^{n} \left[y_i x_{ir} - x_{ir} \frac{e_{ir}^{x} \beta_r}{1 + e_{ir}^{x} \beta_r} \right] = \sum_{i=1}^{n} x_{ir}(y_i - \pi_i) = 0$$

The Jeffreys prior is a non-informative prior distribution proportional to the square root of the Fisher information and is invariant under reparameterization of the model (Jeffreys 1946):

$$p(\theta) \propto \sqrt{I(\theta|y)}$$

One could use this prior in ordinary Bayesian algorithms for calculation and sampling from the posterior distribution. However, it can also be used as a penalty to maximum likelihood. It has been shown that by modifying the score function using Jeffreys invariant prior, one can remove the first-order term from the asymptotic bias

of the maximum likelihood estimates (Firth 1993). In order to reduce the bias in these estimates that occurs in small samples, Firth(1993) suggested using the Jeffreys prior to penalize the likelihood in the following way:

$$L(\beta)^* = L(\beta)|I(\beta)|^{\frac{1}{2}},$$

where $|I(\beta)|^{\frac{1}{2}}$ is precisely the Jeffreys invariant prior for this problem (Heinze and Schemper 2002). This modifies the log likelihood formulation as follows:

$$\log L(\beta)^* = \log L(\beta) + 1/2 \log|I(\beta)|$$

Firth (1993) showed that the $O(n^{-1})$ bias of the maximum likelihood estimates $\hat{\beta}$ is removed by using this penalty. For the logistic model:

$$P(y_i = 1|x_i, \beta) = \pi_i = \left[1 + \exp\left(-\sum_{r=1}^{p} x_{ir}\beta_r\right)\right]^{-1}$$

the original score equation is:

$$U(\beta_r) = \sum_{i=1}^{n}(y_i - \pi_i)x_{ir} = 0 \quad (r = 1, ..., p)$$

and the modified score equation is:

$$\begin{aligned} U(\beta_r)^* &= U(\beta_r) + 1/2\text{trace}[I(\beta)^{-1}\{\partial I(\beta)/\partial\beta_r\}] = 0 \\ &= \sum_{i=1}^{n}\left[y_i - \pi_i + h_i(1/2 - \pi_i)\right]x_{ir} = 0, \quad (r = 1, ..., p) \end{aligned}$$

where h_i are the diagonal elements of the 'hat' matrix $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$ with $W = \text{diag}\{\pi_i(1 - \pi_i)\}$(Heinze and Schemper 2002). The estimates $\hat{\beta}$ are obtained iteratively until convergence at iteration k by:

$$\beta^{(k+1)} = \beta^{(k)} + I^{-1}(\beta^{(k)})U(\beta^{(k)})^*.$$

Using Jeffreys prior to penalize the maximum likelihood function in small samples is not very expensive computationally, making it a common technique for logistic

and other types of generalized linear models. Its performance for both the logistic regression of an outcome on treatment and covariates and for the estimation of the propensity score will be evaluated in the simulation study discussed in the next chapter.

3.2.2 A Fully Bayesian Approach

In logistic regression, separation occurs when a predictor or a linear combination of predictors perfectly predicts the response. In this case, the likelihood converges, but at least one parameter estimate diverges to plus or minus infinity. Separation is often arises when logistic regression is used in small samples. For the problem of separation in logistic regression, Gelman et al. (2008) suggest using Bayesian inference. They propose an adaptation of the iteratively weighted least squares algorithm to estimate logistic regression coefficients using independent t prior distributions. This approach includes more prior information than the one discussed above and results in more smoothing than bias-reduced logistic regression, and, according to Gelman et al. (2008), yields more stable estimates when applied to sparse data. A key idea behind the choice for the prior is that the effects of an input into the regression model are unlikely to fall beyond a certain range. In particular, the authors use the example that a change of 5.0 on the logistic regression scale moves a probability from 0.01 to 0.5 or from 0.5 to 0.99. Because the switch from 0.01 to 0.99, say, due to an input is unlikely, the authors pick a prior that assigns low probabilities to changes of 10.0 on the logistic scale. The approach consists of the following two steps:

Step 1: Standardization

Each input variable has to be standardized to be in accord with the scale of the chosen prior distribution.

- Binary input variables are recomputed to have a mean of 0 and to differ by 1 between their lower and upper condition
- Other input variables are shifted to have a mean of 0 and standard deviation of 0.5.

Step 2: A Weakly informative t family of prior distributions

Gelman et al. (2008) suggest using the Student-t family of distributions with mean zero, degrees of freedom being equal to one and a scale parameter of 2.5. This also corresponds to a Cauchy distribution with a scale parameter of 2.5. These independent priors are assigned to each of the coefficients in the logistic regression, except for the constant term, which gets assigned a Cauchy distribution with center zero and scale 10.

Computation

Ordinarily, a logistic regression algorithm might utilize Iteratively Reweighted Least Squares (IRLS) to estimate the coefficient vector β in the logistic regression model $Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta)$. Each iteration of this algorithm involves computing an adjusted response vector that is based on the linear prediction using current parameter estimates ($X\beta$), observed values of the response y, predicted probabilities of the response being present, $\text{logit}^{-1}(X_i\beta)$, and the derivative of the link function, which happens to have a connection to the predicted variance at each point. The adjusted response vector is calculated as:

$$y_i^* = X_i\hat{\beta} + (\sigma_i^z)^{-2}\left(y_i - \frac{e^{X_i\hat{\beta}}}{1 + e^{X_i\hat{\beta}}}\right),$$

and the variances used for inverse weighting are:

$$(\sigma_i^z)^2 = \frac{e^{X_i\hat{\beta}}}{(1 + e^{X_i\hat{\beta}})^2}.$$

An ordinary logistic regression algorithm might regress z on X using weights $(\sigma_i^{y*})^{-2}$ and use the resulting estimates of β in the calculation of the new adjusted response vector z. The process can then iterate until convergence.

Under a normal prior distribution for the model's coefficients, estimation is straightforward. However, as mentioned in Step 2 above, Gelman et al. (2008) suggest a more flexible t-distribution for robust inference. In order to incorporate this information in the estimation process, the authors suggest placing the iteratively weighted least squares described above within an approximate EM algorithm to estimate the coefficients (Dempster et al. 1977).

For the approximate EM algorithm with a t prior distribution, the authors recommend using the formulation

$$\beta_j \sim N(\mu_j, \sigma_j^2),\ \sigma_j^2 \sim \text{Inv-}\chi^2(\nu_j, s_j r),$$

to express the t prior distribution for each β as a mixture of normal distributions with unknown scale σ_j. It results in the following augmented dataset for weighted linear regression:

$$z_* = \begin{pmatrix} z \\ \mu \end{pmatrix},\ X_* = \begin{pmatrix} X \\ I_J \end{pmatrix},\ w_* = (\sigma^z, \sigma)^{-2} \tag{3.1}$$

with z and σ^z vectors of z_i and σ_i^z as defined above, X, the design matrix, and μ and σ vectors of μ_j and σ_j. Gelman et al. (2008) recommend setting $\mu_j = 0\ \forall j$ as the center of the Cauchy distribution, and also setting $\nu_j = 1$ and $s_j = 2.5$. The resulting approximate posterior density is:

$$\begin{aligned} \log p(\beta, \sigma|y) \approx & -\frac{1}{2}\sum_{i=1}^{n} \frac{1}{(\sigma_i^z)^2}(z_i - X_i\beta)^2 - \frac{1}{2}\sum_{j=1}^{J}\left(\frac{1}{\sigma_j^2}\beta_j^2 + \log(\sigma_j^2)\right) \\ & - p(\sigma_j) + \text{constant} \end{aligned} \tag{3.2}$$

Algorithm

First, set each σ_j to s_j and guess a plausible value for β. Then, proceed:

1. Using the current estimate of β, determine the vectors z and σ^z. Then perform weighted least squares to get the new estimates for β and the variance matrix V_β.

2. Use the EM algorithm to estimate the unknown scale σ_j. In the E-step, use the new $\hat{\beta}$ to find the approximate expected value of the log posterior:

$$E\left((\beta_j - \mu_j)^2 \,|\sigma, y\right) \approx \left(\hat{\beta}_j - \mu_j\right)^2 + (V_\beta)_{jj}$$

3. M-step: maximize the above expected value of the log-posterior density and get the estimate

$$\sigma_j^2 = \frac{(\hat{\beta}_j - \mu)^2 + (V_\beta)_{jj} + \nu_j s_j^2}{1 + \nu_j}$$

4. Reset the augmented dataset using the new $\hat{\beta}$ and σ_j.

At convergence, report β and V_β.

3.3 Model Selection

For penalized and Bayesian approaches to regression modeling, common model selection criteria may not apply. For example, in both of the above approaches, the likelihood of the full model does not have to be bigger than the likelihood of the smaller model, so the usual likelihood ratio test no longer is expected to follow a χ^2 distribution. Further, as was discussed in Chapter 2, such a test suggests that we are interested in making inference about a superpopulation, which, in case of balance testing, is difficult to define. For these reasons, we opt for model selection using permutation testing. We will discuss three different models for balance testing below; each of them uses a permutation approach to model selection.

3.4 Suggested Tests

In each of the following sections, we discuss a test that compares the following models:

$$Pr(Z = 1) = \text{logit}^{-1}(\gamma_1 s_1 + \cdots + \gamma_k s_k)$$
$$Pr(Z = 1) = \text{logit}^{-1}(\beta_1 x_1 + \cdots + \beta_p x_p + \gamma_1 s_1 + \cdots + \gamma_k s_k)$$

First, the models are fit using either ordinary logistic regression, bias-reduced logistic regression, or using a Cauchy prior, as discussed above. Then a test statistic to choose between the two models is computed and recorded. In the next step of the algorithm, within strata or matched sets, treatment assignment is randomly permuted, and both models are refit and the test statistic is recalculated and recorded. This step is repeated to obtain a permutation distribution of the test statistic. The p-value for balance is computed by comparing the observed test statistic to its permutation distribution. One can look at this as comparing the test statistic from our study to that of many possible block-randomized experiments that could have taken place with this data and stratification structure. We suggest the following three model-fitting procedures and related test-statistics:

Generalized Linear Model Method

- Model: generalized linear model with a logit link
- Test Statistic: Likelihood Ratio Test statistic between the two models.
- Expectations: We expect that even though we will use permutation rather than a χ^2 distribution for the test statistic, this method may not perform well in small samples. In large samples, however, we expect that this method will perform well. The advantage of this method over the others is the ease and speed of the

computing process.

Bias-Reduced Logistic Regression Method

- Model: Penalized Likelihood using Jeffreys prior, as described above.
- Test Statistic: Difference in the AIC [1] between the two models.
- Expectations: We expect this method to perform better that the GLM-based method in small samples. Because this approach only provides limited smoothing, it may not prove to be enough to correct for separation in logistic regression due to overcrowding. A limitation of this method is computing speed - it is the slowest of the three.

3.4.1 Bayesian Modeling Method with a Cauchy Prior

- Model: Bayesian with a Cauchy prior for each coefficient, with center zero and scale 2.5, with the exception of the intercept, which has scale 10.
- Test Statistic: Likelihood Ratio Test statistic between the two models.
- Expectations: We expect this method to perform better than either of the two above methods in small samples. Because this approach provides the most smoothing, we expect it to help to correct for overfitting of the logistic regression model. Computing speed in large samples is a limitation of this approach, but it is our expectation that in large samples the results of this method may be comparable with those of the GLM-based method, rendering this approach only necessary in small samples.

[1] The Akaike's Information Criterion (AIC) is defined as

$$\text{AIC} = 2k - 2ln(L)$$

where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model. It was proposed by Akaike (1974), as a tool for statistical model selection describing the tradeoff between the model's complexity and precision. Given a data set, and several competing models, the model having the lowest AIC is usually considered the best.

CHAPTER IV

Case Studies

This chapter contains two case studies which use balance testing as an important diagnostic analysis prior to estimating treatment effects. We start with an application to a small quasi-experimental dataset, and conclude with a large observational study. In both case studies, we illustrate the important role of balance tests introduced in Chapter 3 in determining the analysis strategy.

4.1 Comparison of Faith-Based and Biopsychosocial Treatments for Substance Abuse

4.1.1 Dataset

The Puerto Rico (PR) dataset is from a prospective study of men in faith-based (treatment) and biopsychosocial (control) substance abuse treatment programs in Southern Puerto Rico. These men were court mandated to treatment and asked to report recent use of cocaine and/or opiates (Hansen et al. 2004). The causal question of interest is whether the faith-based treatment is more effective than the biopsychosocial treatment on the binary outcome measure of whether the subject was retained in the program after three months. Because subjects had little discretion over the choice of program, selection into treatment is likely to be weak making this data quasi-experimental, in the sense that assignment to treatment was neither random-

ized nor in the hands of the subjects. The number of covariates in this dataset rivals the sample size; this feature can severely undermine regression adjustment. Our methods, however, allow us to utilize this rich, well-validated psychometric data for correction of pretreatment differences. At the beginning of treatment, interviewers administered three instruments: the Religious Background and Behavior Questionnaire (RBBQ), the University of Rhode Island Change Assessment Scale (URICA), and the Addiction Severity Index (ASI). Subjects were asked additional questions about demographics and treatment histories, and administered carbon monoxide breathalyzer tests for cigarette smoke (Hansen et al. 2004). The outcome of interest is binary and measures if the subject was retained in treatment after three months or if the subject left the program. After tackling some missing data issues, we use principal components analysis and Bayesian modeling to reduce the dimension of the covariate matrix by summarizing it with propensity and prognostic scores. We then match treatment and control subjects on the above two scores and other important features of the dataset, using optimal full matching. We evaluate post-matching balance on observed covariates and the resulting setting is closer, we maintain, to the classical randomized experiment framework and proceed to estimate the treatment effect.

4.1.2 Dimension Reduction

Treated and control groups differ with respect to their pre-treatment characteristics. In order to determine the discrepancy in attrition between the experimental and control groups, we need to compare them in an equitable manner by homogenizing joint covariate distributions. Though Rosenbaum and Rubin (1985) and Rubin and Thomas (2000) suggest match on a Mahalanobis distance based on covariates within propensity calipers, we are cautious to follow their advice literally because

matching on a Mahalanobis distance works best when the variables used to construct the metric are normally distributed, which is decidedly not the case in this dataset, which can be seen in Figure 5.1. We are further restricted by the small sample size in the study, which introduces the necessity to distinguish those covariates that matter most to the response from those that matter less.

Because it is undesirable in our data to form a Mahalanobis distance from all of the covariates, we instead reduce their dimension and prioritize among them before forming our distance matrix. We start by reducing the dimension of the data by principal components. Then we compute several data-based scores which reduce the dimensionality and summarize the covariates, thereby resulting in just a few measures easily combined into a Mahalanobis distance to help discriminate between the matches. Although the propensity score is perhaps the most important contributor to bias reduction, the other scores will help focus the comparison on lines of especial interest.

We arrange to compare subsets of the treatment and control group that are more similar in their joint covariate distributions by the use of optimal matching (Rosenbaum 1991), propensity scores (Rosenbaum and Rubin 1985), and related techniques (Hansen 2008a).

Preliminary Dimension Reduction

After tackling some missing data issues (the procedures are described in Kleyman and Hansen (2008)), we start with a preliminary measure to address the problem of large covariate-to-sample size ratio (53 covariates and 33 and 34 subjects in the treatment and control groups, respectively). Principal components analysis is a widely-used technique for reducing the dimension of the covariate matrix so as to parsimoniously capture the most variability (Harrell 2001). To decrease the number

of covariates going into the models we fit, we use principal components to summarize variables describing prior drug treatment and incarceration. These steps reduced the number of pre-treatment variables from 53 to 27. In order to match subjects from the treatment and control groups, we needed to further reduce the dimension of the covariates down to a more manageable number.

Propensity Score

Instead of using the covariates directly in estimation of treatment effects, we utilize them in the computation of the unidimensional propensity score, which was introduced earlier in Section 1.2.7. The propensity score is usually estimated by logistic regression of the treatment indicator on observed covariates relevant to treatment assignment. To estimate the propensity score, we model the treatment indicator as a function of the demographics, principal components of treatment histories, summary RBBQ and URICA variables, ASI variables, and all the relevant missing dummy variables. However, using ordinary logistic regression here can be expected to result in overfitting, since the number of covariates predicting treatment assignment is large compared to the sample size. When the data are modeled using ordinary logistic regression, the two groups differ on the estimated propensity score by 1.64 pooled standard deviations. Given that there were 27 covariates and only 67 observations, such a difference might be expected to occur even if the true propensities were similar for treated and control subjects and vice versa, because of errors of estimation. To address this apparent overfitting, we try fitting the propensity model using methods based on Jeffrey's prior from Firth (1993). This penalized maximum likelihood procedure (`brglm` function in `R`) overfits less, resulting in a smaller standardized difference. Still, this procedure results in several individuals having estimated propensities approximately equal to 0 and 1. As a final attempt to re-

duce overfitting, we fit the propensity model using a Bayesian procedure which puts Cauchy priors on the regression coefficients (Gelman et al. 2008). In R, this can be done using the function `bayesglm`. This model results in some overlap on the estimated propensity score and no propensity values extremely close to 0 or 1. The overlap on the estimated propensity scores as well as overlapping plots of some of the covariates that went into their estimation, can be seen in the Figure 4.1. The fact that individual covariates have overlapping distributions somewhat substantiates our position that our lack of overlap on the estimated propensity score is due to overfitting of the logistic model rather than lack of common support on the true propensity score. Nevertheless, we sought to further improve the quality of our match

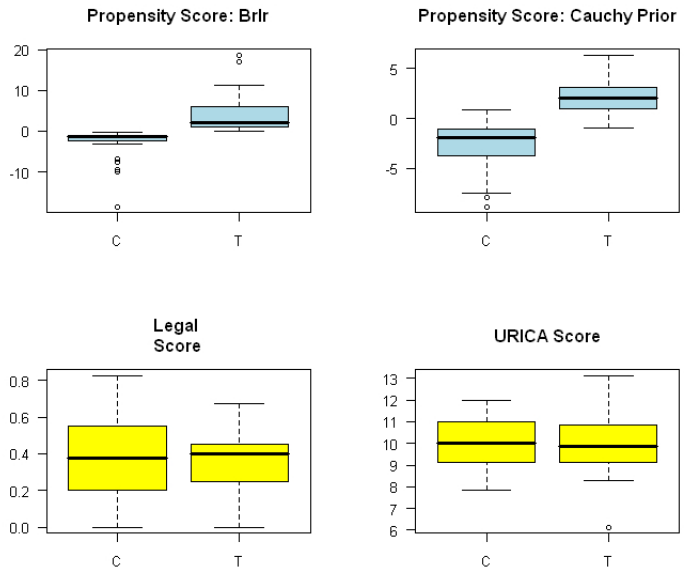


Figure 4.1: Estimated Linear Propensity Score and Covariate Plots for the PR Dataset

by including other dimension-reducing scores in the distance measure.

Prognostic Score

Like the propensity score, the prognostic score reduces the dimension of the covariates while distilling the information most relevant to the treatment effect estimation, in this case, covariate information that is most relevant to the outcome (Hansen 2008a). Matching on both propensity and prognostic scores may control bias better than matching on the propensity score alone and can enhance the precision of estimation (Hansen 2008a). Subclassification on a combination of propensity and prognostic score is, then, likely to result in an effective adjustment for pre-treatment variable differences. Due to similar concerns as in propensity estimation, to estimate the prognostic scores, we fit a penalized logistic regression model using Jeffreys' prior (`brlr` function in R) with attrition as a response variable. The predictors in this model were the demographics, principal components of treatment histories, summary RBBQ and URICA variables, ASI variables, and all the relevant missing dummy variables. There is a substantial amount of overlap between the prognostic scores on the logit scale in the control and treatment groups.

4.1.3 Additional Matching Criteria

A Second Prognostic Score

As is common in biomedically based addiction treatment, the control group was permitted to smoke, so as not to compound the challenge of quitting an illegal addictive drug with the challenge of quitting smoking. The faith-based treatment, on the other hand, saw it as to promote bodily purification, and forbade smoking. Given the programs' agreement on the importance of smoking to the outcome, we treat it as an important intermediate outcome. We thus sought to match on covariates predictive of the need to smoke, in addition to matching on propensity and prognostic factors. Since the faith-based group was not allowed to smoke, we model this test variable

inside the control group and then use the coefficients to predict the CO test scores in the treatment group. Then, we use the predicted scores for both groups in our analysis. We model this variable with ordinary least squares and use the AIC for model selection. The final variables included in the model are: age, medical, drug, alcohol, and financial information, and the Formal Practices component of the RBBQ score. The treatment and control group overlap substantially on their predicted CO test scores.

Religiosity

Because the goal of matching is adjustment for pre-treatment differences between treatment and control groups, it may be important to the researcher to match closely on specific variables in combination with matching closely on the scores we already discussed. This is especially relevant if there is a difference between treatment and control groups related to individual patients' characteristics. Since our investigation concerns the extent to which faith-based treatment affects one's attrition, we must consider the effect of this persuasion method on each subject. Though subjects' religiosity was measured in a multi-dimensional manner, those measurements are easily summarized based on existing literature (Connors et al. 1996b). The RBBQ generates two summary religiosity variables - formal practices and God consciousness (Connors et al. 1996b) and we match on these alongside the other scores.

4.1.4 Prognostic Propensity Score

Even after somewhat heavily penalizing the propensity model, it is likely to overfit, so we consider a different approach to modeling the propensity score. To estimate the causal effect of treatment on outcome we seek to remove the association between covariates and potential outcomes. Some covariates matter more than others for this

purpose so we focus on balancing those. This can be done through a prognostic propensity score. We fit the propensity model using only the prognostic scores (one based on the responses, and one based on smoking) for its estimation. If the treatment assignment depends only on observed covariates, then adjusting for the true prognostic score is sufficient (Hansen 2008a). Inference based on matching on the prognostic propensity score should result in unbiased estimation. This model only has two independent variables, and we fit the prognostic propensity model using the penalized likelihood method, utilizing the `brglm` function in `R`.

4.1.5 Matched Analysis

Matching: Rationale and Implementation

We seek to match on characteristics that distinguish treatment and control groups and are important to the outcome. Reducing the dimension of the covariates by using the propensity and related scores helps us make use of much of our data and helps obtain more good matches than matching on unreduced covariates. In this dataset, if one were to subclassify the data into a set of propensity score quantiles, at most one subclass would consist of both treatments and controls while subjects in all other subclasses would be thrown away, because there would be no subjects from the other group in their subclass. To make use of all of the advantageous matches available, we implement optimal full matching, which was introduced in section 1.2.7.

Matching: Diagnostics

To check if the distributions of observed covariates for the treatment and (matched) control groups are balanced, we follow the permutation-based balance assessment method suggested in (Hansen and Bowers 2008) and the Bayesian modeling approach to balance testing discussed in Chapter 3. The former balance test is imple-

mented using the existing package `RItools` in `R`. In both assessments, we compare balance on covariates in our post-stratified study to the balance one would expect had the same strata been blocks of a randomized experiment (Rosenbaum and Rubin 1984). Variations of matching on the combinations of the propensity score with other scores computed above improved covariate balance between the treatment and control groups; the p-value for balance goes from the marginally significant value of 0.051 to more clearly acceptable values, which will be presented below.

Matching: Results

The following tables summarize some results from full matching. The proportions of treatment and control subjects vary over sets because of the original differences on the scores included in the distance matrix. Matching on the propensity score alone (summarized in Table 4.1) resulted in one set with 28 treated subjects and 1 control, one set with 1 treated subject and 29 controls, and three sets in between. Because the estimated propensity score overlap was small, this was to be expected. The advantage of this approach is that the p-value for the balance test goes from marginally significant (0.05) before matching to 0.998 by the test from Hansen and Bowers (2008), and the Bayesian GLM approach gives a p-value of 1 after this match. Table 4.2 displays the results from matching on a combination of the prognostic

28:1	2:1	1:1	1:2	1:29
1	1	1	1	1

Table 4.1: Results of the Propensity Match for the PR Study

propensity score with the ordinary propensity score. Again, there are two sets with extreme treated-control ratios, but less so than matching on the original propensity score, because that original propensity score now plays a smaller role in the matching distance. The p-value for balance according to Hansen and Bowers (2008) is 0.98,

and according to the suggested Bayesian GLM-based test, it is 0.951. Table 4.3 shows

11:1	8:1	6:1	2:1	1:1	1:2	1:7	1:18
1	1	1	1	3	1	1	1

Table 4.2: Results of the Prognostic and Ordinary Propensity Match for the PR Study

the results from matching on a propensity, prognostic scores, and religiosity. Because the propensity score is weighed less in the dissimilarity measure, the matched sets are of a less extreme structure. The trade-off is that reduction in weight on the propensity score negatively affects balance. The p-value for balance in this setup is 0.292 according to Hansen and Bowers (2008), and 0.038 according to the Bayesian GLM method.

8:1	2:1	1:1	1:2	1:3	1:7
1	1	20	1	1	1

Table 4.3: Results of the Matching on All Data Scores for the PR Study

4.1.6 Recommendations for the Applied Problem

Taking the matching diagnostics from the previous section into account, we came up with the following recommendation for the analysis of our dataset. We consider matching on the prognostic propensity score in combination with the ordinary propensity score. This approach resulted in acceptable Type I error rates and high power, but the structure of the strata was too extreme. This might imply extrapolation - the extreme structure of the matched sets is an indicator that subjects being compared within those strata differ considerably on their estimated propensity scores. We found that by limiting the treated-control ratio to 8 and $\frac{1}{8}$ we can obtain a more desirable structure and maintain the balance p-value at an acceptable level of 0.67, according to Hansen and Bowers (2008), and 0.62, according to the Bayesian GLM measure. The matched structure for this approach appears in

8:1	7:1	2:1	1:1	1:3	1:7	1:8
2	1	1	4	1	1	2

Table 4.4: Results of the Matching on All Data Scores for the PR Study

Table 4.4. As shown in the Figure 4.2, matching on this particular combination of propensity and prognostic propensity scores with the above restriction has (for the most part) decreased the standardized differences between the two groups.

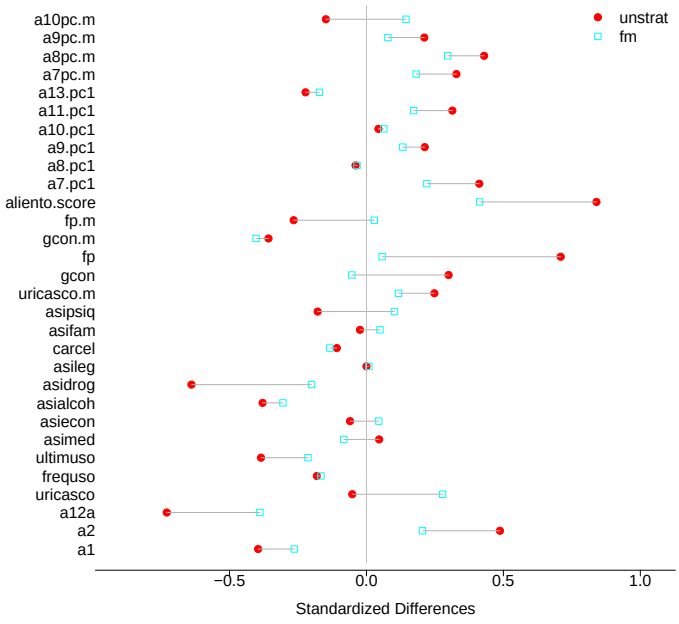


Figure 4.2: Standardized Bias Plot Pre- and Post-Matching in the PR Study

4.1.7 Treatment Effect Estimation

The Mantel-Haenszel chi-squared test gives an estimate for the odds ratio of 4.167 with a corresponding p-value of 0.03163 indicating that the odds ratio of staying in treatment is significantly different from 1. Giving similar results, a two-way ANOVA with treatment and stratification indicators as factors results in a treatment effect coefficient of 0.424, with a permutation-based null standard deviation of 0.1704 (for details on this calculation, see Hájek et al. (1999) and Hansen and Bowers (2008)).

Based on results from Hansen and Bowers (2008), this gives a double sided p-value of 0.01284. Both techniques provide statistically significant evidence for a significant treatment effect.

4.1.8 Summary

In order to balance many observed covariates in a small quasi-experiment, we use a combination of several dimension-reduction techniques with matching on the propensity and related scores. The first step in our analysis was to estimate the propensity score using a Bayesian approach to address overfitting in this small but richly observed dataset. Then we estimate the predicted attrition in both groups (the prognostic score), using the model from the control group (Hansen 2008a). Aside from these two scores, we attempt to match on other variables strongly related to treatment assignment and outcome. The results of carbon monoxide breathalyzer tests for cigarette smoke (variable "CO test score") are predictive of treatment assignment, so we sought to adjust for variables that help explain it, and this adjustment is reflected in our third computed score. These scores represent the dimensions of the covariates which will help focus matching on bias reduction and on the target of inference. Finally, we compute a prognostic propensity score by modeling treatment assignment as a function of the prognostic and smoking scores. To use our entire sample to adjust for pre-treatment differences, we full match on different combinations of the scores explained above. Before estimating the treatment effect, we assess post-matching covariate balance as a diagnostic technique to evaluate the effectiveness of our adjustment. Based on the results of balance tests, we choose a match with restrictions on the combination of propensity and prognostic propensity scores. We then proceed to estimate the treatment effect of the faith-based treatment as compared to the control biopsychosocial treatment.

4.2 Study of Effectiveness of Right Heart Catheterization

4.2.1 Introduction

Right Heart Catheterization (RHC), also known as pulmonary artery catheterization, is the passing of a catheter into the right side of the heart, used to continuously monitor the heart's function and the patient's blood flow. Cardiologists use the information gathered from RHC to guide therapy, usually in critically ill patients, whose organs are at risk of failure. There are risks associated with the procedure, including infections, bleeding, collapse of the lungs, and others. These risks may negatively affect patient outcomes in various ways, from increased hospital stay and medical bills to death. Because the procedure is generally believed to be propitious in terms of patient outcomes, the effects of the procedure have mostly been studied in observational studies rather than randomized experiments. Observational data were used by Gore et al. (1987) and Zion et al. (1990) to investigate the medical consequences of the procedure, concluding that the treatment was associated with negative patient outcomes. Still, even the authors themselves (Zion et al. 1990), questioned whether the studies adequately adjusted for important confounders. In an attempt to give a more thorough evaluation, Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT), a prospective cohort observational study of 5735 critically ill adult patients, was conducted in five US teaching hospitals between 1989 and 1994. The aim of the analysis of these data by (Connors et al. 1996b) was to investigate the effects of RHC within 24 hours of admittance into ICU on important patient outcomes, such as survival time after the procedure, cost and intensity of care, length of stay in the hospital and in the ICU and others. Because the decision whether or not to use RHC was left up to the physician, selection into treatment is likely to be confounded by patient characteristics that also affect the

outcomes, such as low blood pressure or heart abnormalities. Both are likely to prompt a doctor to use RHC and both are also likely to result in a patient's death. Because of this apparent selection bias, it is important in this observational study to balance covariates important to the outcome between the treatment and control groups before evaluating the causal effect of RHC.

4.2.2 The Dataset

The dataset is provided in the `Hmisc` (Alzola 2002) package of the `R` software. It contains the relevant covariates, such as patients' demographic and medical information, treatment information (the indicator of the performance of RHC within the first 24 hours in the ICU), and some outcome measures, such as length of stay in the hospital and an indicator of death within 180 days of performing the procedure. A summary of some covariates as well as their differences between the treated and control groups are provided in Table 4.5. A few covariates appear balanced between the two groups, according to the measure from Hansen and Bowers (2008), but many appear statistically significantly different between treated and control subjects.

4.2.3 Analysis

Propensity Score

To adjust for selection bias, we first construct a propensity score - the probability of being selected into treatment given the observed covariates, as outlined in Rosenbaum and Rubin (1983). The propensity model will include covariates important to treatment assignment and outcomes. Because of the balancing property of the propensity score, subsequent matching on it should result in similar covariate distributions between the treated and control groups. According to Connors et al. (1996b), a panel of 7 specialists in critical care specified the covariates that should

	Control (No RHC)	Treated (RHC)	P-value
age	61.76	60.75	0.03
Female	0.46	0.41	0.00
White	0.78	0.78	0.58
Black	0.16	0.15	0.26
No insurance	0.05	0.06	0.11
Private insurance	0.27	0.34	0.00
Medicare	0.27	0.23	0.01
Medicaid	0.13	0.09	0.00
income Under $11k	0.59	0.52	0.00
Income $11-$25k	0.20	0.21	0.57
Income $25-$50k	0.14	0.18	0.00
Income > $50k	0.07	0.09	0.02
Cancer	0.18	0.15	0.01
Education (in years)	11.57	11.86	0.00
Cardiovascular Symptoms	0.16	0.20	0.00
Congestive Heart Failure	0.17	0.20	0.01
Dementi	0.12	0.07	0.00
Psychiatric Symptoms	0.08	0.05	0.00
Pulmonary Symptoms	0.22	0.14	0.00
Renal Failure	0.04	0.05	0.24
Cirrhosis	0.07	0.06	0.07

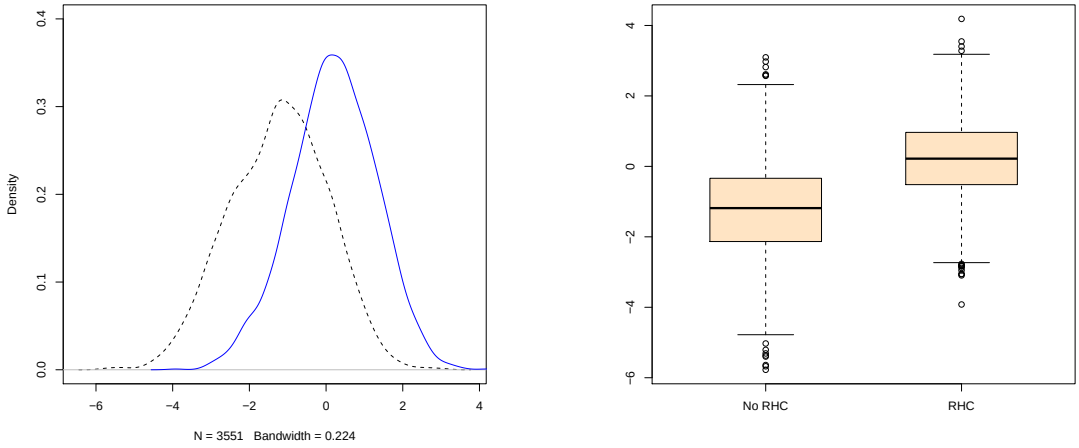
Table 4.5: RHC dataset: some pre-treatment covariates

influence doctors' decisions to use or not to use RHC. These covariates, as well as other variables deemed important to the patients' outcomes, were used as predictors in the logistic regression model with an indicator of RHC as a response. Because the number of subjects in this dataset (5735) is quite large compared to the number of parameters in the propensity model (50), overfitting is not a concern here, so we use a regular generalized linear model approach. The covariates that went into the model included age, sex, race, insurance and income information, as well as the health indicators and disease categories for the subjects provided in the `Hmisc` software.

Table 4.6 and Figure 4.3 summarize and show plots of the propensity scores on the logit scale. Note that there is a slight lack of overlap on the estimated propensity score; this is a concern that we will take on at the matching stage of the analysis.

	No RHC	RHC
Minimum	-5.77	-3.92
1st Quartile	-2.13	-0.52
Median	-1.18	0.22
Mean	-1.23	0.20
3rd Quartile	-0.34	0.96
Maximum	3.10	4.19

Table 4.6: Estimated Linear Propensity Scores for the RHC dataset



(a) Density plots for the estimated propensity score, the dashed line is for control subjects, the solid line - for treated subjects

(b) Box plots for the estimated linear propensity score

Figure 4.3: RHC Estimated Propensity Plots

Prognostic Scores

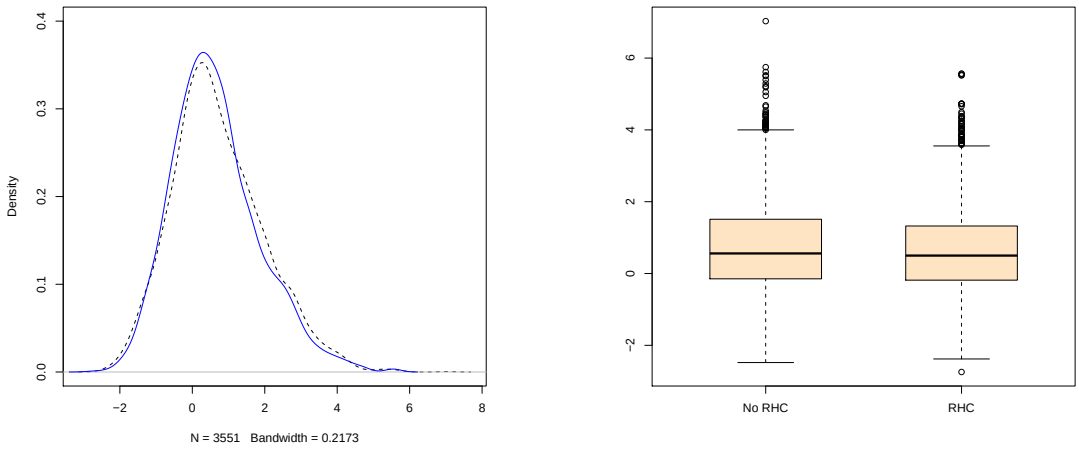
Hansen and Bowers (2008) and Hansen (2008a) emphasize the importance of adjusting for the covariates most important to the outcome. In this dataset, there are two outcomes that are worth investigating: death within 180 days of the hospital admittance and length of stay in the hospital. Although our final analysis will only evaluate the effect of treatment on patients' death outcomes, we seek to adjust for covariates to the extent that they affect both death and length of stay, since both are indicators of important medical consequences to the patient of administering RHC. First, we address covariates important to our main outcome of interest. We

use a generalized linear model to predict death within 180 days of the admittance in the control group using all the 50 covariates provided in the dataset. Then we use the coefficients from this model to predict the outcome in the treated group. The resulting fitted values in both groups are the prognostic scores. Table 4.7 gives a summary of the prognostic scores for the two groups. Figure 4.4 shows the plots

	No RHC	RHC
Minimum	-2.48	-2.74
1st Quartile	-0.15	-0.18
Median	0.56	0.50
Mean	0.72	0.66
3rd Quartile	1.51	1.32
Maximum	7.03	5.56

Table 4.7: Estimated Prognostic Scores for Death Within 180 Days of Admittance

of these prognostic scores on the logit scale. Similarly, we model the length of stay



(a) Density plots for prognostic scores in the RHC data. The dashed line is for control subjects, the solid line - for treated subjects

(b) Box plots for prognostic scores in the RHC data

Figure 4.4: RHC Estimated Prognostic Plots

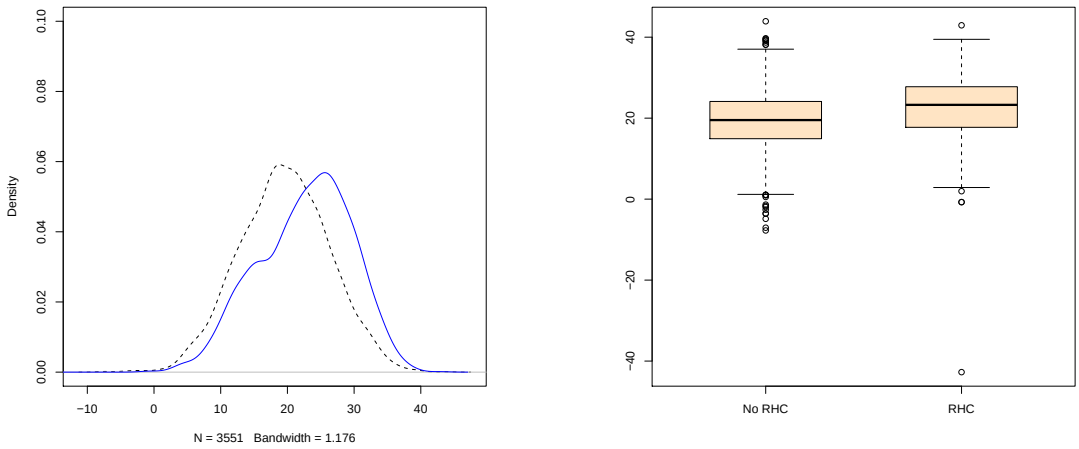
in the hospital in the control group, and use the model to predict length of stay in the treated group. However, it should be noted that short length of stay in the hospital may not always imply a positive outcome; for some subjects it is connected

to death at the hospital. The length of stay is measured in days and modeled using a linear model. Table 4.8 and figure 4.5 show the summary and plots of the second

	No RHC	RHC
Minimum	-7.74	-42.73
1st Quartile	14.92	17.74
Median	19.53	23.28
Mean	19.53	22.59
3rd Quartile	24.13	27.76
Maximum	43.93	42.94

Table 4.8: Estimated Second Prognostic Scores for the RHC dataset

prognostic score in the RHC dataset. There is one subject in the treated group with an outlying prediction. This is due to an unusual linear combination of covariates and not to any one particular covariate. We perform the matched analysis with and without this subject, to determine his impact on the final result, which turns out to be minimal.



(a) Density plots for the prognostic scores in the RHC data based on length of stay. The dashed line is for control subjects, the solid line is for treated subjects

(b) Box plots for prognostic scores in the RHC data

Figure 4.5: RHC: Second Set of Estimated Prognostic Plots

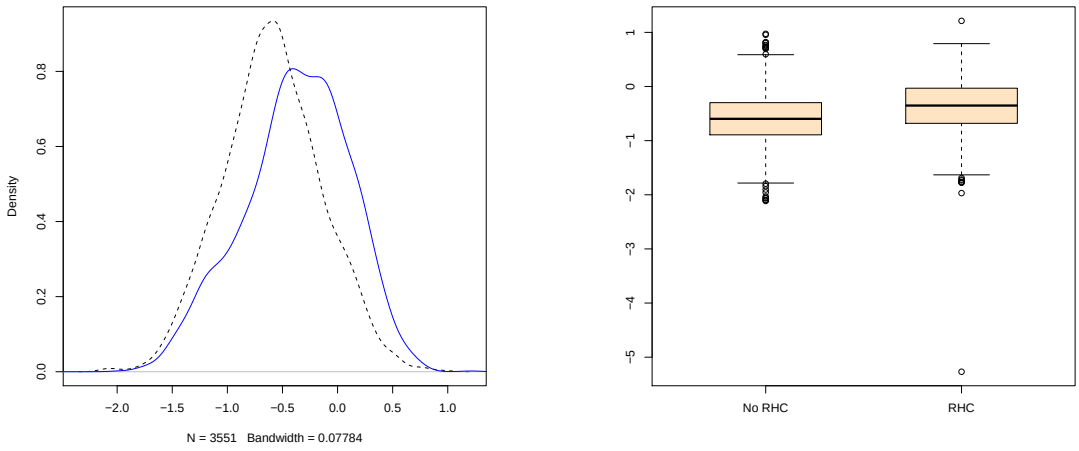
Prognostic Propensity Score

Next, we model the propensity score in an alternative way, by using as predictors the two prognostic scores described above. This should help balance those covariates which are more predictive of the outcome. The resulting fitted linear prognostic propensity score is summarized in Table 4.9.

	No RHC	RHC
Min.	-2.11	-5.27
1st Qu.	-0.89	-0.68
Median	-0.60	-0.35
Mean	-0.59	-0.38
3rd Qu.	-0.30	-0.03
Max.	0.97	1.21

Table 4.9: Estimated Prognostic Propensity Scores for the RHC dataset

Figure 4.6 shows the prognostic propensity scores for the two groups on the logit scale:



(a) Density plots for the prognostic propensity scores in the RHC data based on length of stay. The dashed line is for control subjects, the solid line - for treated subjects

(b) Box plots for prognostic propensity scores in the RHC data

Figure 4.6: RHC: Estimated Prognostic Propensity Plots

Matching and Diagnostics

To make use of most of our subjects and to optimize the similarity of matched subjects, we use optimal full matching (Rosenbaum 1991) to create matched sets of treated and control subjects. Because of some outlying values, matching on the original propensity score, although achieving a balance p-value of 1 by both measures in Hansen and Bowers (2008), and the Bayesian-based measure discussed earlier in Chapter 3, results in a matched set structure that is undesirable. The extreme high and low treated-to-control subject ratios in the structure are indicative of extrapolation - the comparison of subjects who are not matched closely on the estimated propensity scores, which makes our matched dataset unlikely to have properties similar to those of a block-randomized experiment (Hansen 2009). For this reason, we try several matching configurations to try to minimize extrapolation while maximizing balance. We impose various restrictions on our match. For example, we try limiting the treatment-control ratios to improve our effective sample size. This usually comes with a balance trade-off, since by limiting the treatment-control ratios, we deviate from matching as close as possible overall on the propensity score. Another restriction we can impose is matching within calipers on the estimated propensity score. This has the advantage of bringing subjects closer on the propensity score, thus improving balance, but is disadvantageous in that by matching within tight calipers, the technique leaves some subjects unmatched, including several treated subjects, which may change the meaning of the causal estimand, and is warned against in Rosenbaum and Rubin (1985). Table 4.10 presents some of our matching results to illustrate the trade-offs of different matching methods.

The table presents results for matching on the propensity score alone, the prognostic propensity score alone, and on the Mahalanobis distance based on the two scores.

The first row illustrates the match one obtains by matching on the propensity score alone without restrictions. As mentioned before, this match, although it results in balance on observed covariates with a p-value of 1, has undesirable matched set configurations, with treated-control ratios as extreme as 1 control subject matched up to 144 treated subjects, which we refer to as extrapolation in the table. As part of our next step, to reduce extrapolation and insure comparability of the subjects, we impose the "overlap restriction". That is, we discard subjects from that study who are not in the region of overlap on the estimated propensity score (for more on the history of restrictions like these, see Section 1.2.7). Once we imposed the restrictions on the treated-control ratio as well as overlap on the estimated propensity score (second row of the table), the balance p-value drops to 0.745, which is not a source for concern, but our overlap restriction results in excluding 8 treated and 275 control subjects. Matching just on the prognostic propensity score without restrictions does not balance observed covariates, as indicated by the 0 p-value in row 3. Putting a caliper and other restrictions on this match to avoid extrapolation results in improvement in balance to an acceptable level of 0.304, but comes at a cost of losing 3 treated and 710 control subjects. Finally, matching on a combination of the prognostic propensity and ordinary propensity score, with no restrictions, balances the covariates nicely, but has an undesirable matched set structure. By using `optmatch` to exclude 10% of control subjects from the analysis to minimize the total discrepancy on the propensity score, we are able to achieve good balance on observed covariates (p-value of 0.98 by measure from Hansen and Bowers (2008), p-value of 1 from the GLM-based method from Chapter 3) without sacrificing any of the information on treated subjects. Furthermore, this match corresponds to an effective sample size of 1820 matched pairs, which is the maximum effective sample

size accomplished by any of the configurations listed. The structure of this preferred matched dataset follows in Table 4.11. On the extremes, the matched sets have treated-control ratios of as small as 1/10 and as large as 10. The effective sample size is equivalent to 1820 matched pairs.

Match	Prop. Caliper	T Lost	C Lost	Restricted	Extrapolation	Balance p
Prop	∞	0	0	No	Yes	1
Prop	∞	8	275	Yes	No	0.745
Prog Prop	∞	0	0	No	Yes	0
Prog Prop	.2	3	710	Yes	No	0.304
Prog Prop and Prop	∞	0	0	No	Yes	1
Prog Prop and Prop	∞	0	355	Yes	No	.98

Table 4.10: Matching Results for the RHC Dataset. The description and discussion of this table are provided in section 4.2.3 on page 93.

5-10:1	4:1	3:1	2:1	1:1	1:2	1:3	1:4	1:5-10
72	36	61	112	613	163	90	60	188

Table 4.11: RHC Matched Sets Configuration

Balance on observed covariates started out in this study with a p-value of approximately zero. Using full matching on prognostic and ordinary propensity scores and imposing restrictions has raised the balance p-value to 0.98 by the Hansen-Bowers measure and 1 by the Bayesian-based measure from Chapter 3. Figure 4.7 shows the striking improvements in standardized bias from the unstratified study to the matched dataset.

Outcome Analysis

The Mantel-Haenszel test results in a conditional odds ratio for death of 1.38 with a 95% confidence interval of (1.199 1.588). The test statistic is $\chi^2 = 20.177$, with a corresponding p-value of 0.000007. These results confirm the pessimistic findings of Connors et al. (1996a) and are consistent with the theory that the RHC procedure has a negative impact on the patients' odds of survival within 180 days of admittance. Another way of estimating the treatment effect gives a similar result:

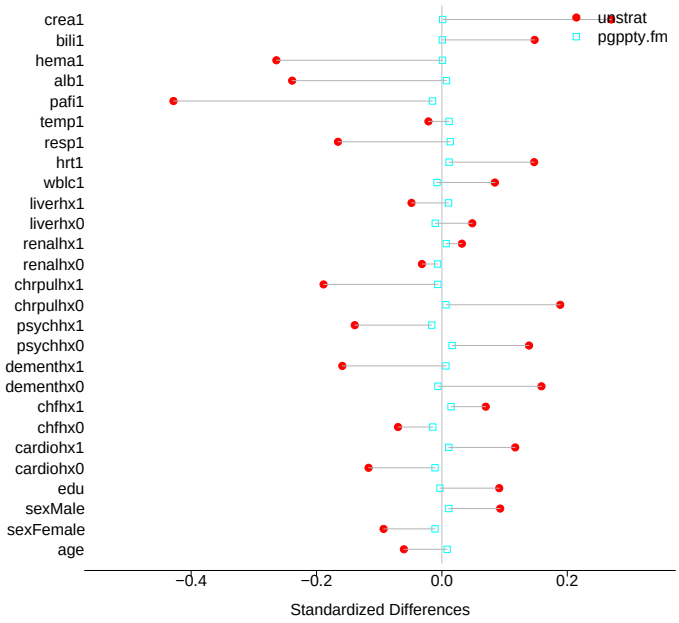


Figure 4.7: Standardized Bias Plot Pre- and Post-Matching in the RHC Dataset

the treatment coefficient from the two-way ANOVA (with treatment and matched set indicator as factors), is 0.070, with a permutation-based null standard deviation of 0.016, resulting, based on a procedure from Hansen and Bowers (2008), in a 2-sided p-value of 0.000012.

Summary

We sought to balance the treated and control groups in this dataset with respect to important confounders to investigate the effect of Right Heart Catheterization on patient death within 180 days of admittance. To match treated and control patients as closely as possible on important characteristics, we reduced the dimension of the covariates in the ways important to treatment assignment and the outcome. First, we fit a propensity score - the predicted probability of assignment to treatment based on observed covariates. Next, we estimated two prognostic scores in the control group and used them to predict the values in the treated group, as described

in Hansen (2008a). The two scores were based on the patient death outcome, as well as the length of stay in the hospital. We also computed a prognostic propensity score by modeling treatment assignment as a function of the two prognostic scores. By full matching on several distances with and without restrictions, and using balance diagnostics as well as effective sample size to evaluate the quality of resulting stratifications, we arrived at one possible solution to the matching problem. Our final match is based on a combination of the prognostic propensity and ordinary propensity scores, with restrictions on the treated-control ratio within matched sets and an omission of 10% of the control subjects. This match resulted in good balance by both measures we used, as well as a high effective sample size. We used two different methods for treatment effect estimation, each being consistent with previous research indicating a possible harmful effect of RHC on patient survival.

4.3 Limitations

Our results in both case studies are limited by the possibility of omitting important covariates. For this reason, Rosenbaum (2002a) and Hosman et al. (2009) give recommendations for sensitivity analyses for estimates of causal effects in observational studies.

CHAPTER V

Simulation Study

5.1 Aims and Objectives

Though the procedures for balance assessment discussed in Chapter 2 are used by many practitioners, the literature still lacks a thorough investigation of their statistical properties. We aim to evaluate the statistical properties of a variety of procedures currently used in literature for testing balance on covariates in comparative studies with a simulation study. We also seek to assess the performance of the new balance-testing procedures described in Chapter 3. We are interested in the performance of these procedures in different data-induced settings, and in combination with common statistical adjustments. Information obtained from studying these techniques will help us to determine the advantages and weaknesses of currently widely-used techniques as well as to evaluate the three newly-proposed techniques. We seek to design simulation conditions that bear some resemblance to datasets that arise in practice rather than restricting our attention to completely simulated data. Below follows a description of both the computing and statistical sides of the simulation.

5.2 Datasets

All the datasets for our simulation stem from real data. They are created based on two original datasets: the Puerto Rico study and the Right Heart Catheterization

study which were described in detail and analyzed in the previous chapter. The following sections provide a refresher on the two datasets as well as new information about them to help describe the simulation setup.

5.2.1 Puerto Rico Dataset

These data were collected from a prospective observational study of 67 Puerto-Rican men who were court-mandated to either biopsychosocial (control) or faith-based (treatment) substance abuse rehabilitation programs. Since the Puerto-Rican court system does not distinguish between religion-based and secular rehabilitation facilities, and because subjects had little discretion over their treatment assignment, selection into treatment in this dataset is likely to be weak. The dataset is richly observed, with 27 covariates after some statistical adjustments, as was discussed in the previous chapter. The covariate information collected includes demographic descriptives, subjects' histories of substance abuse, economic and legal information, a smoking test result and measures of religiosity. The primary outcome of interest is a binary measure of program retention after three months.

As described in section 4.1.2, the main challenge in analyzing this data is the small number of events relative to the number of covariates. To address this issue, we reduced the covariates in this dataset into five scores: the ordinary propensity score, the prognostic score (discussed earlier in section 1.3.1), and three more dimensions of the data. This dataset is particularly useful for evaluating techniques in a small-sample setting.

5.2.2 Right Heart Catheterization Dataset

The other dataset that we use is a prospective cohort study of 5735 clinically ill adult patients that examined the association between the use of right heart catheter-

izaton (RHC) during the first 24 hours in the intensive care unit and subsequent outcomes, of which we will be using the indicator of death within 180 days of admittance to the ICU. Because RHC is believed by some physicians to be necessary to guide therapy for certain kinds of patients, treatment assignment is confounded with patient characteristics related to the outcome, and selection bias is an important consideration in this observational study. A list of possible important confounders was specified by a panel of specialists, and all the variables on that list were included in the analysis. They include demographic and health-related information for each subject. The covariate can be reduced in dimension in the following ways: the propensity score, and two prognostic scores - one based on the main outcome (death within 180 days of the procedure), and one based on an intermediate outcome (length of stay in the hospital). In this dataset the number of confounders is small compared to sample size, so we will use it to study procedures in a medium-sample setting. We will separately take random samples of subjects from this dataset to illustrate our techniques in a modest-sized and a medium-sized study.

5.3 Software / Computing

The simulations were run using the Center for Advanced Computing at the UM Engineering clusters to shorten computing time. We use `R 2.7.2` for the simulations. Each simulation run is independent of previous runs. Random number generation was based on the `.Random.seed` function in R. While in some cases each run of the simulation was run separately to speed up computing, we ensured that the random seed to start each of those runs did not repeat. 1000 simulations were performed for each scenario. Afterwards, all the results were assembled together into a large dataset for each procedure to facilitate analysis of each technique's statistical properties. R

has the advantage of hosting software necessary for key parts of this simulation. In particular we use the following packages:

- We use the package `Hmisc` to obtain the Right Heart Catheterization data and for missing data imputation in the Puerto Rico dataset.
- We use the package and function `brglm` for modeling propensity and prognostic scores, and the outcome using bias-reduced logistic regression, penalizing the likelihood function by Jeffreys prior
- We use the R package `arm` and its function `bayesglm` for modeling the propensity and prognostic scores, and the outcome using the Bayesian method described in Gelman et al. (2008)
- We use the package `optmatch` for optimal full matching (Hansen 2007).
- The package RItools (Bowers and Hansen 2006) is used for checking balance with the procedure described in Hansen and Bowers (2008)
- The `lmer` function within the package `lme4` is used for mixed effects modeling of the response.
- Other `R` packages that are required by the packages listed above were also used.

5.4 Scenarios to be Investigated

5.4.1 Sample Size

1. We will use the Puerto Rico (PR) study with 67 subjects and 27 confounders to investigate the small sample properties of our techniques.
2. We will also use a random subset of 500 subjects with 50 confounders from the Right Hearth Catheterization (RHC) dataset from Connors et al. (1996a) to investigate modestly sized sample properties. Because the literature suggests that in logistic regression (relevant for modeling the propensity score, for example),

about 10 observations per confounder are necessary, we chose a sample size of 500 to reflect this recommended boundary condition.

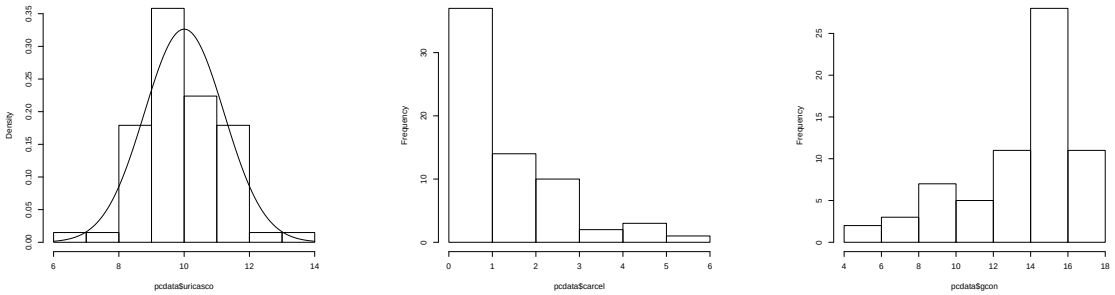
3. We will use a random subset of 1433 subjects (1/4 of the RHC dataset) to investigate how the procedures perform in a medium sample.

5.4.2 Distributions of Observed Covariates

Instead of simulating normally distributed covariates, we use existing features of our data which include variables, the empirical distributions of which appear close to the normal distribution, as well as some naturally-occurring departures. Below are some examples of the data we use.

Puerto Rico Data

Figure 5.1 shows some examples of distributions of quantitative observed covariates in the Puerto Rico dataset. In the first panel, we show the distribution of the variable URICA (University of Rhode Island Change Assessment scale). The shape is close to the normal distribution. In contrast, the next two variables show the number of times in prison and God consciousness variables, which are right-skewed and left-skewed, respectively.



(a) URICA Scores for PR Subjects

(b) Number of Times in Prison for PR Subjects

(c) God Consciousness for PR Subjects

Figure 5.1: Distributions of some quantitative variables in the PR study

Right Heart Catheterization Data

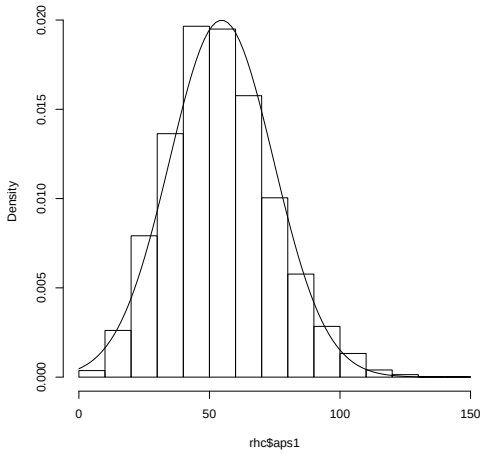
Figure 5.2 shows some examples of distributions of quantitative observed covariates in the RHC dataset. The first is the APACHE II score ("Acute Physiology and Chronic Health Evaluation II"), which is a severity of disease classification system. It follows roughly a normal distribution. It is followed by the distribution of age which looks left-skewed. The next variable that is right-skewed is $PaCO_2$ which denotes the partial pressure of carbon dioxide in the arterial blood. Finally, the variable Mean Blood Pressure (over the observation period) is bimodal. All these distributions occur naturally in our dataset and represent departures from normality that one might expect to see in another study. This variability in distributions of observed covariates as opposed to simulation of normally-distributed variables allows us to create several interesting data settings for the simulation study.

5.4.3 Using Observed Treatment Assignment and Outcome Data

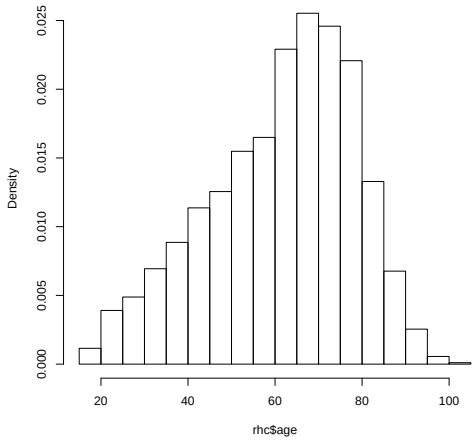
Our simulation study relies on real rather than manufactured data sources. In this section we explain how we can use observed information on treatment assignment and outcomes to our advantage in the simulation setting.

Propensity Score and Overlap

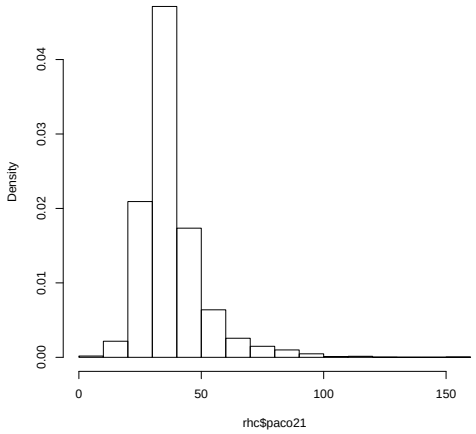
Before the start of the simulation, we estimate the propensity and prognostic scores for each dataset, as was described in the previous chapter (in the PR study, the propensity scores were estimated using Bayesian methods, with a Cauchy prior on logistic regression coefficients). We take the fitted values from these models and store them as true propensity and prognostic scores for future simulation runs. This allows us to simulate several overlap conditions on the true propensity score which might arise in practice as well as base our simulation treatment assignment mechanism on



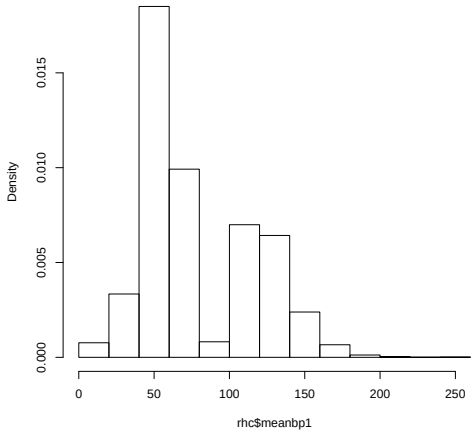
(a) APACHE II Scores for RHC Subjects



(b) Age Variable for RHC Subjects



(c) $PaCO_2$ Variable for RHC Subjects

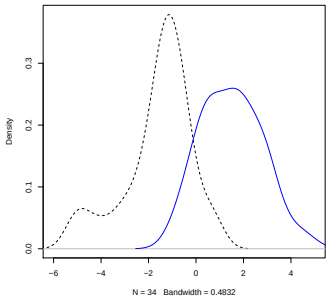


(d) Mean Blood Pressure for RHC Subjects

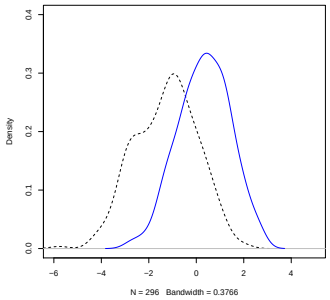
Figure 5.2: Distributions of some quantitative variables in the RHC study

one that stems from real data. Our three datasets embody different structures of overlap on the true propensity score. In the PR dataset, likely because the logistic propensity model is overcrowded by covariates, there is little overlap between the distributions of estimated propensity scores in the treated and control groups. In the RHC datasets, in contrast, the overlap is quite large, as shown below in Figure 5.3. In all cases, we will first consider matching all the observations regardless of overlap to preserve all the subjects in the sample, and then introducing a propensity

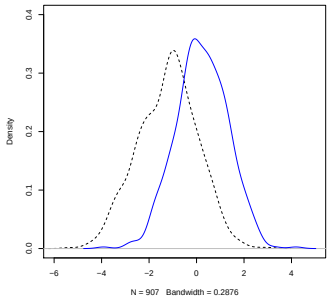
caliper to improve the measures of balance on observed covariates. In the presence of a propensity-score caliper, those subjects with extreme propensity scores are likely to remain unmatched and not considered in treatment effect estimation.



(a) PR Dataset (67 subjects)



(b) RHC Study, Modest Sample (500 subjects)



(c) RHC Study, Medium Sample (1433 subjects)

Figure 5.3: Simulation 'True' Propensity Score Density Plots. The dashed line is for control subjects, the solid line is for treated subjects

In our simulation study, the "true" model for the propensity score is known, and it is of the general nature that is hypothesized in the fitting of the propensity scores within simulation runs. This will allow us to mostly address the "errors of aggregation" rather than "errors of specification" as they were introduced and described in Section 2.2. In other words, we will be particularly interested in covariate balance as an indicator of the quality of the post-stratification procedure, as opposed to a way to evaluate the specification of the propensity model.

Outcomes and Treatment Effects

In our simulation study, we plan to use the original observed outcomes from the existing data sources. Specifically, we designate the vector of observed responses to be the vector of subjects' responses to the control condition (this was denoted by Y_c in Section 1.2.5 on potential outcomes). Some important statistical methods can be addressed having just the information on the vector of Y_c; we use it to simulate the

condition of no treatment effect which allows us to evaluate the test of no effect. To simulate treatment effects of ± 5, we will change the binary Y_c values for 5 subjects selected at random.

5.4.4 Relationship of X with Z and with Y

In our two original datasets, we are able to observe variables that are weakly and strongly related to treatment assignment and the outcome. To describe that relationship, we use Pearson correlation coefficients. Tables 5.1 and 5.2 provides some examples of our observed covariates and their correlations with treatment assignment and the outcome for the PR and the RHC studies.

Variable	Correlation with Z	Correlation with Y
Other substance abuse treatments	-0.347	-0.182
Formal practices	0.343	0.190
Number of grades completed	0.240	0.054
Age	-0.197	0.075
God Consciousness	0.150	0.210
URICA Score	-0.026	-0.111

Table 5.1: Correlations of Covariates with Treatment and Outcome in the PR Study

Variable	Correlation with Z	Correlation with Y
APACHE Score	0.239	0.192
Mean Blood Pressure	-0.212	-0.104
PaO2/FIO2 ratio	-0.203	0.020
Weight	0.122	-0.045
Number of Years of Education	0.044	-0.025
Age	0.029	0.220

Table 5.2: Correlations of Covariates with Treatment and Outcome in the RHC Study

5.5 Statistical Methods to be Evaluated

Although the main goal of this simulation study is to evaluate balance testing procedures, we are also interested in being able to describe scenarios and research approaches that affect the performance of these balance tests. For this reason we will look at several analysis and estimation techniques in combination with our balance

assessments of interest.

Before the propensity score can be modeled, each simulation run requires a treatment assignment vector. As mentioned in section 5.4.3, we start by estimating the propensity score in the two original datasets, using the Bayesian fitting method in the PR study and a penalized model in the RHC study. We then take these propensity scores to be "true" propensity scores for the rest of the simulations. Using these vectors of "true" propensity scores, we use maximum entropy sampling[1] to assign subjects to treatment and control. We ensure that the total numbers of treated and control subjects are held constant throughout the simulations. This sampling method allows us to preserve some of the relationship between the covariates and treatment assignment that we observe in our datasets.

5.5.1 Modeling the propensity score

Because neither the PR nor the RHC dataset provides us with complete overlap on the true propensity score, we use regularized techniques in modeling the estimated propensity score. In the PR dataset, we adapt a Bayesian approach by putting a Cauchy prior on the coefficients and using the EM algorithm to converge on the estimates as discussed in Chapter 3 and in the paper by Gelman et al. (2008). This provides a large reduction in overfitting the logistic regression model. In the RHC dataset, because of the larger sample size, such a heavy penalty is likely not necessary so we used a penalized likelihood approach from Firth (1993).

5.5.2 Subclassification Techniques

We plan to assess several stratification methods:

[1] Maximum entropy sampling is a way to select a sample from a vector of inclusion probabilities. For details, see the description of the package `sampling` in `R`.

Fixed-Width Stratification

We chose to modify the traditional approach of stratifying in quintiles of the propensity score. Cochran (1968) discussed the reduction in bias on a continuous covariate that results from stratification into quintiles on that covariate. Our modification of this approach is designed to ensure comparability of units that belong to the same subclass. To support subjects within the same stratum being close on the estimated propensity score, we use fixed-width stratification, with width of the propensity range dependent on the particular simulation setting. It should be noted that other approaches to stratification, such as variable-bandwidth stratification methods described in Galdo et al. (2009) may be preferable but will not be discussed here. In this simulation, subjects are divided into strata based on some number of pooled median absolute deviations[2] on the propensity score. The number of MADs starts out at 2 and decreases with an increasing demand on the balance test p-value (this part of the simulation is discussed a bit later in Section 5.5.4). This idea resembles matching within propensity calipers. Closer matching should improve balance, but this comes at the cost of leaving some subjects unmatched and out of the analysis stage. The algorithm starts at the mean of the medians of the propensity score in the treated and control group. Then we build a stratum of the specified width (in terms of the number of MADs) using that datapoint as the center. From there, we continue imposing stratum bounds using the prespecified width criterion on both sides of the resulting interval. This can result in some subjects being placed in a stratum without counterparts from the opposite group, in which case these subjects have to be discarded from treatment effect estimation.

[2]The median absolute deviation (MAD) is a robust measure of the variability:

$$\text{MAD} = \text{median}_i \left(|X_i - \text{median}_j(X_j)| \right)$$

Matching

- We consider the optimal full matching algorithm (Rosenbaum (1991), Hansen (2004)) because it results in the closest possible matches of subjects on a given distance measure.
- Matching will be performed with and without the restrictions of a propensity caliper, depending on the balance requirement within the simulation run
- Matching distances that go into the optimal algorithm will also vary. We will consider matching on metrics based on the following:
  - Estimated propensity score
  - A combination of the propensity and prognostic scores
  - Matching on all data-generated scores (for the PR dataset, propensity, prognostic and other data reduction scores, for the RHC dataset - propensity and two prognostic scores)
  - Prognostic propensity score
  - For the PR dataset - data reduction scores and prognostic score and for the RHC dataset - the prognostic scores (in both cases excluding the propensity score from the matching distance). This is done to study the importance of the propensity score in the matching distance, especially in the small dataset, where the propensity overlap is very limited.

5.5.3 Methods for Balance Testing

- Kolmogorov-Smirnov test
- Smith-Todd method
- Hansen-Bowers test

The next three models are newly suggested in Chapter 3. As a reminder, they use the following comparison for balance assessment. The main idea is to compare the following models:

$$H_0 : Pr(Z = 1) = \text{logit}^{-1}\left(\gamma_1 s_1 + \cdots + \gamma_k s_k\right)$$

$$H_a : Pr(Z = 1) = \text{logit}^{-1}\left(\beta_1 x_1 + \cdots + \beta_p x_p + \gamma_1 s_1 + \cdots + \gamma_k s_k\right)$$

Existing research shows that simply comparing these models as in the ANOVA framework using ordinary asymptotics does not work, and results in extremely high false rejection rates (Lee (2008), Hansen and Bowers (2008). Further, this kind of an approach assumes an underlying superpopulation, which, as discussed in Imai et al. (2008), may not be desirable. We thus suggest the following modifications to the process of choosing between the two models.

- Permutation-Based Likelihood Ratio method through the generalized linear model

  As part of this approach, we record the LRT statistic and then proceed to permute treatment assignments within matched sets, which is reasonable under a block-randomized experiment framework. Then, to obtain the p-value, we compare the LRT from the dataset to its permutation distribution to see how this study compares to other possible studies block-randomized in this fashion.

- Permutation-Based AIC method through bias-reduced logistic regression

  Because logistic regression in the setting discussed above is likely to overfit, we consider a similar approach as above but with penalized likelihood instead of regular GLM used to estimate the models. We use the AIC difference between the two models and its permutation distribution to assess balance on covariates.

- Permutation-Based Likelihood Ratio method through Bayesian inference

Finally, another and a more conservative way to address overfitting in the above models is to use a Bayesian approach in estimation. Our final suggestion for balance assessment first fits the two models with a Cauchy prior on the coefficients, then computes the LRT statistic, and then compares it to its permutation distribution to help assess the similarity of our dataset to other possible datasets with a similar block-randomized structure.

5.5.4 Balance Targets

One way to tell whether a balance test has favorable statistical properties is to impose various balance requirements within our simulation runs. That is, within each simulation run, we will check the post-stratification balance against a preset list of balance targets. If the post-stratification balance p-value is smaller than the requirement, we impose progressively more demanding restrictions on the match in the form of propensity calipers until the balance target is met. If a balance test is successful at detecting large covariate differences between groups, then increasing the balance requirements should positively impact the statistical properties that we will set out to investigate.

5.5.5 Methods for Treatment Effect Estimation

- Two-Way ANOVA with Factors Based on Treatment and Matched Sets

  This method gives an estimate that is based on the coefficient β of the treatment indicator z in the OLS model also adjusting for matched sets $s_1, \ldots, s_M$:

  $$Y = \beta z + \gamma_1 s_1 + \cdots + \gamma_M s_M$$

  Hansen and Bowers (2008) suggest a permutation-based variance estimate based on the null hypothesis of no effect, to go with this treatment effect estimate.

- Mixed effects regression with random effects for the stratum indicator

  This method is suggested in Agresti (2002) for the consistency of treatment effect estimates. Given the estimated propensity score $\hat{e}$, the estimated prognostic score $\hat{\varphi}$, and the stratum membership indicator s, the estimate of interest is the coefficient β of the treatment indicator z in the following regression:

$$Y_{it} = \beta z_{it} + \alpha_1 \hat{e}_{it} + \alpha_2 \hat{\varphi}_{it} + \gamma_i s_i + u_{it}$$

- Ordinary least squares regression

  Finally, one might estimate the treatment effect using linear regression but include information other than simply the treatment indicator and strata indicators. The models we test are

  - Outcome as a function of treatment, stratification, and covariates
  - Outcome as a function of treatment, stratification, and the propensity score
  - Outcome as a function of treatment, stratification, propensity and prognostic scores (in the case of the PR dataset, also other data-reduction scores)

5.6 Values to be Stored for Each Simulation

- An indicator whether or not the null hypothesis of balance was rejected
- The propensity caliper imposed within the simulation run
- A p-value for balance from each condition within each run
- Treatment effect estimate
- Variance of the treatment effect estimate (varies depending on the estimation method)

5.7 Criteria to Evaluate the Performance of Statistical Methods for Different Scenarios

Assessment of Bias

Since treatment effects are simulated by introducing random changes of a fixed size to the response vector, and treatment effect estimates are recorded, it should be easy to assess bias as the difference between the two quantities. An adjustment for a stratification that passes a balance test should result in smaller bias in the treatment effect estimate. Insisting on a high balance p-value should decrease this bias to negligible levels.

Assessment of Coverage

Most previous investigation of the statistical properties of balance tests has been dedicated to the Type I error rate of the balance test, before the treatment effect is estimated. In contrast, we investigate the Type I error rate in estimation of the treatment effect. In the case when there is no treatment effect, Type 1 error for the null hypothesis of no treatment effect will be computed and compared with the nominal rate of 0.05. If a given stratification passes a balance test, we expect that a post-matching adjustment would incorrectly reject the null hypothesis of no effect no more than 5 percent of the time. Higher levels of balance should correspond to better coverage rates.

Assessment of Power

By simulating outcomes so that there is a treatment effect, we will be able to evaluate whether a particular combination of a balance assessment and outcome analysis technique is powerful at determining a treatment effect. Because in some simulation conditions higher balance p-values are accomplished by matching or stratifying

within propensity calipers which might result in exclusion of subjects from the study, we expect that there will be a trade-off between balance and power.

5.7.1 Simulation Algorithm

1. Using maximum entropy sampling Chen et al. (1994), we compute a new vector of assignment to treatment based on the true propensity score.
2. Using the new treatment assignment vector, we compute the new estimated propensity score. Particularly in our small dataset, but also in the others, we seek to avoid overfitting the propensity score to prevent separation. For this reason, we use the Bayesian fitting approach discussed in the previous chapter to fit the estimated scores in the PR study.
3. The response data is either taken as is (for simulating no treatment effect), or 5 subjects selected at random have their response data changed to simulate the treatment effect of ± 5. In a related study in the PR dataset, we also explore a different response surface, in which we assign the new outcome vector based on the prognostic scores estimated from the original PR dataset.
4. Using the new treatment assignment, we compute the prognostic score(s). These are computed using ordinary least squares for continuous outcomes and using penalized likelihood for binary outcomes to avoid overfitting.
5. Next, we post-stratify treated and control subjects into small groups. This is done using either full matching or fixed-width stratification.
6. We evaluate balance for this match
7. We check the p-value from the balance test against the balance requirement
8. If the p-value from the balance test is greater than or equal to the requirement, we proceed to estimate the treatment effect, and record it (this is done in several

ways described in section 5.5.5). Then, we move on to the next balance target and go back to the previous step.

9. Otherwise, impose a (stricter) caliper in case of matching or stratify within a smaller number of MADs and go back to step 7.

5.8 Results

5.8.1 Small Sample Results: PR Study

Type 1 Error Rates

Tables 5.3, 5.4, and 5.5 present Type 1 error rates for the Hansen-Bowers approach, GLM-based permutation test and Bayesian-based permutation tests. Here we only show results for estimating treatment effects using ANOVA with a permutation-based standard deviation. The other methods of estimation and their statistical properties don't fare quite as well, and are presented in the Appendices A-D. The first column of the table is the balance p-value requirement. The Hansen-Bowers procedure gives correct Type 1 error rates when subjects are matched on the propensity score or the prognostic propensity score (margin of error due to natural variation in the simulation results is 0.01, so 0.06 Type 1 error rate is within that error of the nominal rate of 0.05). Matching on propensity and prognostic score, all data-reducing scores, and matching without the propensity score only perform well when the balance requirement is quite strong at 0.50. Importantly, going down the rows of the table, as the balance requirement gets stricter and subjects are matched within smaller calipers on the estimated propensity score, the Type 1 error rates get smaller. This indicates the relationship between covariate balance and error rates that we would expect: balance on observed covariates helps assure correct rejection rates for the hypothesis of no effect, though this finding may be confounded by the decrease in sample size associated with imposing propensity calipers. Another finding from

this table is that despite poor overlap on the estimated propensity score, matching on a distance that does not incorporate the propensity score results in highest false rejection rates.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.05	0.08	0.16	0.19	0.06
0.05	0.05	0.08	0.16	0.19	0.06
0.2	0.05	0.08	0.14	0.17	0.06
0.3	0.05	0.08	0.10	0.11	0.06
0.5	0.05	0.08	0.04	0.05	0.06

Table 5.3: Type 1 Error Rates for PR Study Using the Hansen-Bowers Method

The GLM-based method for evaluating balance in this small study still has high rejection rates. This occurs because the logistic regression model used in the balance test strongly overfits and fails to converge; it does so both for our balance model and in the permutation distribution. Because of estimation problems in the overfit models, the test tends to return very high p-values, indicating that the study under consideration is almost always balanced on observed covariates. For this reason, there is no improvement as we move down the table across balance targets. Regardless of the matching distance, the Type 1 error rates are too high.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.07	0.09	0.19	0.22	0.08
0.05	0.07	0.09	0.19	0.22	0.08
0.2	0.07	0.09	0.19	0.22	0.08
0.3	0.07	0.09	0.19	0.22	0.08
0.5	0.07	0.09	0.19	0.22	0.08

Table 5.4: Type 1 Error Rates for PR Study Using the GLM-based Method

Using Bayesian inference in the logistic regression models for the balance test, we find that matching on the propensity score gives the closest Type 1 error rates to the nominal levels. In contrast to the Hansen-Bowers balance measure, this approach also gives acceptable false rejection rates when we insist on a balance p-value of 0.20 for matching on all data-based scores and matching without the propensity score.

Also, again, in contrast to the Hansen-Bowers measure, matching on the combination of prognostic and ordinary propensity score, just misses the mark for correct Type 1 error rates.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.06	0.09	0.16	0.19	0.07
0.05	0.05	0.08	0.07	0.07	0.07
0.2	0.05	0.08	0.05	0.05	0.07
0.3	0.05	0.08	0.05	0.05	0.07
0.5	0.05	0.08	0.05	0.05	0.07

Table 5.5: Type 1 Error Rates for PR Study Using the Bayesian-based Method

Treatment Effect Estimates

Tables 5.6 - 5.9 summarize the treatment effect estimates and their permutational null standard deviations for the PR study, using permutational ANOVA and two balance-testing methods - Hansen-Bowers and Bayesian-based method. The true treatment effect is zero in all cases, so these results help assess bias in the estimates. Insisting on a large balance p-value results in a slight increase in the variance of the estimates because it leads to a decrease in sample size. Based on the estimates and their standard deviations listed below, there does not appear to be any sizeable difference between methods.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.00	0.09	0.14	0.15	0.06
0.05	0.00	0.09	0.14	0.15	0.06
0.2	0.00	0.09	0.12	0.13	0.06
0.3	0.00	0.09	0.10	0.10	0.06
0.5	0.00	0.09	0.03	0.01	0.06

Table 5.6: Treatment Effect Estimates for PR Study Using the Hansen-Bowers Method

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.25	0.17	0.12	0.11	0.17
0.05	0.25	0.17	0.12	0.11	0.17
0.2	0.25	0.17	0.14	0.13	0.17
0.3	0.25	0.17	0.15	0.15	0.17
0.5	0.25	0.18	0.29	0.30	0.19

Table 5.7: Treatment Effect Standard Deviations for PR Study Using the Hansen-Bowers Method

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.02	0.10	0.14	0.15	0.07
0.05	0.01	0.10	0.08	0.07	0.07
0.2	0.01	0.10	0.07	0.06	0.07
0.3	0.01	0.10	0.06	0.06	0.06
0.5	0.01	0.09	0.06	0.06	0.06

Table 5.8: Treatment Effect Estimates for PR Study Using the Bayesian-based Method

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.25	0.18	0.11	0.11	0.18
0.05	0.25	0.18	0.16	0.17	0.18
0.2	0.25	0.18	0.17	0.17	0.18
0.3	0.25	0.18	0.17	0.18	0.18
0.5	0.25	0.18	0.17	0.18	0.18

Table 5.9: Treatment Effect Standard Deviations for PR Study Using the Bayesian-Based Method

Power Estimates

The power estimates for the common treatment effect of ± 5 subjects are given in tables 5.10 and 5.11. If we restrict our attention to the modes of balance testing and inference that provided acceptable false rejection rates, we see that there is not much difference between the Hansen-Bowers and Bayesian permutation approach, whether matching is done on the propensity score or if we compare matching on the propensity and prognostic score for the Hansen-Bowers measure with matching on all scores or without the propensity scores combined with the Bayesian-based permutation test.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.26	0.35	0.46	0.47	0.34
0.05	0.26	0.35	0.46	0.47	0.34
0.2	0.26	0.35	0.44	0.45	0.34
0.3	0.26	0.35	0.40	0.40	0.34
0.5	0.26	0.35	0.27	0.25	0.34

Table 5.10: Power Estimates for PR Study Using the Hansen-Bowers Method

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.25	0.37	0.47	0.48	0.36
0.05	0.24	0.37	0.37	0.36	0.35
0.2	0.24	0.37	0.34	0.34	0.35
0.3	0.24	0.36	0.34	0.34	0.34
0.5	0.24	0.36	0.33	0.34	0.34

Table 5.11: Power Estimates for PR Study Using the Bayesian-Based Method

5.8.2 Modest Sample Results: RHC Study

Bayesian Method

Table 5.12 presents Type 1 error rates for estimating the treatment effect using the ANOVA approach with a permutation-based variance, described above, with covariate balance evaluated using the Bayesian-based permutation test. The first line indicates the lack of a balance requirement. Matching on any combination of propensity and prognostic scores seems to work well, in contrast to matching without the propensity score, which has a Type I error rate of 0.12, exceeding its nominal level. Insisting on the balance p-value being at least 0.05 by imposing propensity calipers, however, brings that Type I error rate down to an acceptable level. Detailed results for other types of estimation for the treatment effect can be found in Appendix E; both, the mixed effects regression adjustments and OLS regression (on the treatment, stratification indicator, and the propensity score) produce acceptable Type I error rates, with OLS estimates slightly on the conservative side.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.03	0.05	0.04	0.12	0.03
0.05	0.03	0.05	0.04	0.06	0.03
0.2	0.03	0.05	0.04	0.06	0.03
0.3	0.03	0.05	0.04	0.06	0.03
0.5	0.03	0.05	0.04	0.05	0.03

Table 5.12: Type I Error Rates for the Small Sample from the RHC Study Using the Bayesian-Based Method

Table 5.13 gives the treatment effect estimates based on ANOVA, which help us assess bias. The true treatment effect is zero in this case. The standard deviation for

the estimates is roughly 0.05. It can easily be seen that all estimates have negligible bias, regardless of the matching distance. A similar result is observed for other methods of treatment effect estimation, details are described in the Appendix E.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.01	0.00	0.00	-0.01	-0.00
0.05	-0.01	0.00	0.00	-0.01	-0.00
0.2	-0.01	0.00	0.00	-0.01	-0.00
0.3	-0.01	0.00	0.00	-0.01	-0.00
0.5	-0.01	0.00	0.00	-0.01	-0.00

Table 5.13: Treatment Effect Estimates for the Small Sample from the RHC Study Using the Bayesian-Based Method

Table 5.14 provides power results for detecting a treatment effect of ± 5, again, based on the ANOVA estimation. It appears that matching on a combination of propensity and prognostic score(s) as well as matching without the propensity score gives slightly better power than the competing approaches we tested. This may be a result of cases in which there is not full support on the propensity score, in those cases matching distances which place less emphasis on the propensity score result in matched sets with less extreme structure and a higher effective sample size. Results for other treatment effect estimation methods are given in Appendix F, both resulting in substantially lower power than the ANOVA estimator.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.10	0.14	0.14	0.22	0.11
0.05	0.10	0.14	0.14	0.17	0.11
0.2	0.10	0.14	0.14	0.16	0.11
0.3	0.10	0.14	0.14	0.16	0.11
0.5	0.10	0.14	0.14	0.15	0.11

Table 5.14: Power Estimates for the Small Sample from the RHC Study Using the Bayesian-Based Method

Results for the Hansen-Bowers procedure are quite similar to these results in all respects, including variation of the treatment effect estimates. For this reason, we do not present those results separately here.

Smith and Todd Method

Below, in Tables 5.15 and 5.16 we present results for Type 1 error rates for estimators based on matching within a caliper of 0.4 pooled standard deviations on the propensity score. The first table is for simulation runs that passed the Smith and Todd pre-stratification test, and the second table is for those that did not. For all the methods used to estimate the treatment effect, the Type 1 error rates tend to be within the simulation margin of error of 0.01 when we compare the results from passing and failing the test. They also tend to be within the margin of error of the nominal rate of 0.05. Though these rejection rates were not usually at or below their nominal value for less restrictive calipers, we point out that matching closely on the propensity score is the method most similar to the stratification and matching techniques suggested in Smith and Todd (2005b) for the analysis of the NSW data.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.03	0.06	0.07	0.07	0.04
Mixed Effects	0.06	0.07	0.06	0.06	0.06
OLS	0.01	0.03	0.04	0.04	0.02

Table 5.15: Type 1 Error Rates for the Small Sample from the RHC Study, after Passing the Smith-and-Todd Test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.03	0.05	0.05	0.05	0.04
Mixed Effects	0.05	0.05	0.05	0.05	0.05
OLS	0.01	0.02	0.01	0.01	0.01

Table 5.16: Type 1 Error Rates for the Small Sample from the RHC Study, after Failing the Smith-and-Todd Test

As seen below, the treatment effects are estimated without bias regardless of the method and regardless of passing or failing this balance test. Likely, close matching on the propensity score is the cause of these unbiased results. The standard deviation is about 0.05 for the ANOVA and OLS estimates and about 0.30 for the mixed effects estimates in both tables.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	-0.01	-0.00	-0.00	-0.00	-0.01
Mixed Effects	0.00	0.00	0.00	0.00	0.00
OLS	-0.01	-0.00	-0.00	-0.00	-0.01

Table 5.17: Treatment Effect Estimates for the Small Sample from the RHC Study, after Passing the Smith-and-Todd Test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	-0.01	-0.00	-0.01	-0.01	-0.01
Mixed Effects	-0.01	-0.01	-0.01	-0.01	-0.01
OLS	-0.01	-0.00	-0.00	-0.00	-0.01

Table 5.18: Treatment Effect Estimates for the Small Sample from the RHC Study, after Failing the Smith-and-Todd Test

The power of the test to detect a treatment effect of ± 5 does not appear to be affected by the pre-stratification balance test result. However, an emerging pattern shows that power is highest when the treatment effects are estimated using permutational ANOVA and is lowest for OLS-based estimation.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.10	0.14	0.16	0.16	0.12
Mixed Effects	0.09	0.09	0.09	0.09	0.09
OLS	0.02	0.04	0.04	0.04	0.02

Table 5.19: Power Estimates for the Small Sample from the RHC Study, after Passing the Smith-and-Todd Test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.11	0.16	0.14	0.15	0.11
Mixed Effects	0.08	0.08	0.08	0.08	0.08
OLS	0.01	0.03	0.02	0.03	0.02

Table 5.20: Power Estimates for the Small Sample from the RHC Study, after Failing the Smith-and-Todd Test

The Smith-and-Todd test was introduced by the authors to detect errors of specification in the propensity score. In this simulation study, we only tested its performance in data with errors of aggregation. The test performs reasonably well with a somewhat tight caliper matching, and not so well without it, so it is possible that the performance of the analyses relying on the Smith-and-Todd method to measure balance is confounded with matching closely on the propensity score. For these reason,

future investigations will be helpful at more clearly identifying the test's statistical properties.

5.8.3 Medium Sample Results: RHC Study

In this section, we present medium-sample results for the RHC Study using the Bayesian-based balance test. Recall that the sample size is now 1433. With this larger sample size, differences between balance assessments virtually disappear, and the results associated with the other methods are presented in Appendix F.

As can be seen in table 5.21, the ANOVA estimates result in acceptable false rejection rates, once there is not significant imbalance (starting at line 2). Matching distance does not make a bit of difference for this measure. Treatment effect estimates are all unbiased, though the ones based on matching just on the propensity score have a slightly smaller variance. The variances for the estimates were not drastically different from each other, so we do not present that information here.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.03	0.06	0.06	0.08	0.06
0.05	0.03	0.06	0.06	0.05	0.06
0.2	0.03	0.06	0.06	0.05	0.06
0.3	0.03	0.06	0.06	0.05	0.06
0.5	0.03	0.06	0.06	0.05	0.06

Table 5.21: Type 1 Error Rates for the Medium Sample from the RHC Study, using Bayesian-based Test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.00	-0.00	-0.00	-0.01	-0.00
0.05	-0.00	-0.00	-0.00	-0.00	-0.00
0.2	-0.00	-0.00	-0.00	-0.00	-0.00
0.3	-0.00	-0.00	-0.00	-0.00	-0.00
0.5	-0.00	-0.00	-0.00	-0.00	-0.00

Table 5.22: Treatment Effect Estimates for the Medium Sample from the RHC Study, using Bayesian-based Test. The standard deviation is roughly 0.03.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.09	0.12	0.12	0.17	0.12
0.05	0.09	0.12	0.12	0.13	0.12
0.2	0.09	0.12	0.12	0.13	0.12
0.3	0.09	0.12	0.12	0.12	0.12
0.5	0.09	0.12	0.12	0.12	0.12

Table 5.23: Power Estimates for the Medium Sample from the RHC Study, using Bayesian-based Test

5.8.4 Fixed-Width Stratification

Small Sample - PR Study

In this small study, there was very limited overlap on the estimated propensity scores, likely due to overfitting. For this reason, stratification on the propensity score frequently resulted in a significant reduction in sample size by leaving many subjects without a counterpart. In many cases, this happened to all subjects. In practice, with limited overlap on the estimated propensity score, the researcher should be aware and cautious about such exclusion of subjects, as it may reduce the power and the generalizability of her results. As can be seen from Table 5.24, only the Hansen-Bowers method gave acceptable Type 1 error rates for this combination of conditions. We suspect that the Bayesian approach, despite a somewhat heavy penalty, still failed to make up for the overfitting of the logistic model that occurred due to a reduction of sample size from stratification. It is likely that this overfitting is the cause of excessive false rejections of the hypothesis of no effect for this test. We present treatment effect estimates with their standard deviations, as well as power estimates. While the estimates for the treatment effects are slightly higher (more biased) than with matching, both the variability of the estimates and power to determine a treatment effect are similar to the results we got for matching on the prognostic propensity score with the Hansen-Bowers balance-testing procedure.

	Hansen-Bowers	Bayes
0	0.10	0.08
0.05	0.06	0.08
0.2	0.06	0.08
0.3	0.06	0.08
0.5	0.06	0.08

Table 5.24: Type 1 Error Rates for the PR study using Stratification and ANOVA

	TE Estimate	TE SD	Power
0	0.10	0.16	0.35
0.05	0.07	0.17	0.35
0.2	0.07	0.17	0.35
0.3	0.07	0.17	0.35
0.5	0.07	0.17	0.35

Table 5.25: Results using the Hansen-Bowers Test for the PR study using Stratification and ANOVA

Stratification in the RHC Study

For the modest and medium samples from the RHC study, the statistical properties of the tests we investigated were quite similar to those when matching was used to post-stratify. For this reason, we omit the results of fixed-width stratification for these larger samples.

5.9 Discussion

5.9.1 Sample Size

The small sample results are more telling of the differences between various combinations of procedures that we tested. In the modest and medium samples, the differences between techniques nearly vanish, and the choice of an approach does not tend to make a difference on any of the statistical indicators we investigated; most of them perform adequately. The following discussion of choosing a methodology bears more importance for small samples where overfitting may be an issue.

5.9.2 Subclassification

We investigated optimal full matching on various distances as well as fixed-width stratification on the estimated propensity score. In modest and medium samples,

the subclassification approach did not make a sizeable difference on the results. In a small sample, however, approaches based on matching appeared to be more reliable. This was largely due to the small overlap on the propensity score, which, in the case of stratification, often lead to configurations which excluded a large proportion (and sometimes all) of the subjects. Although in combination with the Hansen-Bowers balance test, stratification still resulted in acceptable rejection rates, estimates and power, a practitioner should be aware of the decrease in sample size and its effects.

Optimal full matching fared well in the small sample case, though the matching distance tended to make a difference. Matching with little emphasis on the propensity score or without the propensity score performed well when combined with the Bayesian-based test for covariate balance, while matching on the combination of the propensity and prognostic propensity scores outperformed other matching distances when combined with the Hansen-Bowers balance measure. In combination with either balance test, matching just on the propensity score was a reliable approach, even in the small-sample setting with incomplete overlap.

5.9.3 Balance Test

Again, in the modest and medium-sample cases, the choice of a balance test made little difference, although the Bayesian-based test is computationally expensive in this case, making it less advantageous. In the small sample, the Hansen-Bowers and the Bayesian-based test outperform the rest, complementing each other, in some cases, on the compatible matching distances.

5.9.4 Estimation of Treatment Effects

Based on the simulation results, the estimator based on permutational ANOVA (Hansen and Bowers 2008) provides for the best combination of unbiasedness, vari-

ance, and power for estimating a common treatment effect. The variance of the estimates did not appear to be related to the choice of the balance test, but rather was sensitive to the estimator. In particular, estimation using mixed effects regression consistently showed much higher variance than the other approaches, and sometimes higher bias as well. Estimates based on covariate adjustment tended to exhibit very low bias, which was advantageous, but unfortunately, that was coupled with having the least power as compared to other estimators, regardless of other data or analysis conditions.

5.9.5 Omitted Results

As mentioned previously, in a related study, we experimented with the response surface for the PR study by imputing the responses for subjects based on their estimated prognostic scores. Because in most cases the results of this change did not have an effect on our conclusions, we do not report them here. Also, to avoid overcrowding the results section, we omit some results which do not add information to our overall conclusions.

In a separate study, working with only the Hansen-Bowers balance testing procedure, we investigated different ways to fit the propensity score in the small PR study - using ordinary logistic regression, the penalized likelihood approach, and the fully Bayesian method. Although we do not include the details of our results, the overall conclusion was that only the fully Bayesian approach to fitting the propensity model results in acceptable statistical indicators, while other methods of fitting it fail to correct for overfitting of the logistic model.

5.10 Conclusions

Overall, we are able to recommend the Hansen-Bowers and the Bayesian-based approaches to balance testing, as they performed reliably in most simulation conditions.

The Hansen-Bowers approach fared especially well in combination with matching on the propensity and prognostic propensity scores in our small sample case, even when stratification was used instead of full matching. On the other hand, false rejection rates for the hypothesis of no effect were too high when this test was combined with matching distances which did not directly incorporate the propensity score or which placed little emphasis on the propensity score, by combining it with several other matching metrics. This method is also reliable and computationally inexpensive in large samples.

The Bayesian-based balance-testing approach performed well in small samples when combined with matching on the propensity score or when matching was done without the propensity score or placing little emphasis on it in the matching metric. In some situations, matching on the propensity score may be less desirable than in others if, for example, there is little support on the estimated propensity score and the researcher believes that this is an indicator of little support on the true propensity score. Also, it may be important to match on several other covariates in addition to the propensity score, decreasing the importance of the propensity scores in the matching metric. Our simulation results show that this kind of matching combined with the Bayesian-based balance result, produces reliable, unbiased, and powerful estimates.

Our simulations demonstrated two limitations of the Bayesian-based approach to

balance testing. First, the test is computationally expensive. In the small-sample case, this did not create a significant disadvantage, but in the medium-sample case, it became a rather large problem. Second, even the heavy penalty that the Bayesian inference implies for the logistic model, cannot make up for data that are sufficiently unsuitable for logistic regression. In our simulations, this was the case with fixed-width stratification on the propensity score in the small-sample case. The stratification frequently reduced the sample size to the point where even our penalized models for the balance tests overfit, giving an overall high rejection rate for the hypothesis of no treatment effect. Another limiting data setting, which did not occur here, might be one in which assignment to treatment is very rare. It is easy to imagine a situation extreme enough in this regard that the Bayesian inference in logistic regression might fail to correct for overfitting.

To conclude, covariate balance, regardless of sample size and mostly regardless of the type of test we used, generally produced false rejection rates near or below nominal, as well as unbiased treatment effect estimates. To the extent that insisting on covariate balance was related to the decrease in sample size by imposing matching or stratification restrictions, it was also connected to increasing the variance of the estimates and decreasing the power in estimating treatment effects.

We sought to design a simulation in which the generated data are realistic. While this approach might have some generalizability advantages over simulating data from known distributions, it also has its limitations. For example, we cannot affirm that the CIA holds nor do we control the nature or extent of treatment effect heterogeneity. To offer general expectations with finite-sample data, we direct the reader to the paper by Hansen (2009), in which the author provides asymptotic analyses that corroborate our results for propensity-matched data.

Appendix A: Small Sample Size Case - PR Study, Hansen-Bowers Test

5.10.1 Discussion

We look at the combination of Hansen-Bowers test with full matching and estimating treatment effects using mixed effects and covariate adjustment using OLS regression. Type 1 error rates are too high in the mixed effects case, and are close to nominal levels for the OLS case when the matching metric is based on the propensity and prognostic propensity scores as well as. Also, matching without the propensity scores gives correct Type 1 error results when the balance p-value is at least 0.2. Treatment effects are estimated with less bias and smaller variance using OLS rather than mixed effects regression.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.10	0.11	0.11	0.10	0.10
0.05	0.10	0.11	0.11	0.10	0.10
0.2	0.10	0.11	0.10	0.10	0.10
0.3	0.10	0.11	0.08	0.07	0.10
0.5	0.10	0.11	0.04	0.04	0.10

Table 5.26: Type 1 Error Rates in the PR study for the Hansen-Bowers test. Treatment effects are estimated using Mixed Effects Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.03	0.07	0.10	0.07	0.05
0.05	0.03	0.07	0.10	0.07	0.05
0.2	0.03	0.07	0.08	0.06	0.05
0.3	0.03	0.07	0.06	0.05	0.05
0.5	0.03	0.07	0.05	0.06	0.05

Table 5.27: Type 1 Error Rates in the PR study for the Hansen-Bowers test. Treatment effects are estimated using OLS

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.46	0.48	0.51	0.51	0.48
0.05	0.46	0.48	0.51	0.51	0.48
0.2	0.46	0.48	0.30	0.38	0.48
0.3	0.46	0.48	0.11	0.13	0.48
0.5	0.35	0.37	-31.99	-26.51	-0.61

Table 5.28: Treatment Effect Estimates in the PR study for the Hansen-Bowers test. Treatment effects are estimated using Mixed Effects Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	1.07	1.13	1.00	0.99	1.03
0.05	1.07	1.13	1.00	0.99	1.03
0.2	1.07	1.13	1.93	1.11	1.03
0.3	1.07	1.13	2.03	1.29	1.04
0.5	3.17	18.49	671.11	681.11	26.13

Table 5.29: Standard Deviations for Treatment Effect Estimates in the PR study for the Hansen-Bowers test. Treatment effects are estimated using Mixed Effects Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.00	0.08	0.13	0.11	0.06
0.05	0.00	0.08	0.13	0.11	0.06
0.2	0.00	0.08	0.10	0.09	0.06
0.3	0.00	0.08	0.05	0.05	0.06
0.5	0.00	0.08	0.01	0.01	0.06

Table 5.30: Treatment Effect Estimates in the PR study for the Hansen-Bowers test. Treatment effects are estimated using OLS Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.25	0.20	0.20	0.19	0.19
0.05	0.25	0.20	0.20	0.19	0.19
0.2	0.25	0.20	0.25	0.24	0.19
0.3	0.25	0.20	0.30	0.30	0.19
0.5	0.25	0.21	0.45	0.49	0.20

Table 5.31: Standard Deviations for Treatment Effect Estimates in the PR study for the Hansen-Bowers test. Treatment effects are estimated using OLS Regression

Appendix B: Small Sample Size Case - PR Study, Bayesian-Based Test, Penalized Likelihood Test

5.10.2 Discussion: Bayesian-Based Test

In combination with mixed effects regression, the Bayesian-based test gives near-nominal Type 1 error rates when matching is done on all data-reducing scores or without the propensity score, once a balance p-value of at least 0.05 has been established. In combination with OLS, on the other hand, all matching configurations except for matching on the propensity and prognostic score result in correct false rejection rates once a p-value for balance of at least 0.05 is present. Consistent with the other results, the OLS method of estimating treatment effect produces less biased and less variable estimates of the treatment effect.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.09	0.10	0.09	0.09	0.10
0.05	0.08	0.10	0.06	0.06	0.09
0.2	0.08	0.10	0.05	0.05	0.09
0.3	0.08	0.09	0.05	0.05	0.09
0.5	0.08	0.09	0.04	0.05	0.08

Table 5.32: Type 1 Error Rates in the PR study for the Bayesian-based test. Treatment effects are estimated using Mixed Effects Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.04	0.08	0.09	0.08	0.05
0.05	0.03	0.08	0.04	0.03	0.05
0.2	0.03	0.07	0.03	0.02	0.05
0.3	0.03	0.07	0.03	0.02	0.05
0.5	0.03	0.07	0.02	0.02	0.04

Table 5.33: Type 1 Error Rates in the PR study for the Bayesian-based test. Treatment effects are estimated using OLS

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.52	0.50	0.54	0.54	0.50
0.05	1.16	0.50	0.68	0.57	0.47
0.2	1.16	1.11	0.55	0.48	1.05
0.3	1.16	1.10	0.52	0.47	1.04
0.5	1.16	1.04	0.49	0.46	0.98

Table 5.34: Treatment Effect Estimates in the PR study for the Bayesian-based test. Treatment effects are estimated using Mixed Effects Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	1.51	1.25	1.04	1.04	1.06
0.05	15.32	1.38	14.38	14.39	1.24
0.2	15.32	14.36	14.39	14.40	14.31
0.3	15.32	14.36	14.39	14.40	14.32
0.5	15.32	14.37	14.39	14.40	14.33

Table 5.35: Standard Deviations for Treatment Effect Estimates in the PR study for the Bayesian-based test. Treatment effects are estimated using Mixed Effects Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.01	0.08	0.13	0.12	0.06
0.05	0.01	0.08	0.03	0.02	0.06
0.2	0.01	0.07	0.00	0.00	0.05
0.3	0.01	0.07	-0.00	-0.00	0.04
0.5	0.01	0.05	-0.01	0.00	0.03

Table 5.36: Treatment Effect Estimates in the PR study for the Bayesian-based test. Treatment effects are estimated using OLS

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.26	0.20	0.21	0.20	0.19
0.05	0.54	0.24	0.44	0.44	0.22
0.2	0.54	0.26	0.45	0.44	0.25
0.3	0.54	0.26	0.45	0.44	0.26
0.5	0.54	0.31	0.46	0.45	0.28

Table 5.37: Standard Deviations for Treatment Effect Estimates in the PR study for the Bayesian-based test. Treatment effects are estimated using OLS

5.10.3 Discussion: Penalized Likelihood-Based Test

In combination with this approach to balance testing, we obtain acceptable Type I error rates by insisting on balance and combining matching on the propensity score with ANOVA, matching on all data scores with mixed effects regression, and matching on propensity or propensity with prognostic propensity with OLS, as well as combining other matching distances with OLS while insisting on a balance p-value of at least 0.3. Estimation of treatment effects is similar in terms of bias and variance for the ANOVA and OLS-based estimators, and the mixed effects estimates, again, are more biased and variable.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.07	0.09	0.18	0.20	0.08
0.05	0.06	0.09	0.14	0.20	0.07
0.2	0.06	0.08	0.09	0.16	0.08
0.3	0.05	0.08	0.07	0.12	0.07
0.5	0.05	0.07	0.05	0.07	0.07

Table 5.38: Type 1 Error Rates in the PR study for the Penalized Likelihood-based test. Treatment effects are estimated using ANOVA.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.09	0.10	0.10	0.10	0.10
0.05	0.09	0.10	0.08	0.09	0.10
0.2	0.08	0.09	0.06	0.08	0.10
0.3	0.08	0.09	0.05	0.07	0.09
0.5	0.08	0.08	0.04	0.05	0.09

[!ht]

Table 5.39: Type 1 Error Rates in the PR study for the Penalized Likelihood-based test. Treatment effects are estimated using Mixed Effects Regression.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.04	0.08	0.10	0.09	0.06
0.05	0.04	0.08	0.08	0.08	0.06
0.2	0.04	0.07	0.06	0.07	0.06
0.3	0.04	0.06	0.04	0.06	0.06
0.5	0.04	0.06	0.03	0.05	0.05

Table 5.40: Type 1 Error Rates in the PR study for the Penalized Likelihood-based test. Treatment effects are estimated using OLS.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.02	0.10	0.15	0.16	0.08
0.05	0.02	0.10	0.13	0.15	0.07
0.2	0.02	0.10	0.08	0.10	0.07
0.3	0.02	0.10	0.07	0.08	0.07
0.5	0.01	0.09	0.05	0.04	0.06

Table 5.41: Treatment Effect Estimates in the PR study for the Penalized Likelihood-based test. Treatment effects are estimated using ANOVA.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.25	0.18	0.12	0.11	0.18
0.05	0.25	0.18	0.14	0.12	0.18
0.2	0.25	0.18	0.17	0.16	0.18
0.3	0.25	0.19	0.18	0.18	0.19
0.5	0.25	0.20	0.21	0.22	0.20

Table 5.42: Standard Deviations for Treatment Effect Estimates in the PR study for the Penalized Likelihood-based test. Treatment effects are estimated using ANOVA.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.54	0.52	0.56	0.56	0.52
0.05	0.95	0.75	0.73	0.74	0.49
0.2	0.84	1.14	0.85	0.99	0.51
0.3	-0.05	0.22	1.29	0.59	0.33
0.5	0.49	0.18	2.10	1.46	-0.46

Table 5.43: Treatment Effect Estimates in the PR study for the Penalized Likelihood-based test. Treatment effects are estimated using Mixed Effects Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	1.58	1.29	1.06	1.06	1.08
0.05	13.46	7.58	7.90	7.56	1.12
0.2	15.04	15.35	15.89	15.47	2.57
0.3	32.05	32.65	28.76	14.38	4.08
0.5	33.26	46.29	38.52	47.07	33.83

Table 5.44: Standard Deviations for Treatment Effect Estimates in the PR study for the Penalized Likelihood-based test. Treatment effects are estimated using Mixed Effects Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.01	0.08	0.13	0.12	0.07
0.05	0.01	0.08	0.11	0.11	0.06
0.2	-0.00	0.07	0.05	0.07	0.06
0.3	0.00	0.06	0.03	0.06	0.05
0.5	0.01	0.06	0.01	0.02	0.04

Table 5.45: Treatment Effect Estimates in the PR study for the Penalized Likelihood-based test. Treatment effects are estimated using OLS.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.26	0.20	0.21	0.20	0.19
0.05	0.29	0.20	0.34	0.27	0.20
0.2	0.39	0.25	0.42	0.34	0.24
0.3	0.42	0.29	0.43	0.38	0.27
0.5	0.44	0.32	0.45	0.48	0.31

Table 5.46: Standard Deviations for Treatment Effect Estimates in the PR study for the Penalized Likelihood-based test. Treatment effects are estimated using OLS.

Appendix C: Small Sample Size Case - PR Study, Smith and Todd Method, Matching within 0.25 caliper

5.10.4 Discussion

It can be seen from these results that passing or failing the Smith and Todd balance assessment with subsequent matching, does not make a clear difference for the Type 1 error rate. However, passing the pre-stratification test does result in less variable estimates for the treatment effect when mixed effects regression is used for estimation. Power to detect a treatment effect does not appear to be affected by the results of the balance test.

5.10.5 Type I Error Rates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.01	0.02	0.01	0.01	0.01
Mixed Effects	0.00	0.00	0.00	0.00	0.00
OLS	0.01	0.01	0.01	0.01	0.01

Table 5.47: Type 1 Error Rates in the PR study after passing the Smith and Todd test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.01	0.02	0.01	0.01	0.01
Mixed Effects	0.00	0.00	0.01	0.01	0.00
OLS	0.01	0.02	0.02	0.02	0.01

Table 5.48: Type 1 Error Rates in the PR study after failing the Smith and Todd test

5.10.6 Treatment Effect Estimates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	-0.00	-0.01	-0.01	-0.01	-0.01
Mixed Effects	56.50	56.49	56.36	56.44	56.35
OLS	0.01	-0.01	0.01	0.00	0.00

Table 5.49: Treatment Effect Estimates in the PR study after passing the Smith and Todd test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.25	0.25	0.25	0.25	0.25
Mixed Effects	320.17	320.17	320.16	320.16	320.16
OLS	0.31	0.34	0.36	0.36	0.31

Table 5.50: Standard Deviation Estimates in the PR study after passing the Smith and Todd test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	-0.00	-0.01	-0.01	-0.00	-0.00
Mixed Effects	-8.02	-7.68	-7.94	-7.81	-8.24
OLS	0.01	-0.00	0.01	0.01	0.01

Table 5.51: Treatment Effect Estimates in the PR study after failing the Smith and Todd test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.25	0.25	0.25	0.25	0.25
Mixed Effects	244.44	244.43	244.45	244.44	244.46
OLS	0.40	0.40	0.43	0.43	0.44

Table 5.52: Standard Deviation Estimates in the PR study after failing the Smith and Todd test

5.10.7 Power Estimates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.11	0.17	0.14	0.14	0.12
Mixed Effects	0.03	0.03	0.03	0.03	0.03
OLS	0.03	0.02	0.03	0.02	0.02

Table 5.53: Power Estimates in the PR study after passing the Smith and Todd test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.14	0.18	0.16	0.15	0.15
Mixed Effects	0.03	0.03	0.03	0.03	0.03
OLS	0.03	0.03	0.03	0.03	0.03

Table 5.54: Power Estimates in the PR study after failing the Smith and Todd test

Appendix D: Small Sample Size Case - PR Study; Kolmogorov-Smirnov Method, within 0.25 caliper

5.10.8 Discussion

Passing or failing this balance test does not appear to affect Type 1 error rates, estimates, or power in estimation of treatment effects.

5.10.9 Type 1 Error Rates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.01	0.02	0.01	0.01	0.01
Mixed Effects	0.00	0.00	0.01	0.01	0.00
OLS	0.00	0.00	0.00	0.00	0.00

Table 5.55: Type 1 Error Rates in the PR study after passing the K-S test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.01	0.02	0.01	0.01	0.01
Mixed Effects	0.00	0.00	0.00	0.00	0.00
OLS	0.01	0.01	0.01	0.01	0.01

Table 5.56: Type 1 Error Rates in the PR study after failing the K-S test

5.10.10 Treatment Effect Estimates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	-0.01	-0.02	-0.01	-0.01	-0.01
Mixed Effects	-2.29	-1.92	-1.83	-1.77	-2.61
OLS	0.01	-0.01	0.01	0.01	0.01

Table 5.57: Treatment Effect Estimates in the PR study after passing the K-S test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.23	0.23	0.23	0.23	0.24
Mixed Effects	271.35	271.36	271.36	271.36	271.34
OLS	0.49	0.48	0.45	0.45	0.49

Table 5.58: Standard Deviation Estimates in the PR study after passing the K-S test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.00	-0.01	-0.00	-0.00	-0.00
Mixed Effects	22.55	22.70	22.37	22.50	22.40
OLS	0.01	-0.00	0.01	0.01	0.00

Table 5.59: Treatment Effect Estimates in the PR study after failing the K-S test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.26	0.26	0.26	0.26	0.26
Mixed Effects	278.03	278.02	278.02	278.02	278.04
OLS	0.31	0.33	0.39	0.39	0.35

Table 5.60: Standard Deviation Estimates in the PR study after failing the K-S test

5.10.11 Power Estimates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.21	0.23	0.22	0.22	0.23
Mixed Effects	0.09	0.09	0.09	0.09	0.09
OLS	0.04	0.04	0.04	0.04	0.04

Table 5.61: Power Estimates in the PR study after passing the K-S test

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
ANOVA	0.22	0.24	0.23	0.23	0.23
Mixed Effects	0.05	0.05	0.05	0.05	0.05
OLS	0.12	0.13	0.12	0.12	0.12

Table 5.62: Power Estimates in the PR study after failing the K-S test

Appendix E: Modest Sample Size Case - 500-Subject Sample from the RHC Study

5.10.12 Discussion

With the exception of matching without the propensity score in some cases, regardless of the post-stratification balance test, we are able to obtain correct Type 1 error rates and unbiased treatment effect estimates. Variance of the estimates is the highest for mixed effects regression, as previously. Power to detect a treatment effect is the highest for ANOVA-based estimation.

5.10.13 Type I Error Rates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.05	0.05	0.06	0.05	0.05
0.05	0.05	0.05	0.06	0.05	0.05
0.2	0.05	0.05	0.06	0.05	0.05
0.3	0.05	0.05	0.06	0.05	0.05
0.5	0.05	0.05	0.06	0.05	0.05

Table 5.63: Estimates for the Type I Error Rates for the Bayesian Procedure in the Modest Sample from the RHC Study Treatment effects are estimated using Mixed Effects Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.01	0.03	0.02	0.02	0.01
0.05	0.01	0.03	0.02	0.02	0.01
0.2	0.01	0.03	0.02	0.02	0.01
0.3	0.01	0.03	0.02	0.02	0.01
0.5	0.01	0.03	0.02	0.02	0.01

Table 5.64: Estimates for the Type I Error Rates for the Bayesian Procedure in the Modest Sample from the RHC Study Treatment effects are estimated using OLS

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.02	0.04	0.05	0.13	0.04
0.05	0.02	0.04	0.05	0.06	0.04
0.2	0.02	0.04	0.04	0.05	0.04
0.3	0.02	0.04	0.05	0.06	0.04
0.5	0.02	0.04	0.05	0.05	0.04

Table 5.65: Estimates for the Type I Error Rates for the Penalized Procedure in the Modest Sample from the RHC Study Treatment effects are estimated using ANOVA

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.05	0.05	0.06	0.05	0.06
0.05	0.05	0.05	0.06	0.05	0.06
0.2	0.05	0.05	0.06	0.05	0.06
0.3	0.05	0.05	0.06	0.05	0.06
0.5	0.05	0.05	0.05	0.05	0.06

Table 5.66: Estimates for the Type I Error Rates for the Penalized Procedure in the Modest Sample from the RHC Study Treatment effects are estimated using Mixed Effects Regression

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.00	0.03	0.02	0.02	0.01
0.05	0.00	0.03	0.02	0.01	0.01
0.2	0.00	0.03	0.02	0.02	0.01
0.3	0.00	0.03	0.02	0.02	0.01
0.5	0.00	0.03	0.02	0.02	0.01

Table 5.67: Estimates for the Type I Error Rates for the Penalized Procedure in the Modest Sample from the RHC Study Treatment effects are estimated using OLS

Treatment Effect Estimates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.01	-0.01	-0.01	-0.01	-0.01
0.05	-0.01	-0.01	-0.01	-0.00	-0.01
0.2	-0.01	-0.01	-0.01	-0.00	-0.01
0.3	-0.01	-0.01	-0.01	-0.00	-0.01
0.5	-0.01	-0.01	-0.01	0.00	-0.01

Table 5.68: Treatment Effect Estimates for the Bayesian Procedure in the Modest Sample from the RHC Study Treatment effects are estimated using Mixed Effects Regression. The standard deviation for the estimates is roughly 0.29

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.01	0.00	-0.00	-0.00	-0.00
0.05	-0.01	0.00	-0.00	-0.00	-0.00
0.2	-0.01	0.00	-0.00	-0.00	-0.00
0.3	-0.01	0.00	-0.00	-0.00	-0.00
0.5	-0.01	0.00	-0.00	-0.00	-0.00

Table 5.69: Treatment Effect Estimates for the Bayesian Procedure in the Modest Sample from the RHC Study Treatment effects are estimated using OLS. The standard deviation for the estimates is roughly 0.05

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.01	0.00	0.00	-0.01	-0.00
0.05	-0.01	0.00	0.00	-0.00	-0.00
0.2	-0.01	0.00	0.00	-0.00	-0.00
0.3	-0.01	0.00	0.00	-0.00	-0.00
0.5	-0.01	0.00	0.00	-0.00	-0.00

Table 5.70: Treatment Effect Estimates for the Penalized Procedure in the Modest Sample from the RHC Study Treatment effects are estimated using ANOVA. The standard deviation for the estimates is roughly 0.05

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.01	-0.00	-0.01	-0.01	-0.01
0.05	-0.01	-0.00	-0.01	0.00	-0.01
0.2	-0.01	-0.00	-0.00	0.01	-0.01
0.3	-0.01	-0.00	-0.00	0.01	-0.01
0.5	-0.01	-0.00	-0.00	0.01	-0.01

Table 5.71: Treatment Effect Estimates for the Penalized Procedure in the Modest Sample from the RHC Study Treatment effects are estimated using Mixed Effects Regression. The standard deviation for the estimates is roughly 0.29

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.01	0.00	-0.00	-0.00	-0.00
0.05	-0.01	0.00	-0.00	0.00	-0.00
0.2	-0.01	0.00	-0.00	0.00	-0.00
0.3	-0.01	0.00	-0.00	0.00	-0.00
0.5	-0.01	0.00	-0.00	-0.00	-0.00

Table 5.72: Treatment Effect Estimates for the Penalized Procedure in the Modest Sample from the RHC Study. Treatment effects are estimated using OLS. The standard deviation for the estimates is roughly 0.05.

Power Estimates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.08	0.08	0.08	0.08	0.08
0.05	0.08	0.08	0.08	0.08	0.08
0.2	0.08	0.08	0.08	0.08	0.08
0.3	0.08	0.08	0.08	0.08	0.08
0.5	0.08	0.08	0.08	0.08	0.08

Table 5.73: Power Estimates for the effect of ± 5 for the Bayesian Procedure in the Modest Sample from the RHC Study. Treatment effects are estimated using Mixed Effects Regression.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.02	0.04	0.03	0.03	0.02
0.05	0.02	0.04	0.03	0.03	0.02
0.2	0.02	0.04	0.03	0.03	0.02
0.3	0.02	0.04	0.03	0.03	0.02
0.5	0.02	0.04	0.03	0.03	0.02

Table 5.74: Power Estimates for the effect of ±5 for the Bayesian Procedure in the Modest Sample from the RHC Study. Treatment effects are estimated using OLS.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.10	0.13	0.14	0.23	0.12
0.05	0.10	0.13	0.14	0.15	0.12
0.2	0.10	0.13	0.14	0.14	0.12
0.3	0.10	0.13	0.14	0.14	0.12
0.5	0.10	0.13	0.14	0.14	0.12

Table 5.75: Power Estimates for the effect of ±5 for the Penalized Procedure in the Modest Sample from the RHC Study. Treatment effects are estimated using ANOVA.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.08	0.08	0.08	0.09	0.09
0.05	0.08	0.08	0.08	0.08	0.09
0.2	0.08	0.08	0.08	0.08	0.09
0.3	0.08	0.08	0.09	0.08	0.09
0.5	0.08	0.08	0.08	0.08	0.09

Table 5.76: Power Estimates for the effect of ±5 for the Penalized Procedure in the Modest Sample from the RHC Study. Treatment effects are estimated using Mixed Effects Regression.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.01	0.04	0.03	0.03	0.02
0.05	0.01	0.04	0.03	0.03	0.02
0.2	0.01	0.04	0.03	0.03	0.02
0.3	0.01	0.04	0.03	0.03	0.02
0.5	0.01	0.04	0.03	0.03	0.02

Table 5.77: Power Estimates for the effect of ±5 for the Penalized Procedure in the Modest Sample from the RHC Study. Treatment effects are estimated using OLS.

Appendix F: Medium Sample Size Case - 1433-Subject Sample from the RHC Study

Similar to the modest sample size case, regardless of the post-stratification balance test, we are able to obtain correct Type 1 error rates and unbiased treatment effect estimates. Variance of the estimates is the highest for mixed effects regression, as previously. Power to detect a treatment effect is the highest for ANOVA-based estimation.

Type I Error Rates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.06	0.06	0.06	0.06	0.06
0.05	0.06	0.06	0.06	0.06	0.06
0.2	0.06	0.06	0.06	0.06	0.06
0.3	0.06	0.06	0.06	0.06	0.06
0.5	0.06	0.06	0.06	0.06	0.06

Table 5.78: Estimates for the Type I Error Rates for the Bayesian Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using Mixed Effects Regression.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.01	0.02	0.02	0.02	0.02
0.05	0.01	0.02	0.02	0.02	0.02
0.2	0.01	0.02	0.02	0.01	0.02
0.3	0.01	0.02	0.02	0.02	0.02
0.5	0.01	0.02	0.02	0.02	0.02

Table 5.79: Estimates for the Type I Error Rates for the Bayesian Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using OLS.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.03	0.06	0.06	0.08	0.06
0.05	0.03	0.06	0.06	0.08	0.06
0.2	0.03	0.06	0.06	0.08	0.06
0.3	0.03	0.06	0.06	0.08	0.06
0.5	0.03	0.06	0.06	0.08	0.06

Table 5.80: Estimates for the Type I Error Rates for the GLM-based Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using ANOVA.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.06	0.06	0.06	0.06	0.06
0.05	0.06	0.06	0.06	0.06	0.06
0.2	0.06	0.06	0.06	0.06	0.06
0.3	0.06	0.06	0.06	0.06	0.06
0.5	0.06	0.06	0.06	0.06	0.06

Table 5.81: Estimates for the Type I Error Rates for the GLM-based Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using Mixed Effects Regression.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.01	0.02	0.02	0.02	0.02
0.05	0.01	0.02	0.02	0.02	0.02
0.2	0.01	0.02	0.02	0.02	0.02
0.3	0.01	0.02	0.02	0.02	0.02
0.5	0.01	0.02	0.02	0.02	0.02

Table 5.82: Estimates for the Type I Error Rates for the GLM-based Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using OLS.

Treatment Effect Estimates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.01	-0.01	-0.01	-0.01	-0.01
0.05	-0.01	-0.01	-0.01	-0.01	-0.01
0.2	-0.01	-0.01	-0.01	-0.01	-0.01
0.3	-0.01	-0.01	-0.01	-0.01	-0.01
0.5	-0.01	-0.01	-0.01	-0.01	-0.01

Table 5.83: Treatment Effect Estimates for the Bayesian Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using Mixed Effects Regression. The standard deviation for the estimates is roughly 0.15.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.00	-0.00	-0.00	0.00	-0.00
0.05	-0.00	-0.00	-0.00	-0.00	-0.00
0.2	-0.00	-0.00	-0.00	-0.00	-0.00
0.3	-0.00	-0.00	-0.00	-0.00	-0.00
0.5	-0.00	-0.00	-0.00	-0.00	-0.00

Table 5.84: Treatment Effect Estimates for the Bayesian Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using OLS. The standard deviation for the estimates is roughly 0.03.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.00	-0.00	-0.00	-0.01	-0.00
0.05	-0.00	-0.00	-0.00	-0.01	-0.00
0.2	-0.00	-0.00	-0.00	-0.01	-0.00
0.3	-0.00	-0.00	-0.00	-0.01	-0.00
0.5	-0.00	-0.00	-0.00	-0.01	-0.00

Table 5.85: Treatment Effect Estimates for the GLM-based Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using ANOVA. The standard deviation for the estimates is roughly 0.03.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.01	-0.01	-0.01	-0.01	-0.01
0.05	-0.01	-0.01	-0.01	-0.01	-0.01
0.2	-0.01	-0.01	-0.01	-0.01	-0.01
0.3	-0.01	-0.01	-0.01	-0.01	-0.01
0.5	-0.01	-0.01	-0.01	-0.01	-0.01

Table 5.86: Treatment Effect Estimates for the GLM-based Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using Mixed Effects Regression. The standard deviation for the estimates is roughly 0.15.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	-0.00	-0.00	-0.00	-0.00	-0.00
0.05	-0.00	-0.00	-0.00	-0.00	-0.00
0.2	-0.00	-0.00	-0.00	-0.00	-0.00
0.3	-0.00	-0.00	-0.00	-0.00	-0.00
0.5	-0.00	-0.00	-0.00	-0.00	-0.00

Table 5.87: Treatment Effect Estimates for the GLM-based Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using OLS. The standard deviation for the estimates is roughly 0.03.

Power Estimates

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.06	0.07	0.07	0.06	0.07
0.05	0.06	0.07	0.07	0.07	0.07
0.2	0.06	0.07	0.07	0.07	0.07
0.3	0.06	0.07	0.07	0.07	0.07
0.5	0.06	0.07	0.07	0.07	0.07

Table 5.88: Power Estimates for the effect of ± 5 for the Bayesian Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using Mixed Effects Regression.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.01	0.03	0.03	0.03	0.03
0.05	0.01	0.03	0.03	0.03	0.03
0.2	0.01	0.03	0.03	0.03	0.03
0.3	0.01	0.03	0.03	0.03	0.03
0.5	0.01	0.03	0.03	0.03	0.03

Table 5.89: Power Estimates for the effect of ±5 for the Bayesian Procedure in the Medium Sample from the RHC Study Treatment effects are estimated using OLS.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.09	0.12	0.12	0.17	0.12
0.05	0.09	0.12	0.12	0.17	0.12
0.2	0.09	0.12	0.12	0.17	0.12
0.3	0.09	0.12	0.12	0.17	0.12
0.5	0.09	0.12	0.12	0.17	0.12

Table 5.90: Power Estimates for the effect of ±5 for the GLM-based Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using ANOVA.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.07	0.07	0.07	0.07	0.07
0.05	0.07	0.07	0.07	0.07	0.07
0.2	0.07	0.07	0.07	0.07	0.07
0.3	0.07	0.07	0.07	0.07	0.07
0.5	0.07	0.07	0.07	0.07	0.07

Table 5.91: Power Estimates for the effect of ±5 for the GLM-based Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using Mixed Effects Regression.

	Prop	Prop and Prog	All Scores	No Prop	Prog Prop
0	0.01	0.03	0.03	0.03	0.03
0.05	0.01	0.03	0.03	0.03	0.03
0.2	0.01	0.03	0.03	0.03	0.03
0.3	0.01	0.03	0.03	0.03	0.03
0.5	0.01	0.03	0.03	0.03	0.03

Table 5.92: Power Estimates for the effect of ±5 for the GLM-based Procedure in the Medium Sample from the RHC Study. Treatment effects are estimated using OLS.

BIBLIOGRAPHY

Bibliography

A. Abadie and G.W. Imbens. Simple and bias-corrected matching estimators for average treatment effects. NBER Technical Working Papers 0283, National Bureau of Economic Research, Inc, 2002. available at http://ideas.repec.org/p/nbr/nberte/0283.html.

A. Agresti. *Categorical data analysis*. John Wiley & Sons, 2002.

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

R.P. Althauser and D.B. Rubin. The computerized construction of a matched sample. *American Journal of Sociology*, (76):325–346, 1971.

Harrell F.E. Alzola, C.F. *An Introduction to S and the Hmisc and Design Libraries*. http://hesweb1.med.virginia.edu/biostat/s/doc/splus.pdf, 2002.

P.C. Austin, P. Grootendorst, and G.M. Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine*, 26(4), 2007.

W.A. Belson. A technique for studying the effects of a television broadcast. *Applied Statistics*, 5 (3):195–202, November 1956.

R.A. Berk. *Regression analysis: A constructive critique*. Sage, 2004.

J. Bowers and B.B. Hansen. RITOOLS, 2006. an add-on package for R.

M. Busso, J.E. DiNardo, and J. McCrary. New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators. *IZA Discussion Paper Series*, Discussion Paper No. 3998, 2009.

J.F. Carlisle, Schilling S.G., J. Zeng, K.S. Cortina, and Y.N. Kleyman. Progress in reading first in a michigan school district: A study of lansing elementary schools. Technical Report 4.1, Michigan Department of Education, 2006. http://www.mireadingfirst.org/report-41-progress-reading-first-michigan-school-district-study-lansing-elementary-schools.

X.H. Chen, A.P. Dempster, and J.S. Liu. Weighted finite population sampling to maximize entropy. *Biometrika*, 81(3):457–469, 1994.

W. G. Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society*, 128:234–266, 1965.

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

W.G. Cochran and D.B. Rubin. Controlling bias in observational studies: A review. *Sankhyā, Series A, Indian Journal of Statistics*, 35:417–446, 1973.

A.F. Connors, T. Speroff, NV Dawson, C. Thomas, FE Harrell, D. Wagner, N. Desbiens, L. Goldman, AW Wu, RM Califf, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. *Journal of the American Medical Association*, 276(11):889–897, 1996a.

G. Connors, Tonigan S., and Miller J. A measure of religious background and behavior for use in behavior change research. *Psychology of Addictive Behaviors*, 10:90–96, 1996b.

J. Cornfield, W. Haenszel, E.C. Hammond, A.M. Lilienfeld, M.B. Shimkin, and E.L. Wynder. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 1959.

D.R. Cox. *The Planning of Experiments*. John Wiley, 1958.

A.P. Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95:407–448, 2000. with discussion.

R. Dehejia. Does matching overcome lalonde's critique of non-experimental estimators? a postscript. *Manuscript*, 2005.

R.H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(443):1053–1062, 1999.

R.H. Dehejia and S. Wahba. Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161, 2002.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Reply to comments on "maximum likelihood from incomplete data via the Em algorithm". *Journal of the Royal Statistical Society, Series B, Methodological*, 39:34–37, 1977.

D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27, 1993.

R.A. Fisher. The mathematics of a lady tasting tea. *The world of mathematics*, 3:1512–1521, 1956.

D.A. Freedman. Statistical models and shoe leather. *Sociological methodology*, pages 291–313, 1991.

D.A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.

M. Frölich. Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, 86(1):77–90, 2004.

J. Galdo, J.A. Smith, D. Black, and IZA Bonn. Bandwidth selection and the estimation of treatment effects with unbalanced data. *Annales dEconomie et Statistique*, forthcoming, 2009.

A. Gelman, A. Jakulin, M. Pittau, and Y. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

J.M. Gore, R.J. Goldberg, D.H. Spodick, J.S. Alpert, and J.E. Dalen. A community-wide assessment of the use of pulmonary artery catheters in patients with acute myocardial infarction. *Chest*, 92 (4):721–727, 1987.

X.S. Gu and P.R. Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420, 1993.

J. Hájek, Z. Šidák, and P.K. Sen. *Theory of rank tests*. Academic Press New York, second edition, 1999.

B.B. Hansen. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618, September 2004.

B.B. Hansen. OPTMATCH: Flexible, optimal matching for observational studies. *R News*, 2007.

B.B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, page To appear, 2008a.

B.B. Hansen. The essential role of balance tests in propensity-matched observational studies: Comments on'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003'by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, 27(12), 2008b.

B.B. Hansen. Propensity score matching to recover latent experiments: diagnostics and asymptotics. Technical Report 486, Statistics Department, University of Michigan, April 2009.

B.B. Hansen and J. Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2):219–235, 2008.

B.B. Hansen and S.O. Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

H. Hansen, J. Rodriguez, B. Hansen, L. Rebhun, and T.P. George. Clinical comparison of faith-based with biopsychosocial treatment for male drug abusers in puerto rico. *Manuscript*, 2004.

F.E. Harrell. *Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression*. Springer, 2001.

J.J. Heckman. Dummy endogenous variables in a simultaneous equation system. *Econometrica: Journal of the Econometric Society*, pages 931–959, 1978.

J.J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

J.J. Heckman and V.J. Hotz. Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association*, pages 862–874, 1989.

J.J. Heckman, H. Ichimura, and P.E. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4): 605–654, October 1997.

J.J. Heckman, H. Ichimura, and P.E. Todd. Matching as an Econometric Evaluation Estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

G. Heinze and M. Schemper. A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16), 2002.

J.L. Hill, J.P. Reiter, and E. Zanutto. A comparison of experimental and observational data analyses. In Andrew Gelman and Xiao Li Meng, editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, chapter 5, pages 49–60. Wiley, 2004.

K. Hirano and G.W. Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3):259–278, 2001.

K. Hirano, G.W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

P. W. Holland. Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81:945–970, 1986.

C.A. Hosman, B.B. Hansen, and P.W. Holland. The sensitivity of linear regression coefficients's confidence limits to the omission of a confounder. Technical Report 482, Statistics Department, University of Michigan, April 2009.

K. Imai. Do get-out-the-vote calls reduce turnout? the importance of statistical methods for field experiments. *American Political Science Review*, 99(2):283–300, May 2005.

K. Imai, G. King, and E.A. Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):481–502, 2008.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, pages 453–461, 1946.

J.D.Y. Kang and J.L. Schafer. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4): 523, 2007.

Y. Kleyman and B. Hansen. Deconfounding Small Quasi-Experiments Using Propensity Scores and Other Dimension Reduction Techniques. *Manuscript*, 2008.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76:604–620, 1986.

Wang-Sheng Lee. Propensity score matching and variations on the balancing test. *Manuscript*, 2008.

E. L. Lehmann. Model specification: The views of Fisher and Neyman, and later developments. *Statistical Science*, 5:160–168, 1990.

D.Y. Lin, B.M. Psaty, and R.A. Kronmal. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 54(3):948–963, 1998.

J.K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 2004.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:463–480, 1990. reprint. Transl. by Dabrowska and Speed.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669, 1995.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge U.P., 2000.

P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, and A.R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, 1996.

S.W. Raudenbush and A.S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications Inc, 2002.

J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.

P.R. Rosenbaum. Dropping out of high school in the united states: An observational study. *J. Ed. Statist.*, 11:207–224, 1986.

P.R. Rosenbaum. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society*, 53:597– 610, 1991.

P.R. Rosenbaum. Choice as an alternative to control in observational studies (with discussion). *Statistical Science*, 14(3):259–304, 1999.

P.R. Rosenbaum. *Observational Studies*. Springer-Verlag, second edition, 2002a.

P.R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002b.

P.R. Rosenbaum. Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association*, 97(457):183–192, 2002c.

P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

P.R. Rosenbaum and D.B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524, 1984.

P.R. Rosenbaum and D.B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39:33–38, 1985.

D.B. Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29:185–203, 1973a.

D.B. Rubin. Matching to remove bias in observational studies (corr: V30 p728). *Biometrics*, 29: 159–183, 1973b.

D.B. Rubin. Estimating the causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psych.*, 66:688–701, 1974.

D.B. Rubin. Assignment to treatment group on the basis of a covariate (corr: V3 p384). *Journal of Educational Statistics*, 2:1–26, 1977.

D.B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6:34–58, 1978.

D.B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328, 1979.

D.B. Rubin. Bias reduction using Mahalanobis-metric matching (corr: V36 p752). *Biometrics*, 36: 293–298, 1980.

D.B. Rubin. Comments on "statistics and causal inference". *Journal of the American Statistical Association*, 81:961–962, 1986.

D.B. Rubin. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25:279–292, 1990.

D.B. Rubin. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, 47:1213–1234, 1991.

D.B. Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127:757–763, 1997.

D.B. Rubin and N. Thomas. Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52:249–64, 1996.

D.B. Rubin and N. Thomas. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585, 2000.

J.S. Sekhon. Alternative balance metrics for bias reduction in matching methods for causal inference. *Survey Research Center, University of California, Berkeley*, 2007.

S.J. Senn. Testing for baseline balance in clinical trials. *Statistics in Medicine*, 13(17):1715–26, 1994.

B. Sianesi. An evaluation of the swedish system of active labour market programmes in the 1990s. IFS Working Papers W02/01, Institute for Fiscal Studies, January 2002. URL `http://ideas.repec.org/p/ifs/ifsewp/02-01.html`.

J.A. Smith and P.E. Todd. Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review*, pages 112–118, 2001.

J.A. Smith and P.E. Todd. Does matching overcome lalonde's critique of nonexperimental methods. *Journal of Econometrics*, 125(1–2):305–353, 2005a.

J.A. Smith and P.E. Todd. Rejoinder. *Journal of Econometrics*, 125(1-2):365–375, 2005b.

Z. Zhao. Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence. *The Review of Economics and Statistics*, 86(1):91–107, February 2004.

M.M. Zion, J. Balkin, D. Rosenmann, U. Goldbourt, H. Reicher-Reiss, E. Kaplinsky, and S. Behar. Use of pulmonary artery catheters in patients with acute myocardial infarction. Analysis of experience in 5,841 patients in the SPRINT Registry. SPRINT Study Group. *Chest*, 98(6): 1331–1335, 1990.