2017

Balancing Multiple Goals In Observational Study Design

Samuel Pimentel
University of Pennsylvania, sdbpimentel@gmail.com

Balancing Multiple Goals In Observational Study Design

Abstract
This thesis unites three papers discussing new strategies for matched pair designs using observational data, developed to balance the demands of various disparate design goals. The first chapter introduces a new matching algorithm for large-scale treated-control comparisons when many categorical covariates are present. The algorithm balances covariates and their interactions in a prioritized manner by solving a combinatorial optimization problem, and guarantees computational efficiency through the use of a sparse network representation. The second chapter defines a class of variables called prods which can be ignored when matching in order to strictly attenuate unmeasured bias, if it is present. These variables can be difficult to identify with confidence, so a multiple-control-group strategy is proposed in which investigators match once on all variables, and once ignoring prods; the two treated-control comparisons together give stronger evidence about treatment effects than either one individually. The final paper considers a new version of Fisher's classical lack-of-fit test for regression models, appropriate for data that lack replicated observations. The test uses matched pairs formed by optimal nonbipartite matching as near-replicates, and the model fit is used is used in constructing the matching distance in order to focus attention on variables that are predictive in the null model.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Statistics

First Advisor
Paul Rosenbaum

Keywords
Causal Inference, Combinatorial optimization, Lack of fit, Matching, Network flow optimization, Unmeasured bias

Subject Categories
Statistics and Probability

BALANCING MULTIPLE GOALS IN OBSERVATIONAL STUDY DESIGN

Samuel D. Pimentel

A DISSERTATION

in

Statistics

For the Graduate Group in
Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Paul Rosenbaum, Robert G. Putzel Professor; Professor of Statistics

Graduate Group Chairperson

_____

Catherine Schrand, Celia Z. Moh Professor; Professor of Accounting

Dissertation Committee:

Dylan Small, Class of 1965 Wharton Professor of Statistics

Jeffrey Silber, Nancy Abramson Wolfson Endowed Chair in Health Services Research, Children's Hospital of Philadelphia; Professor of Health Care Management

Abba Krieger, Robert Steinberg Professor; Professor of Statistics, Operations, Information, and Decisions, and Marketing

BALANCING MULTIPLE GOALS IN OBSERVATIONAL STUDY DESIGN

Acknowledgments

I would like to thank my advisor Paul Rosenbaum for his substantial involvement and influence in every stage of my development as a student and scholar. Besides introducing me to a wide range of interesting research problems and helping me week-by-week to break them into smaller, achievable tasks, Paul showed incredible patience through the dry periods of my research, gave heartening encouragement and praise at every step, and immediately provided wise and thoughtful advice on any aspect of my academic life whenever I asked. Paul's perspective and example has been and will surely continuing to be a guiding light in all my academic thought and writing.

Dylan Small also spent countless hours with me during my Ph.D., both as an instructor in foundational courses that underpinned my later research and as an advisor and co-author in more recent years. His creative insights on challenging research problems and his tremendous example of effective collaborative work have been an inspiration.

I am also indebted to the other members of my dissertation committee. Jeffrey Silber introuduced me to rich datasets and compelling questions in modern medical research, for willingly became an early adopter of my statistical software, and boosted my confidence in my ability to engage with researchers outside of Statistics. Abba

Krieger provided a patient listening ear and an illuminating perspective on my work as it developed, and he also imparted invaluable support and advice as I prepared for and went on the job market.

I was fortunate to work with several other supportive collaborators — notably Rachel Kelz, Luke Keele, and Guy Grossman — who introduced me to new areas of application for my methods, invested many hours in our joint projects, and helped me develop confidence as an independent collaborator. The members of the Center for Causal Inference at Penn also provided a stimulating intellectual community, helping me workshop my own ideas and bringing me up to speed on important work across our field.

The NDSEG program of the Department of Defense and the Penn Family Grant program both provided me with generous funding during my Ph.D., and I greatly appreciate the support of those programs and their staff.

Maggie Saia and Gidget Murray of Wharton Doctoral Programs cheerfully guided me through the administrative process of obtaining my degree and provided me with many fulfilling opportunities to give service to the University outside of my department.

I owe much to the faculty of the Statistics Department, particularly Mark Low, Mike Steele, Ed George, Larry Brown, Linda Zhao, and Andreas Buja, and to its staff, particularly Adam Greenberg, Noelle Felipe, Carol Reich, Sarin Sieng, and Tanya Winder. All made themselves approachable from my first day at Penn and provided much-needed advice and support in diverse aspects of my life as a graduate student.

I am also very grateful for the friendship of my fellow graduate students in the Statistics Department during our shared journey. My cohort-mates Zijian Guo, Kwonsang Lee, Tengyuan Liang, Dan McCarthy, Peichao Peng, and Xin Lu Tan have pro-

vided a wonderful community, both rejoicing in my successes with me and helping me work through more challenging times. Several former students of the department have also played key roles as mentors to me, including Jose Zubizarreta, Frank Yoon, Mike Baiocchi, and Adam Kapelner.

Finally, I am very grateful for the love and support of my immediate and extended family. Above all, my wife Maren and my sons Caleb and Jeremy have been my biggest supporters, and have demonstrated the highest levels of patience and personal sacrifice in helping me complete my degree. Thank you for everything.

ABSTRACT

BALANCING MULTIPLE GOALS IN OBSERVATIONAL STUDY DESIGN

Samuel D. Pimentel

Paul Rosenbaum

This thesis unites three papers discussing new strategies for matched pair designs using observational data, developed to balance the demands of various disparate design goals. The first chapter introduces a new matching algorithm for large-scale treated-control comparisons when many categorical covariates are present. The algorithm balances covariates and their interactions in a prioritized manner by solving a combinatorial optimization problem, and guarantees computational efficiency through the use of a sparse network representation. The second chapter defines a class of variables called prods which can be ignored when matching in order to strictly attenuate unmeasured bias, if it is present. These variables can be difficult to identify with confidence, so a multiple-control-group strategy is proposed in which investigators match once on all variables, and once ignoring prods; the two treated-control comparisons together give stronger evidence about treatment effects than either one individually. The final paper considers a new version of Fisher's classical lack-of-fit test for regression models, appropriate for data that lack replicated observations. The test

uses matched pairs formed by optimal nonbipartite matching as near-replicates, and the model fit is used is used in constructing the matching distance in order to focus attention on variables that are predictive in the null model.

Contents

List of Tables

List of Figures

1 Introduction

This thesis is based on three papers, all of which address the need to balance several possibly disparate goals in forming matched pair designs using observational data. The first paper considers a large-scale comparison of the health outcomes of patients treated by new surgeons and those of patients treated by experienced surgeons, using data from Medicare claims. Matching new surgeon patients to similar experienced surgeon patients in this observational study provides a special analytic and computational challenge, because of the volume of data but also because of the large number of complex categorical covariates measured for the patients. A new network flow algorithm is presented for matching in this setting, incorporating novel balancing constraints that remove pre-treatment group differences on categorical covariates and their interactions in order of scientific priority. Furthermore, the algorithm represents large observational studies as sparse network flow problem, allowing matches of unprecedented size to be constructed efficiently. In the surgical study, the algorithm produced very desirable levels of balance on interactions of many nominal covariates. This project is joint work with Rachel Kelz, Jeffrey Silber, and Paul Rosenbaum, and was published in 2015 in Volume 110, Issue 510 of the *Journal of the American Statistical Association*. It was produced with support by Grant SBS 1260782 from the MMS Program of the US National Science Foundation, Grant AG032963 from the US

National Institute of Aging, and by Fellowship FA9550-11-C-0028 from the Department of Defense, Army Research Office, National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, 32 CFR 168a4.

The second paper addresses an important question, frequently arising in practice, about which variables to use for matching in observational studies. Randomized trials balance all covariates, observed and unobserved; matching analyses generally try to balance all observed covariates (since it cannot directly balance unobserved covariates). However, informal arguments in the applied medical literature claim that matched analyses would admit less bias due to unobserved covariates if certain observed variables were left unbalanced. This work formalizes that argument and prove that a certain class of variables, called "prods" to receive treatment, can be left unbalanced in the match to strictly lower the degree of unmeasured bias. In practice it is difficult to identify these variables, since they are defined by uncheckable conditions. It is suggested the result is most useful in the computerized construction of a second control group, where the investigator can see more in available data without necessarily believing the required conditions. One of the two control groups controls for the possibly irrelevant observed covariate, the other control group either leaves it uncontrolled or forces separation; therefore, the investigator views one situation from two angles under different assumptions. A pair of sensitivity analyses for the two control groups is coordinated by a weighted Holm or recycling procedure built around the possibility of slight attenuation of bias in one control group. Issues are illustrated using an observational study of the possible effects of cigarette smoking as a cause of increased homocysteine levels, a risk factor for cardiovascular disease. This is joint work with Dylan Small and Paul Rosenbaum, and was published in 2015 in Volume 111, Issue 515 of the *Journal of the American Statistical Association*. It was conducted with support by the Measurement, Methodology, and Statistics Pro-

gram of the National Science Foundation and by Fellowship FA9550-11-C-0028 from the Department of Defense, Army Research Office, National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, 32 CFR 168a4.

The final paper considers an alternative use of matching, not for formation of comparison groups in an observational study but for use in testing the fit of a linear regression model. Fisher's classical lack-of-fit test uses perfectly replicated observations to produce a second estimate of the model standard error and conduct a goodness-of-fit test, but in practice observational datasets rarely contain perfectly replicated observations. A new test is presented, which identifies near-replicates for use in Fisher's testing framework via optimal nonbipartite matching. In particular, a distance is defined for use in the matching algorithm that focuses on predictors important in the original model, betting that model failures involve variables important in the original fit. The test is shown to be exact, despite its use of the original fitted model, and to have reasonable power even when the true set of predictors is hidden within a large collection of spurious ones. This is joint work with Dylan Small and Paul Rosenbaum, and is forthcoming in *Technometrics*, and was conducted with support by National Science Foundation Grant SES-1260782 from the Measurement, Methodology and Statistics Program of the NSF and by Fellowship FA9550-11-C-0028 from the Department of Defense, Army Research Office, National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, 32 CFR 168a4.

2 Large, Sparse Optimal Matching with Refined Covariate Balance

2.1 Introduction: Matching within natural blocks

2.1.1 What are natural blocks?

In observational studies of treatment effects, we often wish to compare treated and control subjects from the same natural block. Familiar examples of natural blocks are twins, siblings, surgical patients in the same hospital, or students in the same school. Important unmeasured covariates may be more similar within a natural block than between blocks: the genes of siblings; the nursing staff and intensive care unit in the same hospital; the teaching staff and socioeconomic conditions within the same school.

There can be a tension between the desire to compare treated and control individuals within natural blocks and the desire to compare treated and control groups with similar distributions of measured covariates. In our study in §3.1 comparing new and experienced surgeons, there are 1252 natural blocks of a new and experienced surgeon performing similar types of surgery working in the same hospital. Additionally there are many categories of measured covariates, including 176 surgical procedures,

ultimately nearly 2.9 million categories defined by measured covariates. With many categories, it is difficult if not impossible to find similar patients inside the same natural block.

Attempts to balance many covariates by pairing individuals who are nearly identical almost invariably fail because nearly identical people do not exist. This is illustrated in Zubizarreta et al. (2011, Table 6; 2014, §2.4) where close individual pairs are not available but covariate balance is attainable. Matching for a scalar propensity score can balance many covariates such as age or gender, but this approach can perform poorly with sparse nominal covariates having many categories, for instance the 176 surgical procedures and their interactions with comorbidities. Like randomization, matching on propensity scores balances covariates stochastically with the aid of the law of large numbers, whereas a nominal covariate with many categories may have small sample sizes in most categories.

Our algorithm pairs patients within a natural block, trying to pick individual pairs that are close on covariates. There is a limit to what can be achieved by finding individually close pairs on many variables, so a separate effort is made to balance distributions of covariates when individuals within a pair may differ. The approach comes as close as possible to balance for a sequence of nested nominal variables, starting with the 176 surgical procedures, gradually subdividing these 176 categories to finally reach nearly 2.9 million categories involving comorbidities and admission source, obtaining the best possible balance at each successive stage of the subdivision. This new objective, "refined covariate balance," is defined in §2.4.4, where it is proved in Theorem 6 that our new network optimization algorithm yields a minimum distance match subject to the constraint of refined covariate balance. This new approach is made practical by exploiting network sparsity.

2.1.2 Natural blocks and network sparsity

Optimal matching in observational studies (Rosenbaum, 1989; Hansen, 2007) is often implemented using network optimization, a collection of mathematical and computational techniques originally developed to solve problems in operations research; see the review of network optimization in §2.4.3. A network is a set of nodes together with a set of directed edges or ordered pairs of nodes. Think of the nodes as subjects and the edges as candidate pairings of two subjects. A network with N nodes might have N^2 edges with loops or $N(N-1)$ edges if with no loops; that is, it might have $O(N^2)$ edges as $N \to \infty$ and in this case the network is said to be dense. A network is said to be sparse if the number of edges is $O(N)$ rather than $O(N^2)$. Matching within natural blocks, such as within hospital-surgeon-pairs, drastically restricts the number of permitted pairings of patients, resulting in a sparse network. The time and space required for optimization is much greater in dense than in sparse networks (e.g., Korte and Vygen 2008, Theorem 9.17).

Typical uses of optimal matching in observational studies do not exploit sparsity, in part because a network defined by measured covariates without natural blocks is likely to be dense. A program such as Hansen's (2007) `optmatch` package in `R` can match thousands of individuals at once in a dense network. In current practice, if a problem has many more than thousands of individuals, then it is divided into smaller problems each consisting of thousands of individuals by matching exactly for several important covariates. This strategy often works well for measured covariates. However, with natural blocks, there may be relatively few choices within blocks, so more of the work needs to be done through balancing covariate distributions. By working with a network that is naturally sparse because of natural blocks, we are able to match hundreds of thousands of individuals at once, thereby making much more effective use of balancing techniques.

2.1.3 Outline: an example; a new objective; a new algorithm; the benefits of sparsity

The surgical example is discussed in §3.1 and §2.5. The general problem is described informally in §2.2 and developed precisely in §2.4. All new results and methods are contained in §2.4. Notation is introduced in §2.4.1, key concepts such as refined balance are defined in §2.4.2, and existing literature on network optimization is briefly reviewed in §2.4.3. The matching network for refined balance is defined in §2.4.4. The main theorem in §2.4.5 says that a minimum cost flow in the network defined in §2.4.4 is the closest possible match that exhibits refined balance while respecting the natural blocks. Sparsity is discussed in §2.4.7. The discussion in §2.6 considers how the proposed methods might be applied in other contexts.

For discussion of matching, see Baiocchi et al. (2012), Hansen and Klopfer (2006); Hansen (2007), Heller et al. (2009), Lu et al. (2011), Rosenbaum (1989, 2010), Rosenbaum and Rubin (1985), Stuart (2010), Yang et al. (2012), and Zubizarreta et al. (2011, 2014). For recent applications of optimal matching, see Silber et al. (2013) and Neuman et al. (2014).

2.2 Abstract problem; intuition behind its solution; other applications

2.2.1 The abstract problem: refined balance in a sparse match

In a sparse matching problem, each treated subject has a short list of potential controls. When there are natural blocks, this short list consists of controls from the same block; however, sparse networks arise or can be produced in other ways; see §2.6.2. As the sample size increases, the length of the list of potential controls for

each given treated subject does not increase. As you add more and more families or schools or hospitals or zip codes to the study, you have more and more subjects to match, but individual families or schools or hospitals or zip codes do not become larger. If the number of blocks increases in constant proportion to the increase in total sample size, then block effects are not consistently estimable without assumptions about their form (Kiefer and Wolfowitz, 1956, p. 888); however, it is possible to match within blocks.

In addition to picking for each treated subject a control from the short list of candidates, the matching must balance many observed covariates. We would be satisfied if the balance on observed covariates after matching were similar to the balance on observed covariates in a completely randomized experiment, but this may not be possible in an observational study. Randomization also balances unmeasured covariates whereas matching for observed covariates cannot be expected to do this.

Because the list of candidate controls for a given treated subject is short, it is rarely possible to find a control on the short list who is identical to the treated subject with respect to many covariates. So the matching algorithm tolerates a mismatch in one pair providing it can counterbalance that mismatch in another pair. If it is necessary to match a treated male to a control female in one block, then a treated female will be matched to a control male in another block, so the final treated and control groups have exactly the same number of males and the same number of females. Exact counterbalancing is called "fine balance"; see Rosenbaum et al. (2007). Fine balance means that the marginal distribution of a categorical covariate is exactly the same in treated and control groups, and in the surgical example the 176 surgical procedures are finely balanced. Counterbalancing is a familiar strategy in experimental design, for example in Latin square designs or crossover designs. Sometimes exact fine balance is not achievable: for instance, it is not possible in the surgical example to exactly

balance all 2.9 million categories of patients. "Near fine balance" means that the marginal distributions of a categorical covariate in matched samples are "as close as possible" to fine balance given the data available; see Yang et al. (2012). In defining near fine balance, one may define "as close as possible" in various ways, but one natural and familiar measure is the total variation distance, the sum of the absolute treated-minus-control differences in category percents. See Arratia et al. (1990, §3) for several attractive equivalent definitions of the total variation distance. If the matched treated group is 51% male and the matched control group is 49% male, then the total variation distance in gender is $|0.51 - 0.49| + |0.49 - 0.51| = 0.04$ reflecting the 2% mismatch for males plus the corresponding 2% mismatch for females. One form of near fine matching minimizes the total variation distance in matched samples, and it achieves exact fine balance whenever this is achievable.

Refined balance is an extension of fine or near-fine balance. One defines a sequence of nested nominal variables, $\nu_1, \ldots, \nu_K$, so ν_{k+1} subdivides ν_k. Refined balance comes as close as possible to fine balance for ν_1, and among all matches that do that, it comes as close as possible to fine balance for ν_2, and so on. In the surgical example, ν_1 consists of the 176 surgical procedures and these are finely balanced, ν_2 interacts the 176 surgical procedures with two types of hospital to make 352 categories for which the minimum total variation distance is 0.001 or one tenth of 1%, ..., and v_K for $K = 6$ has 2.9 million categories. Among all matched samples that exhibit refined covariate balance, the algorithm finds pairings from the short lists to minimize the total covariate distance within pairs.

2.2.2 Intuition behind the solution

In §2.4, the matching problem is represented by a network or directed graph. For each category of each of the nested nominal variables, ν_k, the network has two routes

to a match. One route is free of charge, and a pair can take this route if it leaves this category balanced. The other route has a large toll or penalty, and a pair can take this route without balancing the category but must pay the penalty. The penalty for ν_1 is much larger than for ν_2, and so on. The objective function is the sum of all of these penalties plus the sum of the within-pair covariate distances. The penalization of certain paths is developed in detail in §2.4.4 and it involves a parameter Υ. Network optimization minimizes this penalized objective function. If the penalties are both sufficiently large and sufficiently different for ν_k and ν_{k+1}, then they override all other considerations, producing refined balance. Among all matches that minimize the penalties, the optimal match minimizes the sum of the covariate distances. In the example, among matches that are equally good in terms of refined covariate balance, the algorithm tried to pair individuals with similar ages and estimated risks of death, two variables that were not explicitly balanced. Section 2.4 states the algorithm precisely and proves that it works.

Refined balance and sparsity are separate ideas that work well together. In a sparse network, it is difficult to find close individual pairs, and more of the work must be done by covariate balancing; hence, the attraction of refined balance for sparse problems. Conversely, balancing of rare categories is easier in very large problems, and computations for large problems require less computer time and storage if the problem is sparse; hence the attraction of sparsity for refined balance. Sparsity is discussed in §2.4.7.

2.3 Patient outcomes achieved by new and experienced surgeons

2.3.1 Background

Are the patient outcomes of newly trained surgeons comparable to the outcomes of experienced surgeons performing the same types of surgery at the same hospitals? If the typical patient of the typical new surgeon were instead treated by an experienced surgeon, would the patient's outcomes be different? The data describe patients in Medicare in six states between 2004 and 2007 who had Medicare Part B, were not in a Medicare HMO, and had surgery performed at a hospital rather than on an outpatient basis at an ambulatory surgical center. Here, we look at 6260 patients of 1252 new surgeons and 6260 patients of 1252 experienced surgeons at the same hospitals, 5 patients per surgeon.

Surgical skill varies from surgeon to surgeon. Are the worst surgeons also the new surgeons? A typical hospital might have one new surgeon and a group of experienced surgeons. We expect that the performance of individual new surgeons will be more variable, more extreme, than the average performance of a group of experienced surgeons, simply because averages are more stable than individuals. Surgeons specialize, focusing on particular types of surgery, and the 30-day mortality rate following, say, elective orthopedic surgery is much lower than for some types of cancer surgery. These considerations, together with desire for a simple, transparent study design, led us to pair each new surgeon with an experienced surgeon performing similar types of surgery at the same hospital.

New surgeons gradually become experienced surgeons. As they become more experienced, they perform more surgery. Most of the population of patients of new surgeons are the patients of the most experienced of the new surgeons, but we are most

interested in new surgeons when they are starting out, when most of their experience is from surgical training. For these reasons, we decided to give equal weight to each young surgeon, rather than weighting surgeons by the number of operations they performed. We considered only new and experienced surgeons who had performed at least five operations in our data. We sampled at random five surgical patients of each new surgeon as the treated group. For many newer new surgeons, five patients was a large part of the portion of the overlap of their surgical practice with our data. Our analysis describes the typical patient of the typical new surgeon, not the typical patient of new surgeons as a group, the latter being weighted towards the most experienced new surgeons.

2.3.2 Matching the patients of new and experienced surgeons within the same hospital

Surgical data are characterized by quite a bit of detail, much of it recorded in nominal variables. Using ICD-9 codes, we distinguish 176 surgical procedures (listed in Table 2.1 as Procedure). In addition, we distinguish among 498 hospitals, whose performance varies for reasons unrelated to surgical performance. Patients often have existing medical problems, called comorbidities, besides those treated by the current surgery, such as congestive heart failure (CHF) or chronic obstructive pulmonary disease (COPD), and these may increase the risk of death following surgery. We distinguish hospitals with many new surgeons or few new surgeons (Hospital Group). Patients are matched within surgeon pairs within the same hospital.

Table 2.1 lists covariates that structure the match, and additional covariates appear in Table 2.2. Table 2.1 includes notation that will be defined in §2.4. In the rows of Table 2.1, there are 15 nominal covariates, making 176×2^{14} or about 2.9 million categories of patients. The columns of Table 2.1 define $K = 6$ nominal

covariates, $\nu_1, \ldots, \nu_6$, where ν_1 is simply the $L_1 = 176$ procedures, ν_2 is the 176 procedures crossed with Hospital Group with $L_2 = 176 \times 2 = 352$ categories, ν_3 is the 176 procedures crossed with Hospital Group, male, ER-admission, and Transfer-admission with $L_3 = 176 \times 2^4 = 2816$ categories, ..., and ν_6 crosses all 15 covariates with $176 \times 2^{14} \doteq 2.9$ million categories.

Ideally, the number of patients of new surgeons in each of 2.9 million categories would equal the number of patients of experienced surgeons. That was not quite possible while always also matching patients within the 498 hospitals. Subject to that requirement of matching within hospitals, the match minimized imbalance in a sense to be defined in a moment, and minimized the sum of a covariate distance over 6260 patient pairs.

A nominal covariate with L_k levels yields an $L_k \times 2$ contingency table with two columns for the patients of new and experienced surgeons. In the matched sample, each column contains a total of 6260 patients distributed among L_k categories or rows. How different are the distributions in the two columns? Write $\beta_{k\ell}$ for the difference in counts of ν_k in row ℓ of the table; then $0 = \sum_{\ell=1}^{L_k} \beta_{k\ell}$ and $\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$ is proportional to a standard measure of the difference between two discrete probability distributions, namely the total variation distance. Now, $\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$ could be as small as 0 if the distributions were identical or as large as $2 \times 6260 = 12520$ if they do not overlap. To equalize the two distributions, one would need to switch the categories for $\sum_{\ell=1}^{L_k} |\beta_{k\ell}| /2$ controls or the percentage $(100/6260) \sum_{\ell=1}^{L_k} |\beta_{k\ell}| /2$.

The lower portion of Table 2.1 shows the total imbalance in the six nominal covariates, $\nu_1, \ldots, \nu_6$. For procedures, ν_1, the imbalance was 0, so the distribution of the 176 procedures is identical in the new and experienced groups. The imbalance for ν_1 is as small as possible. For ν_2, the imbalance was 6, meaning that there was a total excess of 3 in some of the rows of the 2×352 table and a total deficit of 3

Table 2.1: The $K = 6$ nominal variables ν_k that were balanced as closely as possible by the matching algorithm, where ν_1 consists of $L_1 = 176$ surgical procedures, and ν_6 is the interaction of 176 surgical procedures with 14 binary covariates, making $L_6 = 176 \times 2^{11}$ categories, or about 2.9 million categories. An $\times$ indicates that the row variable contributes to nominal variable ν_k. The algorithm minimized the total imbalance $\sum_{\ell=1}^{L_{k'}} |\beta_{k'\ell}|$ for $\nu_{k'}$ among all matches that minimized $\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$ for ν_k for $k < k'$. The balance obtained by matching is much better than the best balance obtained in 10,000 simulated randomized experiments with the same marginal totals.

		Nested nominal covariate, ν_k $k = 1, \ldots, 6$					
Covariate	Levels	1	2	3	4	5	6
Procedure	176	×	×	×	×	×	×
Hospital Group	2		×	×	×	×	×
Male	2			×	×	×	×
ER-admit	2			×	×	×	×
Transfer	2			×	×	×	×
Paraplegia	2				×	×	×
Stroke	2				×	×	×
PPF	2				×	×	×
CC	2					×	×
CHF	2					×	×
Dementia	2					×	×
Renal	2					×	×
Liver	2						×
Past A	2						×
Past MI	2						×
# Categories		176	176×2	176×2^4	176×2^7	176×2^{11}	176×2^{14}
L_k		$= 176$	$= 352$	$= 2,816$	$= 22,528$	$= 360,448$	$= 2,883,584$
Imbalance							
$\sum_{\ell=1}^{L_k} \lvert\beta_{k\ell}\rvert$		0	12	52	176	664	1242
% of maximum		0.0%	0.1%	0.4%	1.4%	5.3%	9.9%
Independence χ^2		0.0	4.9	43.3	142.3	588.9	1158.7
Balance in 10,000 simulated randomized experiments with the same margins							
Simulated χ^2 statistics for independence							
Mean χ^2		174.9	302.9	767.5	1062.0	1946.0	2814.0
Minimum χ^2		117.0	226.5	645.7	933.6	1777.0	2645.0
Simulated Total Imbalance $\sum_{\ell=1}^{L_k} \lvert\beta_{k\ell}\rvert$							
Mean $\sum_{\ell=1}^{L_k} \lvert\beta_{k\ell}\rvert$		768	1051	1749	2086	3010	3812
Min. $\sum_{\ell=1}^{L_k} \lvert\beta_{k\ell}\rvert$		540	814	1500	1826	2752	3578

in some other rows. The imbalance for ν_2 is as small as possible among matches that minimize the imbalance in ν_1. And so on. For ν_6, the total absolute imbalance is 1242 for $2 \times 6260 = 12520$ patients in 2.9 million categories, or about 10% of the maximum imbalance. The imbalance for ν_6 is as small as possible subject to minimizing the imbalance in ν_1, ..., ν_5 and matching within surgeon pairs. In addition to producing a small imbalance in $\nu_1, \ldots, \nu_6$, the matching algorithm certifies that the imbalance attained is the smallest possible imbalance when matching new and experienced surgeon patients within the same hospital; that is, there is no point in trying to achieve a smaller imbalance.

The balance described in the previous paragraph is much better than randomization would produce. We computed the usual χ^2-statistic for independence in each of the six $2 \times L_k$ contingency tables. We created 10,000 simulated randomized experiments by simple random sampling without replacement of 6260 patients from the 12520 patients, so row and column margins of the $2 \times L_k$ are unchanged, and computed 10,000 independence χ^2-statistics and imbalances $\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$; see the bottom of Table 2.1. For ν_6 with 2.9 million categories, the actual matched sample had an imbalance of 1242 and χ^2 of 1158.7, and that was much better balance than the best of 10,000 simulated randomized experiments with an imbalance of 3578 and χ^2 of 2645.0.

Subject to the constraints of matching within hospital and minimizing imbalance $\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$ in Table 2.1, the algorithm minimized the total over 6260 patient pairs of a covariate distance within pairs. Table 2.2 looks at the imbalance on the individual matching variables, including age and the risk score, neither of which is in Table 2.1.

Do new surgeons treat the easiest patients? Apparently not. In Table 2.2, before matching, the patients of new surgeons are much more likely to have entered through the emergency room, have higher estimated risks of death based on comorbidities,

Table 2.2: Covariate imbalance before and after matching. The table compares new surgeons to experienced surgeons, before and after matching, in term of covariate means, standardized differences in means as a fraction of the standard deviation before matching, and two-sample P-values. New = new surgeon, Ex-B = experienced surgeon, before matching, Ex-A = experienced surgeon, after matching. Standardized differences above 1/10th of a standard deviation are in **bold**.

	Covariate Mean			Standardized Difference		2-sample P-value	
Covariate	New	Ex-B	Ex-A	Before	After	Before	After
Sample size	6,260	123,846	6,260				
Age	77.883	76.992	77.926	**0.116**	-0.005	0.000	0.617
Male	0.345	0.358	0.346	-0.027	-0.003	0.038	0.880
ER-admit	0.538	0.323	0.537	**0.444**	0.003	0.000	0.886
Transfer	0.008	0.008	0.007	0.000	0.013	1.000	0.532
Risk	0.042	0.030	0.040	**0.214**	0.031	0.000	0.237
CHF	0.149	0.123	0.143	0.076	0.019	0.000	0.311
Liver	0.043	0.036	0.038	0.035	0.026	0.005	0.161
Cancer	0.164	0.175	0.164	-0.029	0.001	0.030	0.981
Past A	0.170	0.171	0.161	-0.002	0.024	0.880	0.178
Diabetes	0.189	0.197	0.199	-0.019	-0.024	0.145	0.198
Renal	0.069	0.058	0.064	0.046	0.020	0.000	0.282
COPD	0.167	0.147	0.160	0.055	0.019	0.000	0.298
CC	0.028	0.028	0.022	-0.006	0.031	0.691	0.075
Dementia	0.101	0.065	0.093	**0.131**	0.032	0.000	0.103
Paraplegia	0.019	0.011	0.015	0.063	0.031	0.000	0.114
Past MI	0.058	0.054	0.051	0.015	0.031	0.265	0.083
PPF	0.023	0.020	0.021	0.023	0.015	0.069	0.429
Stroke	0.068	0.058	0.063	0.041	0.019	0.001	0.312

are more likely to have dementia, and tend to be older. These differences are largely absent after matching. New surgeons are treating a challenging and vulnerable group of patients. In §2.5, we ask: How do outcomes compare for new and experienced surgeons when experienced surgeons treat equally challenging patients?

2.4 A network algorithm for large, sparse optimal matching with refined balance

2.4.1 Notation: acceptable 1-to-m match; covariate imbalance $\beta_{k\ell}$

There are T treated subjects, $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$, and $C \geq T$ potential controls, $\mathcal{C} = \{\kappa_1, \ldots, \kappa_C\}$, with $\emptyset = \mathcal{T} \cap \mathcal{C}$. In §2.3.2, $\mathcal{T}$ contains patients of new surgeons and $\mathcal{C}$ contains patients of experienced surgeons. Write $|\mathcal{S}|$ for the number of elements in a finite set $\mathcal{S}$, so that $T = |\mathcal{T}|$. There were $T = 6260$ patients of new surgeons to be matched and $C = 123846$ candidate control patients of experienced surgeons. Treated subject $\tau_t \in \mathcal{T}$ has observed covariate $\mathbf{x}_{\tau t}$ and potential control $\kappa_c \in \mathcal{C}$ has covariate $\mathbf{x}_{\kappa c}$.

There is a subset of acceptable pairings, $\mathcal{A} \subseteq \mathcal{T} \times \mathcal{C}$, such that (τ_t, κ_c) is an acceptable pairing if and only if $(\tau_t, \kappa_c) \in \mathcal{A}$. In §2.3.2, we had previously paired a new and an experienced surgeon at the same hospital performing similar procedures, and the acceptable pairings $\mathcal{A}$ are only of patients of these paired new and experienced surgeons at the same hospital; that is, $(\tau_t, \kappa_c) \in \mathcal{A}$ if and only if τ_t is a patient of a new surgeon and κ_c is a patient of the experienced surgeon with whom this new surgeon is paired. In §2.3.2, $|\mathcal{A}| = 819230 < 7.75 \times 10^8 = T \times C = |\mathcal{T} \times \mathcal{C}|$.

For each $(\tau_t, \kappa_c) \in \mathcal{A}$ there is a distance δ_{tc} between $\mathbf{x}_{\tau t}$ and $\mathbf{x}_{\kappa c}$, $\delta_{tc} = \delta\left(\mathbf{x}_{\tau t}, \mathbf{x}_{\kappa c}\right)$, with $0 \leq \delta_{tc} < \infty$. We would like to pair individuals who are close on covariates. In

§2.3.2, $\delta_{tc} = \delta(\mathbf{x}_{\tau t}, \mathbf{x}_{\kappa c})$ was a robust, rank-based Mahalanobis distance (Rosenbaum, 2010, §8) based on age, sex, emergency admission, transfer admission, risk score and clusters of procedures. There is competition for controls, so κ_c may be the closest control to both τ_t and $\tau_{t'}$, and an optimal matching will minimize the total distance for matched individuals subject to various constraints on the balance of covariates.

There are K nested nominal variables $\nu_k(\cdot)$, $k = 1, \ldots, K$; that is, $\nu_k(\cdot)$ is a function that assigns one of L_k values in $\mathcal{K}_k = \{\lambda_{k1}, \ldots, \lambda_{k,L_k}\}$ to each subject in $\mathcal{T} \cup \mathcal{C}$, or $\nu_k : \mathcal{T} \cup \mathcal{C} \to \mathcal{K}_k$. In §2.3.2 and Table 2.1, there were $K = 6$ nominal variables. Importantly, ν_{k+1} refines or subdivides ν_k. In other words, these K variables are nested in the sense that all individuals who are the same on ν_{k+1} are the same on ν_k; that is, formally, if $\iota \in \mathcal{T} \cup \mathcal{C}$ with $\nu_{k+1}(\iota) = \lambda_{k+1,\ell}$ and $\iota' \in \mathcal{T} \cup \mathcal{C}$ with $\nu_{k+1}(\iota') = \lambda_{k+1,\ell}$, then $\nu_k(\iota) = \nu_k(\iota')$. Variable $\nu_1(\cdot)$ is the coarsest and most important variable and $\nu_K(\cdot)$ is the finest and least important variable. Expressed informally, the algorithm will do everything possible to balance $\nu_1(\cdot)$ as closely as possible, whereas it will merely do what it can to balance $\nu_K(\cdot)$.

Definition 1 *Acceptable 1-to-m match: An acceptable 1-to-m match is a subset $\mathcal{M} \subseteq \mathcal{A}$ such that every $\tau_t \in \mathcal{T}$ appears in exactly m pairs $(\tau_t, \kappa_c) \in \mathcal{M}$ and every $\kappa_c \in \mathcal{C}$ appears in at most one pair $(\tau_t, \kappa_c) \in \mathcal{M}$.*

If $\mathcal{A} = \mathcal{T} \times \mathcal{C}$, then an acceptable 1-to-m match exists whenever $C \geq mT$. If $\mathcal{A} \subset \mathcal{T} \times \mathcal{C}$, then an 1-to-$m$ acceptable match may not exist even when $C \geq mT$. The algorithm finds an acceptable 1-to-m match if one exists; otherwise it reports that no such match exists. The conditions required for the existence of an acceptable match are stated in a famous theorem in graph theory, Hall's theorem; see Diestel (2010, Theorem 2.1.2, p. 38); however, the algorithm determines whether a match exists.

In addition to having an acceptable match with $\mathcal{M} \subseteq \mathcal{A}$ with a small total distance $\sum_{(\tau_t, \kappa_c) \in \mathcal{M}} \delta_{tc}$, we also want to balance the K nominal variables, emphasizing $\nu_k(\cdot)$

over $\nu_{k+1}(\cdot)$. Write $d_{k\ell}$ for the number of treated individuals τ_t falling in category ℓ of the k^{th} nominal variable $\nu_k(\cdot)$, so $d_{k\ell} = |\{\tau_t \in \mathcal{T} : \nu_k(\tau_t) = \lambda_{k\ell}\}|$. Ideally, an acceptable 1-to-m match $\mathcal{M}$ would have $m \times d_{k\ell}$ matched controls falling in category ℓ of the k^{th} nominal variable $\nu_k(\cdot)$, so the distributions of $\nu_k(\cdot)$ would be identical in matched treated and control groups; however, typically, this is not possible for larger k. That is, ideally $|\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_k(\kappa_c) = \lambda_{k\ell}\}|$ would equal $m \times d_{k\ell}$ for every k and ℓ. Because the K variables are nested, an imbalance in $\nu_k(\cdot)$ is necessarily also an imbalance in $\nu_{k+1}(\cdot)$.

The imbalance $\beta_{k\ell}$ in the ℓ^{th} category of the k^{th} nominal variable is a signed integer that is m times the number of treated subjects τ_t in $\mathcal{M}$ with level $\lambda_{k\ell}$ of the k^{th} nominal variable minus the number of controls κ_c in $\mathcal{M}$ with level $\lambda_{k\ell}$, that is,

$$\beta_{k\ell} = m \times d_{k\ell} - |\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_k(\kappa_c) = \lambda_{k\ell}\}|. \tag{2.1}$$

In (2.1), $\beta_{k\ell}$ depends upon the match $\mathcal{M}$ through $|\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_k(\kappa_c) = \lambda_{k\ell}\}|$, but the notation does not indicate the dependence explicitly; that is, some matches $\mathcal{M}$ exhibit better covariate balance than do others. Here $\beta_{k\ell} > 0$ signifies that we wanted more controls at level ℓ of nominal variable $\nu_k(\cdot)$, and $\beta_{k\ell} < 0$ signifies that we wanted fewer. By the definition of an acceptable 1-to-m match, for each k, the total of the signed imbalances is zero, $0 = \sum_{\ell=1}^{L_k} \beta_{k\ell}$ (i.e., everyone has to go somewhere), but the total of the absolute imbalances $\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$ measures the degree to which matched treated and control subjects have differing distributions of nominal variable $\nu_k(\cdot)$. In fact, $(mT)^{-1} \sum_{\ell=1}^{L_k} |\beta_{k\ell}|$ is the total variation distance between the distribution of $\nu_k(\cdot)$ in matched treated and control groups. In Table 2.1, $\sum_{\ell=1}^{L_3} |\beta_{3\ell}| = 52$. In some sense or other, we would like to pick an acceptable 1-to-m match such that each of the $\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$ is as small as possible and the within-pair distance $\sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc}$ is as small as possible.

The k^{th} nested nominal variable is said to satisfy "fine balance" if $\beta_{k\ell} = 0$ for $\ell = 1, \ldots, L_k$, so $\nu_k(\cdot)$ has the same distribution in matched treated and control groups; see Rosenbaum et al. (2007). Because the K nominal variables are nested, nominal variable $\nu_k(\cdot)$ is finely balanced whenever $\nu_{k+1}(\cdot)$ is finely balanced.

The k^{th} nested nominal variable is said to satisfy "near fine balance" if match $\mathcal{M}$ minimizes $\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$ among all acceptable 1-to-m matches; see Yang et al. (2012). Because the K nominal variables are nested, $\sum_{\ell=1}^{L_{k+1}} |\beta_{k+1,\ell}| \geq \sum_{\ell'=1}^{L_k} |\beta_{k\ell'}|$ for each k, as is seen in Table 2.1 where $\sum_{\ell=1}^{L_1} |\beta_{1\ell}| = 0 \leq 12 = \sum_{\ell=1}^{L_2} |\beta_{2\ell}| \leq 52 \leq \ldots \leq 1242 = \sum_{\ell=1}^{L_6} |\beta_{6\ell}|$.

2.4.2 Two key definitions: What is an optimal refined acceptable 1-to-m match $\mathcal{M}$?

Where fine and near fine balance refer to a single nominal variable, "refined balance" refers to a nested sequence of nominal variables, such as $\nu_k(\cdot)$, $k = 1, \ldots, K$, as in Table 2.1. Stated informally, each of the k levels is as balanced as possible, but level k has priority over level $k+1$. Write $\mathfrak{M}$ for the set of all acceptable 1-to-m matches $\mathcal{M}$. Each element $\mathcal{M} \in \mathfrak{M}$ is one possible match. Each such match $\mathcal{M} \in \mathfrak{M}$ has values for $\beta_{k\ell}$ in (2.1) and a value for the total distance within matched sets, $\sum_{(\tau_t, \kappa_c) \in \mathcal{M}} \delta_{tc}$. The two definitions that follow define a "best" choice of $\mathcal{M} \in \mathfrak{M}$.

Definition 2 (Refined balance): *An acceptable 1-to-m match $\mathcal{M} \in \mathfrak{M}$ has refined balance if: (1) $\sum_{\ell=1}^{L_1} |\beta_{1\ell}|$ is minimized among all acceptable 1-to-m matches $\mathcal{M}' \in \mathfrak{M}$, and (2) among acceptable 1-to-m matches that satisfy (1), $\mathcal{M}$ minimizes $\sum_{\ell=1}^{L_2} |\beta_{2\ell}|$, ..., (k) among acceptable 1-to-m matches that satisfy (k-1), $\mathcal{M}$ minimizes $\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$, ..., (K) among acceptable 1-to-m matches that satisfy (K-1), $\mathcal{M}$ minimizes $\sum_{\ell=1}^{L_K} |\beta_{K\ell}|$.*

For example, in Table 2.1, 52 is the minimum possible value of $\sum_{\ell=1}^{L_3} |\beta_{3\ell}|$ among all acceptable 1-to-1 matches with $\sum_{\ell=1}^{L_1} |\beta_{1\ell}| = 0$ and $\sum_{\ell=1}^{L_2} |\beta_{2\ell}| \leq 12$.

Definition 3 (Optimal refined balance): *An acceptable 1-to-m match $\mathcal{M} \in \mathfrak{M}$ with refined balance is optimal if it minimizes the total distance within pairs, $\sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc}$, among all acceptable 1-to-m matches $\mathcal{M} \in \mathfrak{M}$ with refined balance.*

The goal is to find an optimal refined acceptable 1-to-m match $\mathcal{M}$ if one exists and otherwise determine that the problem is infeasible in that no such match exists.

2.4.3 Review of minimum cost flow in a network

The minimum cost flow problem is a standard combinatorial optimization problem with origins in operations research; see Bertsekas (1991), Cook et al. (1998), and Korte et al. (2008). This problem is a special type of integer program which, unlike most integer programs, can be solved with a worst-case time bound that is a polynomial in the size of the problem; that is, large problems can be solved quickly. A standard way to "solve" a combinatorial optimization problem is to show that it is equivalent to an appropriate minimum cost flow problem and to solve this equivalent problem. (In `R`, a good solver for minimum cost flow problems can be obtained as follows. Hansen's `optmatch` package calls Fortran code `RELAXIV` created by Bertsekas et al. (1994) which solves minimum cost flow problems. Loading `optmatch` makes `RELAXIV` accessible in `R` and callable by imitating Hansen's calls with different calling parameters. Documentation and code for `RELAXIV` are on Bertsekas' web page at MIT.)

Metaphorically, objects are supplied and demanded at locations called nodes and are shipped among nodes along edges connecting pairs of nodes, and the goal is to minimize the total shipping cost while meeting demands subject to capacity constraints.

Objects cannot be cut in half (e.g., TVs cannot be cut in half for shipping) so the solution must ship integer rather than fractional objects. Companies like FedEx solve minimum cost flow problems in a literal rather than metaphorical sense. Optimal matching problems are commonly reexpressed as minimum cost flow problems. We find an optimal refined acceptable 1-to-m match $\mathcal{M}$ by solving an equivalent minimum cost flow problem.

A network is a set of nodes, $\mathcal{N}$, a set of edges $\mathcal{E}$ consisting of ordered pairs of nodes, $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$, so each $e \in \mathcal{E}$ is of the form $e = (n, n')$ where n, $n' \in \mathcal{N}$. One draws a network with a point for each node $n \in \mathcal{N}$ and an arrow connecting pairs of nodes for which there is an edge $e = (n, n') \in \mathcal{E}$, where the tail of the arrow is at n and the point of the arrow is at n'. See Figure 1, where the arrowheads are omitted to limit clutter, but edges that are not horizontal point down and horizontal edges point from right to left. Our network is acyclic or without cycles, so we may speak of the early part of the network — the upper part in Figure 1 — or the late part of the network — the lower part in Figure 1.

Each edge $e \in \mathcal{E}$ has a nonnegative, possibly infinite, integer capacity, $\texttt{cap}\,(e)$ with $0 \leq \texttt{cap}\,(e) \leq \infty$, and a nonnegative real cost, $\texttt{cost}\,(e)$ with $0 \leq \texttt{cost}\,(e) < \infty$. That is, e can carry up to $\texttt{cap}\,(e)$ units of flow and each unit costs $\texttt{cost}\,(e)$ to transport over e. Each node $n \in \mathcal{N}$ has a finite integer demand, $\texttt{demand}\,(n)$ with $-\infty < \texttt{demand}\,(n) < \infty$. Node n absorbs $\texttt{demand}\,(n)$ units of flow and passes the rest on, and $\texttt{demand}\,(n) < 0$ means n creates an excess of $-\texttt{demand}\,(n)$ units of flow (e.g., manufactures $-\texttt{demand}\,(n)$ TVs). A feasible flow f is a function that assigns a nonnegative integer $f\,(e)$ to each edge $e = (n, n') \in \mathcal{E}$, such that: (i) the flow is within the capacity limits, $0 \leq f\,(e) \leq \texttt{cap}\,(e)$ for each $e \in \mathcal{E}$, and the demand at

each node $n \in \mathcal{N}$ is met,

$$\sum_{n':(n',n)\in\mathcal{E}} f\left\{(n',n)\right\} - \sum_{n'':\left(n,n''\right)\in\mathcal{E}} f\left\{\left(n,n''\right)\right\} = \texttt{demand}\left(n\right) \text{ for each } n \in \mathcal{N}. \quad (2.2)$$

The first sum in (2.2) is the total flow into n from neighboring nodes n' with $(n',n) \in \mathcal{E}$, while the second sum is the total flow out from n to neighboring nodes n'' with $\left(n,n''\right) \in \mathcal{E}$, so the equation (2.2) says that node n absorbs $\texttt{demand}\left(n\right)$ units of flow. A feasible flow may or may not exist. The total cost of a feasible flow is $\sum_{e\in\mathcal{E}} f\left(e\right) \texttt{cost}\left(e\right)$. An optimal feasible flow is any feasible flow that minimizes the total cost. The problem of finding a minimum cost flow in a network has several fast widely available solutions.

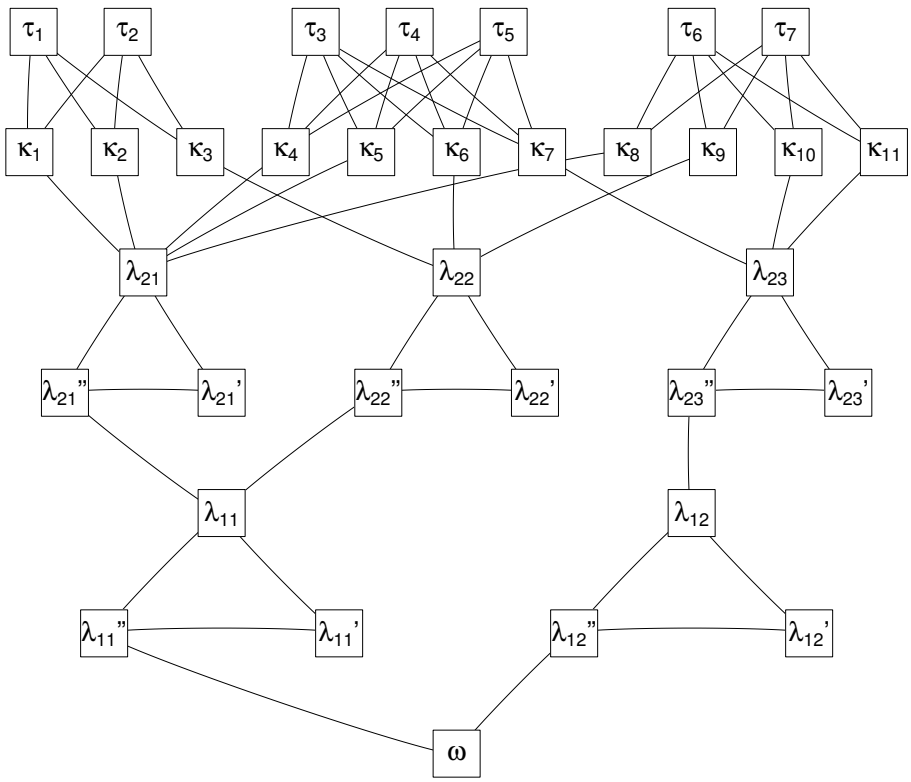
From a practical point of view, finding a minimum cost flow in a network may be regarded by users as a standard mathematical computation, not unlike finding the inverse of a matrix. The user specifies the network and is given a minimum cost flow, as the user of matrix inversion software specifies a matrix and is given its inverse. Not all matrices have inverses, and not all networks have feasible flows, and in both cases competent software announces that the impossible has been requested. A network is dense if $O\left(|\mathcal{E}|\right) = |\mathcal{N}|^2$.

2.4.4 The network for optimal refined acceptable 1-to-m matching

The network involves a penalization parameter, $\Upsilon > 1$. Penalization will increase the cost of a flow when that flow is behaving in a way we wish to avoid. In §2.4.5, it will be shown that if Υ is large enough, then the solution to a certain minimum cost flow problem yields an optimal refined acceptable 1-to-m matching.

The nodes, $\mathcal{N}$, of the network contain the treated subjects $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$,

Figure 2.1: A small network for refined covariate balance with treated subject $\tau_1, \ldots, \tau_7$, potential controls $\kappa_1, \ldots, \kappa_{11}$, two balance layers $\lambda_{1\ell}$ and $\lambda_{2\ell}$, and the sink ω.



the potential controls, $\mathcal{C} = \{\kappa_1, \ldots, \kappa_C\}$, and an additional node ω called a sink. Also the nodes contain all of the possible values of the K nested nominal variables, $\mathcal{K}_k = \{\lambda_{k1}, \ldots, \lambda_{k,L_k}\}$, $k = 1, \ldots, K$. Additionally, the nodes contain a primed copy of values of the nested nominal variables, $\mathcal{K}_k^{'} = \left\{\lambda_{k1}^{'}, \ldots, \lambda_{k,L_k}^{'}\right\}$, $k = 1, \ldots, K$, and a double primed copy of all of the possible values of the nominal variables, $\mathcal{K}_k^{''} = \left\{\lambda_{k1}^{''}, \ldots, \lambda_{k,L_k}^{''}\right\}$, $k = 1, \ldots, K$. That is, the nodes are $\mathcal{N} = \mathcal{T} \cup \mathcal{C} \cup \{\omega\} \cup \bigcup_{k=1}^{K} \mathcal{K}_k \cup \bigcup_{k=1}^{K} \mathcal{K}_k^{'} \cup \bigcup_{k=1}^{K} \mathcal{K}_k^{''}$.

If $(\tau_t, \kappa_c) \in \mathcal{A} \subseteq \mathcal{T} \times \mathcal{C}$ is an acceptable pairing in the sense of §2.4.1, then (τ_t, κ_c) is an edge of the network, $(\tau_t, \kappa_c) \in \mathcal{E}$ with capacity $\texttt{cap}\{(\tau_t, \kappa_c)\} = 1$ and cost $\texttt{cost}\{(\tau_t, \kappa_c)\} = \delta_{tc}$, where δ_{tc} is the covariate distance between τ_t and κ_c introduced in §2.4.1. There is an edge $(\kappa_c, \lambda_{K\ell}) \in \mathcal{E}$ connecting each potential control κ_c to the category $\lambda_{K\ell}$ of the last, most refined nominal variable $\nu_K(\cdot)$ that contains this control; moreover, this edge has capacity 1 and zero cost, $\texttt{cap}\{(\kappa_c, \lambda_{K\ell})\} = 1$ and $\texttt{cost}\{(\kappa_c, \lambda_{K\ell})\} = 0$.

Every category $k\ell$ of every nominal variable $\nu_k(\cdot)$ appears as a small triangle in $\mathcal{E}$ involving $\lambda_{k\ell}$, $\lambda_{k\ell}^{'}$ and $\lambda_{k\ell}^{''}$. These triangles play an important role: each one makes an effort to reduce a corresponding $|\beta_{k\ell}|$ in (2.1), recognizing that it may not be possible to achieve $|\beta_{k\ell}| = 0$. Every node $\lambda_{k\ell}$ is connected to both $\lambda_{k\ell}^{'}$ and $\lambda_{k\ell}^{''}$, so $\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right) \in \mathcal{E}$ and $\left(\lambda_{k\ell}, \lambda_{k\ell}^{''}\right) \in \mathcal{E}$, and $\lambda_{k\ell}^{'}$ is connected to $\lambda_{k\ell}^{''}$ so $\left(\lambda_{k\ell}^{'}, \lambda_{k\ell}^{''}\right) \in \mathcal{E}$ for all k, ℓ; that is, $\lambda_{k\ell}$, $\lambda_{k\ell}^{'}$ and $\lambda_{k\ell}^{''}$ form a triangle. There is, therefore, a direct path from $\lambda_{k\ell}$ to $\lambda_{k\ell}^{''}$ and an indirect path from $\lambda_{k\ell}$ to $\lambda_{k\ell}^{''}$ that passes through $\lambda_{k\ell}^{'}$. As discussed in §2.4.1, we would like to have $m \times d_{k\ell}$ controls in category $\lambda_{k\ell}$ as this would make $\beta_{k\ell} = 0$ in (2.1); however, this may not be possible. The direct path $\left(\lambda_{k\ell}, \lambda_{k\ell}^{''}\right)$ has $\texttt{cap}\left\{\left(\lambda_{k\ell}, \lambda_{k\ell}^{''}\right)\right\} = m \times d_{k\ell}$ and cost $\texttt{cost}\left\{\left(\lambda_{k\ell}, \lambda_{k\ell}^{''}\right)\right\} = 0$, so that up to $m \times d_{k\ell}$ units of flow can move directly from $\lambda_{k\ell}$ to $\lambda_{k\ell}^{''}$ for free, without cost. The indirect path is penalized as we would prefer to use it as little as possible. The

edge $\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right)$ has infinite capacity, $\texttt{cap}\left\{\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right)\right\} = \infty$, and severely penalized cost of $\texttt{cost}\left\{\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right)\right\} = \Upsilon^{K-k+1}$. The last leg of the triangle has infinite capacity and zero cost, $\texttt{cap}\left\{\left(\lambda_{k\ell}^{'}, \lambda_{k\ell}^{''}\right)\right\} = \infty$ and $\texttt{cost}\left\{\left(\lambda_{k\ell}^{'}, \lambda_{k\ell}^{''}\right)\right\} = 0$. Notice that the penalty for $\nu_1(\cdot)$ is Υ^K but this gradually declines to penalty Υ for $\nu_K(\cdot)$. Because the coarse, most important $\nu_1(\cdot)$ is after the fine, less important $\nu_K(\cdot)$, the penalties in triangles increase from Υ for $\nu_K(\cdot)$ to Υ^K for $\nu_1(\cdot)$ as we move from start to the end of the network. Informally, this says that a one-patient imbalance in $v_k(\cdot)$ is worse than a one-patient imbalance in $v_{k+1}(\cdot)$.

The end $\lambda_{k\ell}^{''}$ of a triangle at level k is connected to the beginning $\lambda_{k-1,\ell'}$ of the coarser category $k-1, \ell'$ that contains category $k\ell$. This edge $\left(\lambda_{k\ell}^{''}, \lambda_{k-1,\ell'}\right)$ to a coarsened category has infinite capacity and zero cost, $\texttt{cap}\left\{\left(\lambda_{k\ell}^{''}, \lambda_{k-1,\ell'}\right)\right\} = \infty$ and $\texttt{cost}\left\{\left(\lambda_{k\ell}^{''}, \lambda_{k-1,\ell'}\right)\right\} = 0$. Finally, there is an edge from $\lambda_{1\ell}^{''}$ to the sink ω for each ℓ with infinite capacity and zero cost, $\texttt{cap}\left\{\left(\lambda_{1\ell}^{''}, \omega\right)\right\} = \infty$ and $\texttt{cost}\left\{\left(\lambda_{1\ell}^{''}, \omega\right)\right\} = 0$.

For each $\tau_t \in \mathcal{T}$, $\texttt{demand}(\tau_t) = -m$. The sink has $\texttt{demand}(\omega) = m|\mathcal{T}|$. All other nodes have $\texttt{demand}(n) = 0$. In words, each treated node issues m units of flow, all nodes between the treated nodes and the sink pass on all the flow they receive, and the sink ω collects all mT units of flow issued by the T treated units.

An important property of a feasible flow f in this network is that control node $\kappa_c \in \mathcal{C} \subset \mathcal{N}$ may receive either zero or one unit of flow, because $0 \leq f(\kappa_c, \lambda_{K\ell}) \leq \texttt{cap}\{(\kappa_c, \lambda_{K\ell})\} = 1$, and if $f(\kappa_c, \lambda_{K\ell}) = 1$ then there is only one possible sequence of $\lambda_{k\ell}^{''}$'s along which that unit of flow can pass to the sink ω. For brevity, the network defined in this section will be called "the network $(\mathcal{N}, \mathcal{E})$," omitting explicit reference to the capacities, costs and demands that are also part of its definition.

2.4.5 Main result: A minimum cost flow yields an optimal refined match

Lemma 4 says that the match we seek exists if and only if the minimum cost flow problem is feasible. Proofs are in Appendix A.1.

Lemma 4 *There is a feasible flow f for the network $(\mathcal{N}, \mathcal{E})$ if and only if there is an acceptable 1-to-m match $\mathcal{M}$. In particular, $\mathcal{M} = \{(\tau_t, \kappa_c) \in \mathcal{A} : f\{(\tau_t, \kappa_c)\} = 1\}$.*

Lemma 5 relates total cost to matching quantities, namely total covariate distance within pairs, $\sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc}$, and the imbalance measures $\beta_{k\ell}$ in (2.1).

Lemma 5 *Suppose there is a feasible flow f in $(\mathcal{N}, \mathcal{E})$, let the associated match be $\mathcal{M} = \{(\tau_t, \kappa_c) \in \mathcal{A} : f\{(\tau_t, \kappa_c)\} = 1\}$, and let $\beta_{k\ell}$ be the imbalance measure (2.1) for match $\mathcal{M}$. Then the cost of this flow satisfies*

$$\sum_{e\in\mathcal{E}} f(e)\ \texttt{cost}(e) \geq \sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc} + \sum_{k=1}^{K} \Upsilon^{K-k+1} \sum_{\ell=1}^{L_k} |\beta_{k\ell}|/2. \tag{2.3}$$

If f is a minimum cost feasible flow in $(\mathcal{N}, \mathcal{E})$, then (2.3) holds as an equality.

Theorem 6 says we may find the match in Definition 3 by solving a standard combinatorial optimization problem. There is a finite value (see §2.4.6) of the penalty Υ such that for that value and for all larger values, the resulting match satisfies the constraint of refined balance and minimizes the total covariate distance subject to that constraint.

Theorem 6 *If there exists a feasible flow in $(\mathcal{N}, \mathcal{E})$, then for sufficiently large Υ, a minimum cost flow in $(\mathcal{N}, \mathcal{E})$ yields an optimal refined acceptable 1-to-m match $\mathcal{M}$ given by $\mathcal{M} = \{(\tau_t, \kappa_c) \in \mathcal{A} : f\{(\tau_t, \kappa_c)\} = 1\}$. If there exists no feasible flow in $(\mathcal{N}, \mathcal{E})$, then there is no optimal refined acceptable 1-to-m match.*

2.4.6 Practical issues: deciding about Υ and m

Theorem 6 speaks of "sufficiently large Υ," and in its proof Υ is very large, specifically $\Upsilon > mTK + \sum_{(\tau_t,\kappa_c)\in\mathcal{A}} \delta_{tc}$. For stable computation, use a much smaller Υ, perhaps $\Upsilon = 2\max_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc}$ or smaller. Theorem 6 says that as Υ increases, eventually the imbalances $\sum_{\ell=1}^{L_1} |\beta_{1\ell}|$, ..., $\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$ are the best possible imbalances and further increases in Υ do not change the imbalances, so it is reasonable to match a few times, starting with a small Υ and gradually increasing it until the imbalances stop changing.

How many controls, m, should be matched to each treated unit? Match quality decreases as m increases, so one might match $m = 1$ to 1, examine the resulting average imbalances, $(mT)^{-1}\sum_{\ell=1}^{L_1} |\beta_{1\ell}|$, ..., $(mT)^{-1}\sum_{\ell=1}^{L_k} |\beta_{k\ell}|$, then match $m = 2$ to 1, and so on, stopping when the quality of the match is not acceptable.

2.4.7 Computation in sparse networks

Algorithms are standardly evaluated in terms of an upper bound on the rate of growth of the number of arithmetic steps required to solve them as the size of the problem increases (Cook et al. 1998, §1.2; Korte et al. 2008, §1.2). If steps $= O\left(\text{size}^3\right)$ then the number of arithmetic steps required to solve a problem grows by at most a constant multiple of the cube of the size of the problem. The point we want to make in the current section is that: (i) the new surgeons problem, and more generally the matching-within-natural-blocks problem, is sparse, with far fewer edges than typical matching problems, so (ii) vastly larger problems can be solved in these sparse networks than can be solved in dense networks commonly appearing in statistical matching problems, so (iii) we may balance covariates over an enormous number of natural blocks.

The network $(\mathcal{N}, \mathcal{E})$ is dense if $|\mathcal{E}| = O\left(|\mathcal{N}|^2\right)$ and sparse if $|\mathcal{E}| = O\left(|\mathcal{N}|\right)$. Our network is sparse; see §2.4.1. One can solve the minimum cost flow problem

in $O\left(|\mathcal{E}| \log\left[|\mathcal{E}|\left\{|\mathcal{E}| + |\mathcal{N}| \log\left(|\mathcal{N}|\right)\right\}\right]\right)$ steps; see Korte and Vygen (2008, Theorem 9.17, p. 214). If $|\mathcal{E}| = |\mathcal{N}|^2$, this is $O\left\{|\mathcal{N}|^2 \log\left(|\mathcal{N}|\right)\right\}$, whereas if $|\mathcal{E}| = |\mathcal{N}|$ it is $O\left[|\mathcal{N}| \log\left\{|\mathcal{N}|\right\}\right]$. In §2.4.4, $|\mathcal{N}| > T + C = 130106$ so $|\mathcal{N}|^2 \log\left(|\mathcal{N}|\right)$ is much larger than $|\mathcal{N}| \log\left(|\mathcal{N}|\right)$.

2.5 Do new and experienced surgeons differ?

2.5.1 Brief review of sensitivity analysis and attributable effects

There are $I = 6260$ pairs $i = 1, \ldots, I$ of two patients, $j = 1, 2$, matched for covariates, $\mathbf{x}_{ij}$, one treated with $Z_{ij} = 1$, the other control with $Z_{ij} = 0$, so $Z_{i1} + Z_{i2} = 1$. Write $\mathcal{Z}$ for the event that $Z_{i1} + Z_{i2} = 1$ for each i. Subject ij would exhibit binary response r_{Tij} if treated with $Z_{ij} = 1$ or binary response r_{Cij} if control with $Z_{ij} = 0$, so the observed response from ij is $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\; r_{Cij}$ and the effect of the treatment on ij, namely $\theta_{ij} = r_{Tij} - r_{Cij}$, is not observed; see Neyman et al. (1923) and Rubin (1974). Write $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \ldots \theta_{I2})$ for the $2I$-dimensional parameter and write $\mathcal{F} = \left\{\left(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}\right),\, i = 1, \ldots, I,\;\, j = 1, 2\right\}$. In the current study, $r_{Tij} = 1$ if ij would die within 30 days of surgery performed by the young surgeon in pair i, $r_{Tij} = 0$ otherwise, and $r_{Cij} = 1$ if ij would die within 30 days of surgery performed by the experienced surgeon in pair i, $r_{Cij} = 0$ otherwise. Then $(r_{Tij}, r_{Cij}) = (1, 0)$ if patient ij would die if surgery were performed by the young surgeon in pair i but not if performed by the experienced surgeon in pair i. The notation refers to two specific surgeons in pair i working at the same hospital.

If treatments are randomly assigned, then $\Pr\left(\, Z_{ij} = 1 \,|\, \mathcal{F},\, \mathcal{Z}\right) = 1/2$ with independent assignments in distinct pairs. The sensitivity analysis for nonrandom treatment assignment permits measured deviations from random assignment, specifi-

cally $(1+\Gamma)^{-1} \leq \Pr(Z_{ij}=1 \mid \mathcal{F}, \mathcal{Z}) \leq \Gamma/(1+\Gamma)$ for several $\Gamma \geq 1$; see Rosenbaum (2002a). A calculation in Rosenbaum and Silber (2009a) permits Γ to be interpreted in terms of an unobserved covariate associated with treatment and outcome. In the current paper, for a specified deviation from random assignment, $\Gamma \geq 1$, the sensitivity analysis will yield an upper bound on the P-value testing some hypothesis about treatment effects, so that, if that upper bound is at most α, then a bias of size Γ is too small to lead to acceptance of the hypothesis at level α. A sensitivity analysis asks: How much bias from nonrandom treatment assignment would need to be present to alter the conclusions of a randomization test, that is, to accept a null hypothesis that the randomization test has rejected?

Fisher's (1935) hypothesis of no treatment effect says $H_0 : r_{Tij} = r_{Cij}$ for all ij or equivalently $H_0 : \boldsymbol{\theta} = \mathbf{0}$. If H_0 were false, an interesting quantity is the attributable effect, $A = \sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}\,(r_{Tij} - r_{Cij}) = \sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}\,\theta_{ij}$; it is the number of additional deaths among patients of young surgeons ($Z_{ij}=1$) that would not have occurred had the experienced surgeon in the pair been picked to perform the surgery. If H_0 were true, then $A = 0$. If H_0 were false, then A would be an integer valued random variable. Of course, A is unobservable because $\theta_{ij} = r_{Tij} - r_{Cij}$ is never observed; however, it is possible to draw inferences about A; see Rosenbaum (2002a). This method uses a pivotal argument such that the observed number of deaths among patients of new surgeons, namely $\sum_{ij} Z_{ij}R_{ij}$, minus the unknown true value of A, is a random variable that satisfies the null hypothesis of no effect, $\sum_{ij} Z_{ij}R_{ij} - A = \sum_{ij} Z_{ij}\,r_{Cij}$, so that, for example, in a randomized experiment $\sum_{ij} Z_{ij}\,r_{Cij}$ is a constant plus a binomial random variable, as in McNemar's test. A null hypothesis about A is rejected if the individual null hypotheses $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ compatible with this value of A are all rejected. The calculation involves a binomial tail probability computed from a table of adjusted counts; see Rosenbaum (2002a, §6

and Table 5).

2.5.2 Sensitivity analyses for three-sided tests

Perhaps new surgeons are less capable and cause excess surgical deaths, so that $A > 0$. It is not inconceivable that new surgeons are more capable, having been more recently trained, so $A < 0$. Recent training might be relevant to laparoscopy and related techniques, in which a surgeon inserts a thin robotic surgical tool containing a camera, and manipulates the tool remotely. So it is of interest to test no effect H_0 against a two-sided alternative.

Failure to reject H_0 does not mean H_0 is approximately true. Rather, we wish to be assured that A is tolerably close to zero. For this, some form of equivalence test is needed.

Building upon the work of Bauer and Kieser (1996), Goeman et al. (2010) proposed a "three-sided test" for both difference and equivalence. It combines a two-sided test of no effect with the two-one-sided test procedure for testing inequivalence, all tests being done at the α-level, with no need of correction for multiple testing. Their underlying idea is both simple and clever. Three mutually incompatible hypotheses may be tested at level α without correction for multiple testing, because at most one hypothesis is true, so the α-risk of falsely rejecting a true null hypothesis is incurred at most once despite testing three null hypotheses. In brief, we may perform a two-sided test of no effect to establish both an effect and its direction, and perform a test of the null hypothesis of inequivalence to establish near equivalence, and do this without adjustment for multiple testing.

For sensitivity analyses, one attraction of the three-sided test is that we may use a standard method of sensitivity analysis three times, each time placing an upper bound on the relevant P-value in the presence of a bias in treatment assignment of at

most $\Gamma \geq 1$ for several values of Γ. The standard method says: if the null hypothesis is true and the bias in treatment assignment is at most Γ, then the chance that the upper bound on the P-value exceeds α is at most α. Logically, because at most one of the three null hypotheses is true, the standard method is either saying something trivial if all three null hypotheses are false, or it is referring to the one true null hypothesis despite our ignorance of the identity of that hypothesis. See Rosenbaum and Silber (2009b) for related discussion.

Fisher's $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is tested against a two sided alternative. The null hypothesis of inequivalence in the direction of harm done by new surgeons is defined to be $\boldsymbol{\theta} \geq \mathbf{0}$ (i.e., $\theta_{ij} \geq 0$ for all ij) with $A \geq \iota$ where $\iota > 0$ is a standard of inequivalence. The null hypothesis of inequivalence in the direction of benefit from new surgeons is defined to be $\boldsymbol{\theta} \leq \mathbf{0}$ with $A \leq -\iota$ where again $\iota > 0$. At most one hypothesis is true.

In the US in 2008, the annual mortality rate between age 75 and 76 was 3.95%; see (Arias, 2012). Most people aged 75 in 2008 did not undergo surgery. A risk associated with surgery in Medicare is small if it is small compared with the annual risk faced by the Medicare population. For illustration, we consider two definitions of inequivalence, ι, namely a quarter and a half of the annual mortality in the population at age 75, that is $\iota = 62 = 6260 \times 0.039506/4$ or $\iota = 124 = 6260 \times 0.039506/2$ extra deaths.

2.5.3 Mortality results

The overall 30-day mortality rate among the 2×6260 patients was 3.65%, made up of 3.59% for 6260 patients of experienced surgeons and 3.71% for 6260 patients of new surgeons (see Table 2.3). So the mortality rates for new and experienced surgeons look similar. The randomization test based on McNemar's test has two-sided P-value 0.7689, so the null hypothesis of no effect is plausible even in the absence of

Table 2.3: Mortality in 6260 pairs of matched pairs of patients, one treated by a new surgeon, the other by an experienced surgeon. The table counts pairs, not patients.

Experienced	New Surgeon			
Surgeon	Dead	Alive	Total	Percent
Dead	20	205	225	3.59%
Alive	212	5823	6035	96.41%
Total	232	6028	6260	
Percent	3.70%	96.30%		100.00%

Table 2.4: Sensitivity analysis using the three-sided test of the null hypotheses of no effect and substantial inequivalence with two definitions of inequivalence, $\iota = 62$ and $\iota = 124$. The test of no effect is two-sided, but the equivalence tests are one-sided. The table gives the upper bounds on the P-value for various magnitudes of bias Γ in assignment of patients to surgeons.

	No effect	Definition of inequivalence			
	$A = 0$	$A \geq \iota = 62$		$A \geq \iota = 124$	
	Surgeon type that caused more deaths				
Γ	Equal	Experienced	New	Experienced	New
1.0	0.7689	0.0003	0.0033	0.0000	0.0000
1.1	1.0000	0.0065	0.0379	0.0000	0.0000
1.2	1.0000	0.0521	0.1804	0.0000	0.0000
1.3	1.0000	0.2017	0.4508	0.0000	0.0000
1.4	1.0000	0.4571	0.7276	0.0000	0.0003
1.5	1.0000	0.7147	0.9000	0.0002	0.0026
1.6	1.0000	0.8841	0.9721	0.0012	0.0131
1.7	1.0000	0.9628	0.9938	0.0061	0.0452

unmeasured biases. From §2.3.2, this comparison refers to pairs of surgeons working at the same hospital, with identical distributions of operative procedures, and patients with similar comorbid conditions.

Table 3.4 gives the sensitivity analysis. For $\Gamma = 1$, this is a three-sided randomization test, and in the third column of Table 3.4, the hypothesis that experienced surgeons caused at least 62 extra deaths is rejected with P-value 0.0003, while in the fourth column the hypothesis that new surgeons caused at least an extra 62 deaths is rejected with P-value 0.0033. Biased assignment of patients to new or experienced surgeons might mask a substantial difference in mortality, making it appear to be no difference. In the fifth and sixth columns of Table 3.4, a bias of $\Gamma = 1.7$ is too small to mask a difference of $\iota = 124$ extra deaths in either direction. Using the

calculation in Rosenbaum and Silber (2009a), a bias of $\Gamma = 1.7$ could be produced by an unobserved covariate that more than tripled the odds of treatment by a young surgeon and more than tripled the odds of death.

In short, in the example, there are three findings. There is no evidence that mortality rates for new and experienced surgeons differ. A difference of 62 extra deaths caused by either type of surgeon is rejected in a randomization test, but a small bias of $\Gamma = 1.2$ could mask this difference, making it appear to be no difference. A larger difference of 124 extra deaths is rejected unless the bias is larger than a moderate $\Gamma = 1.7$, that is, the bias that could result from failing to match for an unobserved covariate that tripled the odds of treatment by a young surgeon and tripled the odds of death.

2.6 Discussion of other applications of the methodology

2.6.1 Nested nominal covariates in other applications

The priorities in Table 2.1 were based on the judgment of the surgeon on the research team. Expert judgment is one good way to create and order $\nu_1, \ldots, \nu_K$. Are there other ways?

Important covariates predict both treatment assignment and outcomes. Covariates that predict treatment show up as important in propensity scores estimated from the current data (Rosenbaum and Rubin, 1985), and covariates that predict outcomes show up as important in prognostic or risk scores estimated from external data (Hansen, 2008). The scores suggest covariates deserving priority for balancing, with the distance δ_{tc} seeking close individual pairs on the scores. Traskin and Small

(2011) approximate a propensity score using a regression tree, and such a tree creates a hierarchy of nominal variables to serve as $\nu_1, \ldots, \nu_K$. Alternatively, a lasso fit could prioritize the variables in either score.

A covariate that describes blocks or is constant for each block, such as hospital group in Table 2.1, has a marginal distribution that is balanced simply by matching within hospitals. However, including hospital group in Table 2.1 meant that its interactions with 14 other covariates were also balanced. A subgroup analysis that separately analyzed the two groups of pairs from the two types of hospitals would exhibit covariate balance within each subgroup separately, an important consideration for subgroup analyses.

2.6.2 Other sources of sparsity in optimal balanced matching

In the example, sparsity is created by the desire to match within natural blocks. Sparsity also arises in other ways. If there were one or two important continuous covariates, perhaps a propensity or risk score, then one might restrict the list of potential controls for a given treated subject to the short list comprised of the nearest c controls on those covariates. With fixed c, say $c = 100$, a sparse network is obtained. Refined covariate balance in such a network would obtain pairs that are close on the key covariates while balancing many nominal categories. As discussed by Zubizarreta et al. (2014), a match that reduces the heterogeneity of matched pair differences in outcomes, perhaps by matching closely for predictors of those outcomes, will both increase the power of a randomization test of no effect and increase its insensitivity to unmeasured biases.

With many nominal covariates, one might require exact matches for the most important nominal covariates, merely balancing the rest; then the short list of potential controls is comprised of the exact matches for those most important nominal covari-

ates. If the treatment is applied to everyone in a state or province, then one might wish to match treated subjects near the state boundary to nearby controls just across that boundary, and again this creates sparsity; see Keele et al. (2015) for one such study.

3

Constructed Second Control Groups and Attenuation of Unmeasured Biases

3.1 Introduction: background; motivating example

3.1.1 Is it advantageous to omit adjustments for some measured covariates?

In an observational study of treatment effects, treatments are not randomly assigned to individuals, so treated and control groups are often visibly different in terms of measured pretreatment covariates $\mathbf{x}$, and may differ in terms of unmeasured covariates u. Differing outcomes in treated and control groups after treatment may reflect the lack of comparability of these groups before treatment, rather than an effect caused by the treatment. It is common to adjust for the observed covariates $\mathbf{x}$, perhaps by matching individuals with the same $\mathbf{x}$, and to examine the sensitivity of conclusions to assumptions about unobserved covariates u.

It is sometimes argued informally that parts of $\mathbf{x}$ may be irrelevant, and that there would be less bias from u if adjustments were not made for the parts of $\mathbf{x}$ that

are irrelevant; see Brooks and Ohsfeldt (2013) and Ali et al. (2014) for two general perspectives on this issue, and see Walker (2013) and Zubizarreta et al. (2012) for discussion of a specific situations. The intuitive idea is that it is desirable that something irrelevant decides treatment assignment — that is similar to what happens in a randomized experiment — and if one removes every irrelevant aspect of treatment assignment, one is left with biases from u deciding treatment assignment. Under what circumstances does this line of reasoning have a rigorous basis?

3.1.2 Motivating example: Does smoking increase homocysteine levels?

To permit a tangible discussion, consider an interesting study by Bazzano et al. (2003) concerned with the possibility that cigarette smoking causes an increase in homocysteine levels, a possible risk factor for cardiovascular disease. Bazzano et al. (2003) compared smokers and nonsmokers in NHANES adjusting for certain covariates, $\overline{\mathbf{x}}$, that might have a direct biological connection with homocysteine levels, such as age, race and body mass index. They did not adjust for income and education, $\widetilde{\mathbf{x}}$, two covariates strongly related to smoking. In the US today, smoking is much less common among more educated, higher income individuals than among less educated, lower income individuals. Should one adjust for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ jointly or is it better to adjust for $\overline{\mathbf{x}}$ alone? One might argue that income and education have no known direct biological effect on homocysteine levels, so it makes sense to compare poor, less educated smokers to wealthier, better educated nonsmokers, because then something irrelevant has decided whether an individual smokes or not. Conversely, one might argue that one should adjust for all of $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ because education and income are associated with many aspects of daily life that could affect homocysteine levels, from exercise to diet to the quality of health care. Our goal is to shed some light on this decision and

related options for study design and analysis.

Figure 1 displays 1536 distinct individuals in $I = 512$ matched triples containing one daily smoker and two nonsmokers from NHANES 2005-2006. Smokers smoked every day for the last 30 days and reported smoking at least 10 cigarettes per day (median = 20). Nonsmokers did not smoke at all in the last 30 days and had smoked fewer than 100 cigarettes in their lives. All controls were matched to smokers for biological covariates, $\overline{\mathbf{x}}$, including age, gender, race (black/other), (Hispanic/other), and body mass index (BMI). Controls labeled M were also matched for two socioeconomic (SES) measures, $\widetilde{\mathbf{x}}$, namely education on a five point scale (with 1 meaning $<$ 9th grade, 3 meaning high school graduate, and 5 meaning at least a BA degree) and income recorded as the ratio of income to the poverty level capped at 5 times poverty. Controls labeled P were pushed apart in terms of $\widetilde{\mathbf{x}}$, that is, they had high levels of education and income. Notably in Figure 1, the three groups are similar in terms of biological covariates, the smokers and M-controls are similar in terms of SES, and the P-controls have higher education and income than the smokers. There is an obvious sense in which the M-controls are better than the P-controls: they are similar to smokers in terms of SES. Is there any sense in which the P-controls are better than the M-controls?

Section 3.2 reviews definitions and notation from existing literature. Section 3.3 considers the possibility that ignoring an irrelevant covariate $\widetilde{\mathbf{x}}$ attenuates bias from an unmeasured covariate u, concluding that it is possible, but the assumptions required are heroic and even then the magnitude of the attenuation is meaningful but not large. Also discussed is the possibility that forcing separation on $\widetilde{\mathbf{x}}$ can produce greater attenuation. Section §3.3.2 examines the relationship between an irrelevant covariate $\widetilde{\mathbf{x}}$ and an instrumental variable that might be used with the Wald estimator to estimate a complier-average-causal-effect (CACE). The remainder of the

paper concerns the construction and analysis of two control groups, one controlling for all of $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, the other controlling for $\overline{\mathbf{x}}$ and allowing or forcing separation on $\widetilde{\mathbf{x}}$. In particular, a form of simultaneous inference is proposed in which two sensitivity analyses are conducted for the two control groups, but the power loss for the controls matched for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ is small, so the second analysis adjusted for $\overline{\mathbf{x}}$ comes at little cost. The example uses data from NHANES 2005-2006 to examine the effects of smoking on homocysteine levels, in parallel with Bazzano et al. (2003) who used data from an earlier NHANES.

3.2 Review of notation and definitions

3.2.1 Treatment assignments and treatment effects

There are L individuals $\ell = 1, \ldots, L$ randomly sampled from an infinite population. Individual ℓ is described by $(r_{T\ell}, r_{C\ell}, Z_\ell, \overline{\mathbf{x}}_\ell, \widetilde{\mathbf{x}}_\ell, u_\ell)$, $\ell = 1, \ldots, L$, where $(\overline{\mathbf{x}}_\ell, \widetilde{\mathbf{x}}_\ell)$ are observed covariates, u_ℓ is an unobserved covariate, and individual ℓ exhibits response $r_{T\ell}$ if assigned to treatment, denoted $Z_\ell = 1$, or response $r_{C\ell}$ if assigned to control, denoted $Z_\ell = 0$, so the observed response from individual ℓ is $R_\ell = Z_\ell\, r_{T\ell} + (1 - Z_\ell)\; r_{C\ell}$, and the effect $r_{T\ell} - r_{C\ell}$ caused by the treatment is not observed for any individual ℓ; see Neyman et al. (1923), Welch (1937) and Rubin (1974). Fisher's (1935) sharp null hypothesis of no treatment effect H_0 asserts that $r_{T\ell} = r_{C\ell}$ for all ℓ. When referring to probability distributions in the population, the subscript ℓ is omitted. Following Dawid (1979), conditional independence of A and B given C is written $A \perp\!\!\!\perp B \,\big|\, C$.

When does it suffice to adjust for covariates $\mathbf{v}$ in causal inference? When may a portion of $\mathbf{v}$ safely be omitted from adjustments? We recall two definitions from the literature.

Definition 7 *(Rosenbaum and Rubin, 1983). Treatment assignment Z is said to be strongly ignorable given covariates* $\mathbf{v}$ *if*

$$(r_T,\, r_C) \;\underline{||}\; Z \,\Big|\, \mathbf{v}, \quad \textit{and } 0 < \Pr\left(Z = 1 \,|\, \mathbf{v}\right) < 1, \textit{ for all } \mathbf{v}. \tag{3.1}$$

For brevity and without further mention, the word ignorable is used in place of the term "strongly ignorable." If treatment assignment is ignorable given covariate $\mathbf{v}$, and if $\mathbf{v}$ were observed, then one can estimate causal effects such as $\mathrm{E}\left(r_T - r_C\right)$ or $\mathrm{E}\left(r_T - r_C|\mathbf{v}\right)$ or the average effect of the treatment on the treated, namely $\mathrm{E}\left(r_T - r_C|Z = 1\right)$, by adjusting for $\mathbf{v}$, for instance by matching or stratification; see Rosenbaum and Rubin (1983).

Definition 8 *(Heller et al., 2010). Covariates* $\mathbf{v}_2$ *in* $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$ *are said to be innocuous given* $\mathbf{v}_1$ *if*

$$(r_T,\, r_C) \;\underline{||}\; (Z,\, \mathbf{v}_2) \,\Big|\, \mathbf{v}_1. \tag{3.2}$$

It is straightforward to show that if treatment assignment Z is ignorable given $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$ and if $\mathbf{v}_2$ is innocuous, then treatment assignment is also ignorable given $\mathbf{v}_1$ alone. If $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$ were a measured covariate, if treatment assignment Z were ignorable given $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$, and if $\mathbf{v}_2$ were innocuous given $\mathbf{v}_1$, then causal parameters, such as $\mathrm{E}\left(r_T - r_C\right)$, could be consistently estimated adjusting for $\mathbf{v}_1$, ignoring $\mathbf{v}_2$.

If (3.1) and (3.2) both hold, then causal inference need not include adjustments for $\mathbf{v}_2$. Is there a benefit — not merely absence of harm — from not adjusting for $\mathbf{v}_2$? Claims of benefit in the literature refer to a situation with an unobserved covariate u that cannot be controlled by adjusting for observed covariates, whether $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\overline{\mathbf{x}}$. If treatment assignment were ignorable given $\mathbf{v} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)$ but not given $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\overline{\mathbf{x}}$, then causal effects could not be estimated by matching for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\overline{\mathbf{x}}$ because

u is not controlled. In this case, ask: Is it advantageous to ignore $\widetilde{\mathbf{x}}$ and adjust for $\overline{\mathbf{x}}$ alone? Informal discussions (e.g., Brooks and Ohsfeldt 2013; Ali et al. 2014) debate the possibility that if an innocuous $\widetilde{\mathbf{x}}$ is left unmatched then it decreases the role that u plays in determining treatment assignment, thereby reducing the bias created by our inability to adjust for an unmeasured covariate u. Is this true in any formal sense?

3.2.2 Quantifying the impact of an unobserved covariate on treatment assignment

If $\mathbf{x}$ is some observed covariate, perhaps $\mathbf{x} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\mathbf{x} = \overline{\mathbf{x}}$, then one model for sensitivity to unmeasured bias from u is expressed in terms of the potential influence of u on the odds $\Pr(Z = 1 \,|\, \mathbf{x},\, u) \,/\, \{1 - \Pr(Z = 1 \,|\, \mathbf{x},\, u)\}$ of treatment; see Rosenbaum (1987; 2002b, §4; 2007). This model quantifies bias in treatment assignment in terms of how the propensity score might be different if it took account of the unobserved u in addition to the observed $\mathbf{x}$. Consider two subjects with treatment assignments Z and Z' and unobserved covariates u and u' but the same value of the observed covariate, $\mathbf{x} = \mathbf{x}'$, so these two subjects might be matched when matching for $\mathbf{x}$. Then the odds ratio (for Z given $\mathbf{x}$ and u) or density ratio (for u given $\mathbf{x}$ and Z) linking treatment Z and the unobserved covariate u for these two subjects is:

$$\omega(\mathbf{x}, u, u') =$$
$$\frac{\Pr(Z = 1 \,|\, \mathbf{x},\, u)\ \Pr(Z' = 0 \,|\, \mathbf{x},\, u')}{\Pr(Z = 0 \,|\, \mathbf{x},\, u)\ \Pr(Z' = 1 \,|\, \mathbf{x},\, u')} = \frac{\Pr(u \,|\, \mathbf{x},\, Z = 1)\ \Pr(u' \,|\, \mathbf{x},\, Z' = 0)}{\Pr(u \,|\, \mathbf{x},\, Z = 0)\ \Pr(u' \,|\, \mathbf{x},\, Z' = 1)}, \qquad (3.3)$$

where the second equality follows from Bayes theorem. The sensitivity model says that the impact of failing to control u is at most $\Gamma \geq 1$ in the sense that

$$\frac{1}{\Gamma} \leq \omega\left(\mathbf{x}, u, u'\right) \leq \Gamma \text{ for all } \mathbf{x},\ u,\ u'; \tag{3.4}$$

that is, two subjects with the same $\mathbf{x}$ may differ in their odds of treatment by at most a factor of Γ because they differ in terms of u. Because $\omega\left(\mathbf{x}, u, u'\right) = 1/\omega\left(\mathbf{x}, u', u\right)$, equation (3.4) is actually redundant, and it is equivalent to write

$$\omega\left(\mathbf{x}, u, u'\right) \leq \Gamma \text{ for all } \mathbf{x},\ u,\ u'. \tag{3.5}$$

Typically, one would match a treated subject to a control with the same $\mathbf{x}$, so $Z + Z' = 1$, but they might differ in terms of $u \neq u'$. Conditionally given $Z + Z' = 1$, the probability of $(Z, Z') = (1, 0)$ is

$$\begin{aligned} &\frac{\Pr\left(Z = 1 \,|\, \mathbf{x},\ u\right) \Pr\left(Z' = 0 \,|\, \mathbf{x},\ u'\right)}{\Pr\left(Z = 1 \,|\, \mathbf{x},\ u\right) \Pr\left(Z' = 0 \,|\, \mathbf{x},\ u'\right) + \Pr\left(Z = 0 \,|\, \mathbf{x},\ u\right) \Pr\left(Z' = 1 \,|\, \mathbf{x},\ u'\right)} \\ &\qquad = \frac{\omega\left(\mathbf{x}, u, u'\right)}{\omega\left(\mathbf{x}, u, u'\right) + 1}, \end{aligned}$$

so that (3.4) or (3.5) implies $\varrho\left(\mathbf{x}, u, u'\right) = \Pr\left(Z = 1 \,|\, \mathbf{x},\ u,\ u',\ Z + Z' = 1\right)$ is bounded by

$$\frac{1}{1+\Gamma} \leq \varrho\left(\mathbf{x}, u, u'\right) \leq \frac{\Gamma}{1+\Gamma}, \text{ for all } \mathbf{x}, u, u'. \tag{3.6}$$

The one parameter Γ may be interpreted or amplified into an equivalent formulation in terms of two parameters, Λ and Δ, where Λ controls the relationship between treatment assignment Z and u, Δ controls the relationship between response under control r_C and u, and one sensitivity analysis at Γ is exactly equivalent to an infinite curve of sensitivity analyses with $\Gamma = (\Lambda\Delta + 1) / (\Lambda + \Delta)$; see Rosenbaum and Silber

(2009a) for a precise statement using the semiparametric model introduced by Wolfe (1974). For instance, as $1.25 = (2 \times 2 + 1) / (2 + 2)$, it follows that $\Gamma = 1.25$ is equivalent to an unobserved covariate that doubles the odds of treatment ($\Lambda = 2$) and doubles the odds of a positive treated-minus-control response difference ($\Delta = 2$). In other words, one may calculate and report a one-dimensional sensitivity analysis in terms of Γ but have available the interpretations of a two-dimensional sensitivity analysis in terms of (Λ, Δ).

3.3 When does ignoring an observed covariate attenuate the association between treatment assignment and an unobserved covariate?

3.3.1 Prods to receive treatment

To prod is to "goad, stimulate [or] prompt," according to the *Oxford English Dictionary.*

Definition 9 *The observed covariates* $\widetilde{\mathbf{x}}$ *are a prod to receive treatment given* $(\overline{\mathbf{x}}, u)$ *if*

$$\widetilde{\mathbf{x}} \underline{\,|\,|\,} u \,\Big|\, \overline{\mathbf{x}}, \quad \textit{and } \operatorname{var}\left\{ \Pr\left(Z = 1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u\right) \,|\, \overline{\mathbf{x}}, u \right\} > 0, \textit{ for all } (\overline{\mathbf{x}}, u)\,. \tag{3.7}$$

In (3.7), the condition $\widetilde{\mathbf{x}} \underline{\,|\,|\,} u \,\Big|\, \overline{\mathbf{x}}$ says that, given $\overline{\mathbf{x}}$, there is no information in $\widetilde{\mathbf{x}}$ about u. In other words, trying to remove some bias from the unobserved u by adjusting for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, rather than adjusting for $\overline{\mathbf{x}}$ alone, is not going to work, because $\widetilde{\mathbf{x}}$ is unrelated to u. The requirement in (3.7) that $\Pr\left(Z = 1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u\right)$ varies with $\widetilde{\mathbf{x}}$ for fixed $(\overline{\mathbf{x}}, u)$ says that, although $\widetilde{\mathbf{x}}$ is not informative about u, nonetheless $\widetilde{\mathbf{x}}$ does vary with treatment assignment.

Proposition 10 says that not matching for a prod $\widetilde{\mathbf{x}}$ strictly attenuates the relationship between treatment assignment Z and the unobserved covariate u, or in the notation of §3.2.2 that $\omega(\overline{\mathbf{x}},\, u,\, u')$ is strictly closer to 1 than is $\omega\{(\overline{\mathbf{x}},\, \widetilde{\mathbf{x}}), u, u'\}$.

Proposition 10 *Let $\widetilde{\mathbf{x}}$ be a prod to receive treatment given $(\overline{\mathbf{x}}, u)$. For any fixed $\overline{\mathbf{x}}$, u, u', if*

$$\frac{1}{\Gamma} \leq \omega\{(\overline{\mathbf{x}},\, \widetilde{\mathbf{x}}), u, u'\} \leq \Gamma \text{ for all } \widetilde{\mathbf{x}} \text{ with } \Gamma > 1, \tag{3.8}$$

then there exists an Υ with $1 \leq \Upsilon < \Gamma$ such that

$$\frac{1}{\Upsilon} \leq \omega(\overline{\mathbf{x}},\, u,\, u') \leq \Upsilon. \tag{3.9}$$

Proof. Following Freedman (2008, §9), define $f : (0,1) \to (0,\infty)$ by $f(p) = p/(1-p)$, so that $f(\cdot)$ is strictly increasing and $f^{-1}(v) = v/(1+v)$, and write $h(p) = f^{-1}\{\Gamma f(p)\}$. Freedman shows that $h(\cdot)$ is strictly concave on its domain, the open interval $(0,1)$. Now the second inequality in (3.8) implies

$$f\{\Pr(Z=1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)\} \leq \Gamma f\{\Pr(Z=1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u')\}$$
$$\Pr(Z=1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u) \leq h\{\Pr(Z=1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u')\}.$$

Using this and Jensen's inequality (e.g., Lange 2003, Proposition 3.5.1, page 61) for a strictly concave function yields

$$\begin{aligned}
\Pr(Z=1 \,|\, \overline{\mathbf{x}}, u) &= \int \Pr(Z=1 \,|\, \overline{\mathbf{x}},\, \widetilde{\mathbf{x}},\, u)\, \Pr(\widetilde{\mathbf{x}}|\overline{\mathbf{x}})\, d\widetilde{\mathbf{x}} \\
&\leq \int h\{\Pr(Z=1 \,|\, \overline{\mathbf{x}},\, \widetilde{\mathbf{x}},\, u')\}\, \Pr(\widetilde{\mathbf{x}}|\overline{\mathbf{x}})\, d\widetilde{\mathbf{x}} \\
&< h\left\{\int \Pr(Z=1 \,|\, \overline{\mathbf{x}},\, \widetilde{\mathbf{x}},\, u')\, \Pr(\widetilde{\mathbf{x}}|\overline{\mathbf{x}})\, d\widetilde{\mathbf{x}}\right\} \\
&= h\{\Pr(Z=1 \,|\, \overline{\mathbf{x}},\, u')\}.
\end{aligned} \tag{3.10}$$

Applying the increasing function $f(\cdot)$ to the first and last term in (3.10) yields $f\{\Pr(Z=1\,|\,\overline{\mathbf{x}},u)\} < \Gamma f\{\Pr(Z=1\,|\,\overline{\mathbf{x}},u')\}$ or equivalently $\omega(\overline{\mathbf{x}},u,u') < \Gamma$. Using instead the first inequality in (3.8) and $\omega\{(\overline{\mathbf{x}},\widetilde{\mathbf{x}}),u',u\} = 1/\omega\{(\overline{\mathbf{x}},\widetilde{\mathbf{x}}),u,u'\} \leq \Gamma$, the same argument shows $\omega(\overline{\mathbf{x}},u',u) < \Gamma$, and hence that $\omega(\overline{\mathbf{x}},u,u') = 1/\omega(\overline{\mathbf{x}},u',u)$ satisfies $\omega(\overline{\mathbf{x}},u,u') > 1/\Gamma$. Defining $\Upsilon = \max\{\omega(\overline{\mathbf{x}},u,u'),\,1/\omega(\overline{\mathbf{x}},u,u')\}$ completes the proof. ■

A few technical comments about Proposition 10 follow. First, in the definition of a prod, the requirement that $\text{var}\{\Pr(Z=1\,|\,\overline{\mathbf{x}},\widetilde{\mathbf{x}},u)\,|\,\overline{\mathbf{x}},u\} > 0$ in (3.7) is used to obtain the strict inequality in (3.10) by way of Jensen's inequality (e.g., Lange 2003, Proposition 3.5.1, page 61). Proposition 10 says there is strict attenuation, $\Gamma > \Upsilon$, for each $\overline{\mathbf{x}}$, u, u'; however, the degree of attenuation Υ in (3.9) generally depends upon $\overline{\mathbf{x}}$, u, u'. As a consequence, if the sensitivity model (3.4) were true with $\mathbf{x} = (\overline{\mathbf{x}},\widetilde{\mathbf{x}})$, then (3.8) would hold uniformly in $\overline{\mathbf{x}}$, $\widetilde{\mathbf{x}}$, u, u', but this would not imply that there exists one $\Upsilon < \Gamma$ such that (3.9) holds uniformly in $\overline{\mathbf{x}}$, u, u'. That is, Proposition 10 shows there is strict attentuation at each $\overline{\mathbf{x}}$, u, u', not that there is uniformly strict attenuation. It is clear that if one focused on the subpopulation with $\widetilde{\mathbf{x}} \in \mathcal{C}$ for some subset $\mathcal{C}$, then essentially the same proof shows there is attenuation in every subpopulation defined by $\widetilde{\mathbf{x}}$.

Proposition 10 is of no use on its own. However, if treatment assignment were ignorable given $(\overline{\mathbf{x}},\widetilde{\mathbf{x}},u)$, if $\widetilde{\mathbf{x}}$ were innocuous given $(\overline{\mathbf{x}},u)$ and if $\widetilde{\mathbf{x}}$ were a prod to receive treatment given $(\overline{\mathbf{x}},u)$, then: (i) it is suffices to focus attention on $(\overline{\mathbf{x}},u)$ ignoring $\widetilde{\mathbf{x}}$, because adjustments for $(\overline{\mathbf{x}},u)$ would permit estimation of causal effects, and (ii) it is also advantageous to focus attention on $(\overline{\mathbf{x}},u)$ ignoring $\widetilde{\mathbf{x}}$, because the association between treatment assignment Z and u has been attenuated.

The heavy assumptions required to use Proposition 10 are consequential. Failing to adjust for $\widetilde{\mathbf{x}}$ could increase the bias for either or both of two reasons: (i) if treatment

assignment were ignorable given ($\overline{\mathbf{x}}$, $\widetilde{\mathbf{x}}$, u) but not given ($\overline{\mathbf{x}}$, u), then adjusting for $\widetilde{\mathbf{x}}$ may reduce bias from $\widetilde{\mathbf{x}}$, (ii) even if $\widetilde{\mathbf{x}}$ itself seems to have no direct relevance, adjusting for $\widetilde{\mathbf{x}}$ might possibly reduce bias from u to the extent that $\widetilde{\mathbf{x}}$ and u are associated and the left side of (3.7) fails to hold.

Because Proposition 10 is of no use on its own, its actual usefulness is a matter of speculation. The additional assumptions that would make Proposition 10 useful are stringent assumptions about an unobserved covariate, and any investigator who makes these assumptions can expect an argument from skeptics. Rather than argue for or against the additional assumptions that would make Proposition 10 useful, we suggest conducting two analyses, one with and the other without these assumptions. A simple version of this has two control groups, one matched to treated subjects for ($\overline{\mathbf{x}}$, $\widetilde{\mathbf{x}}$), the other matched for $\overline{\mathbf{x}}$ alone. Heller et al. (2010) observe that if treatment assignment were ignorable given ($\overline{\mathbf{x}}$, $\widetilde{\mathbf{x}}$) and if $\widetilde{\mathbf{x}}$ were innocuous given $\overline{\mathbf{x}}$, then these two comparisons of treated subjects to these two matched control groups would estimate the same parameter, the average effect of the treatment on the treated, so contrasting these two estimates provides a test of these two assumptions. In contrast, Proposition 11 in §3.6 frames the discussion of these two control groups when they may both be affected by bias from an unmeasured covariate u.

3.3.2 Is a prod an instrument?

So far, §3.3 has considered the possibility of comparing outcomes R in treated, $Z = 1$, and control, $Z = 0$, groups without adjustment for a covariate $\widetilde{\mathbf{x}}$ that meets certain additional, fairly speculative, conditions required of a prod. As noted in §3.1.1, this possibility has been discussed in several recent articles concerned with health outcomes research, including Brooks and Ohsfeldt (2013), Ali et al. (2014), Walker (2013), and citetzubizarreta2012contrasting. The method we propose in §3.5 takes

the analysis adjusting for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ as the primary analysis, then adds at negligible cost in power a secondary analysis adjusting for $\overline{\mathbf{x}}$ but not for $\widetilde{\mathbf{x}}$, while controlling the familywise error rate in these two analyses, and making use of controls who might otherwise have been discarded. Could one, instead, view $\widetilde{\mathbf{x}}$ as an instrument or instrumental variable? Viewing $\widetilde{\mathbf{x}}$ as an instrument might suggest a different analysis, say the Wald estimator or two-stage least squares, aimed at estimating the so-called "complier-average causal effect" or CACE.

By definition in the Neyman-Rubin framework, a covariate is a variable whose value is determined prior to treatment assignment Z and hence unaffected by which treatment an individual ultimately receives; that is, a covariate has single version that is the same whether or not $Z = 1$ or $Z = 0$, like $\overline{\mathbf{x}}$ or $\widetilde{\mathbf{x}}$ and unlike R or Z. In this framework, an instrument (recorded in an instrumental variable) is a very special kind of treatment that encourages an experimental subject to take a second treatment over which the experimentor lacks direct control, but the encouragement-treatment affects outcomes only to the extent that it alters acceptance of the second treatment; see Angrist, Imbens and Rubin (1996), Hirano et al. (2000) and Holland (1988). The CACE is the average effect of the second treatment on subjects who would respond to the encouragement treatment by changing their adoption of the second treatment, and Angrist et al. (1996) show that the CACE is the estimand of the Wald estimator. For instance, the Vietnam War draft lottery randomly selected people for the draft, a treatment that "encouraged" some people to serve in the military, though many men served without being drafted and others found ways to dodge the draft; see Angrist et al. (1996). For the draft lottery, the CACE is the average effect of military service on the subset of men who would serve in the military only if drafted.

A substantial literature consistent with the Neyman-Rubin framework cautions against adjusting for certain variables that, unlike $\widetilde{\mathbf{x}}$, are not covariates. In par-

ticular,Rosenbaum (1984, 2015b) cautions against adjusting for other outcomes of treatment, noting that such an adjustment can create a bias that would otherwise be absent. Several authors wisely advise against adjusting for instruments, such as the draft lottery used as an instrument for military service; see, for instance, Wooldridge (2009), Myers et al. (2011), Pearl (2010, 2011), and Bhattacharya and Vogt (2012).

In §3.1.2 and §3.6.2, $\widetilde{\mathbf{x}}$ describes income and education. In the context of NHANES, income and education are plausible covariates for smoking. In particular, we have a clear idea about what it means to be poor and uneducated, and we have no difficulty imagining a person of any fixed income or education choosing to smoke or not smoke. If treatment assignment were ignorable given $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ and if $\widetilde{\mathbf{x}}$ were innocuous given $\overline{\mathbf{x}}$, then the two matched comparisons of the treated group to each of the two control groups would estimate the same parameter, namely the average effect of the treatment on the treated. Although smoking is, in 2015, relatively uncommon among individuals with relatively high income and education, it would be quite a stretch to regard income and education as "treatments" that discourage smoking. For income and education to be instruments, the estimand in instrumental variables estimation, the CACE, would then be the average effect of smoking on people who would change their smoking behavior in response to a substantial change in income and education, a nebulous estimand at best. Within the view of instruments proposed by Angrist et al. (1996), it is not easy to think of income and education as instruments, so within that view, a prod — a type of covariate — is not an instrument — a type of treatment. An older view of instruments defines them in a context-free manner purely in terms of conditional independence or moment conditions. Within this older view, instruments of the type studied by Angrist et al. (1996) and prods might be viewed as two nonoverlapping subsets. A general principle is that an estimand should be clear and intelligible before an investigator sets out to estimate

it. Our sense is that the CACE fails that principle for income and education in the NHANES example in §3.1.2 and §3.6.2. We do not regard the Wald estimator or two-stage least squares as options in this example.

3.4 The magnitude of the attenuation: direct calculation under a simple model

Proposition 10 says that not adjusting for a prod $\widetilde{\mathbf{x}}$ attenuates the bias in (3.8) because the inequality in (3.9) is strict. How large is this attenuation? For fixed (u, u'), how much closer to 1 is $\omega\left(\overline{\mathbf{x}}, u, u'\right)$ than $\omega\{(\overline{\mathbf{x}}, \widetilde{\mathbf{x}}), u, u'\}$? As in §3.3, Bayes theorem permits us to think about the answer in terms of the imbalance in u in treated, $Z = 1$, and control, $Z = 0$, groups. Table 3.1 provides an answer to how large the attenuation is in a simple case in which there is no $\overline{\mathbf{x}}$, $\widetilde{x}$ is a scalar prod with $\widetilde{x} \sim N\left(0, \sigma^2\right)$ for $\sigma = 1/2$ or 1, and treatment assignment probabilities follow a logit model, $\text{logit}\left\{\Pr(Z = 1|\widetilde{x}, u)\right\} = \alpha + \widetilde{x} + \gamma u$, so that for $u = 0$ and $u' = 1$, condition (3.8) holds with equality as $\Gamma = \exp\left(\gamma\right) = \omega(\widetilde{\mathbf{x}}, u, u')$. Under this model, for fixed u and u', the odds of treatment are $\exp\left(2\sigma\right)$ times greater when $\widetilde{x}$ is one standard deviation above its mean than when it is one standard deviation below its mean, or $\exp\left(2\sigma\right) = 2.71$ for $\sigma = 1/2$ and $\exp\left(2\sigma\right) = 7.39$ for $\sigma = 1$, so for both values of σ the prod $\widetilde{x}$ substantially alters the treatment assignment probabilities. Table 3.1 displays the attenuated $\omega\left(\overline{\mathbf{x}}, u, u'\right)$ with $u = 0$ and $u' = 1$, obtained by evaluating (3.10) by numerical integration. For example, for $\alpha = -1$, for $\sigma = 1/2$, a moderate bias of $\Gamma = \exp\left(\gamma\right) = 1.5$ attenuates to 1.47, whereas for $\sigma = 1$ a large bias of $\Gamma = 5$ attenuates to 3.81. The impression from the simple example in Table 3.1 is that: (i) a prod $\widetilde{x}$ must substantially affect the treatment assignment probabilities to produce substantial attenuation, and (ii) even when there is substantial attenuation, the bias

that remains is far from small.

3.5 Two control groups: controlling for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\overline{\mathbf{x}}$

3.5.1 Using two control groups

Proposition 10 reaches an attractive conclusion — a reduction in unmeasured biases in (3.9) — on the basis of heroic assumptions in (3.2) with $\mathbf{v}_1 = (\overline{\mathbf{x}}, u)$ and $\mathbf{v}_2 = \widetilde{\mathbf{x}}$ and (3.7) — the strong influence but total irrelevance of the prod $\widetilde{\mathbf{x}}$. In many applications, investigators will be understandably reluctant to rely on such strong assumptions to achieve the modest level of attenuation seen in Table 3.1. There is, however, a practical way to use Proposition 10 to see a little more in observational data without committing to the strong assumptions in Proposition 10, that is, a way to have it both ways.

The possibility of using two control groups subject to different biases is much discussed in the literature on observational studies; see, for instance, Campbell (1969), Rosenbaum et al. (1987); Rosenbaum (2015a), Meyer (1995), Shadish et al. (2002), Stuart and Rubin (2008), West et al. (2008), Heller et al. (2010) and Lu et al. (2011). Typically, these two control groups are found rather than constructed; that is, the groups existed as groups before the investigation began.

With varied motivations, several recent studies have used the computer to construct two control groups, one matched for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, the other match only for $\overline{\mathbf{x}}$; see Daniel et al. (2008), Heller et al. (2010), and Silber et al. (2012, 2013). These two control groups may be nonoverlapping, perhaps constructed using the tapered matching algorithm of Daniel et al. (2008), or they may share controls. Matched control groups that share controls may be compared to each other using a device known as

the exterior match; see Rosenbaum and Silber (2013). Matching ensures that $\overline{\mathbf{x}}$ has the same distribution in the treated group and both control groups, a helpful fact if the magnitude of the treatment effect varies with $\overline{\mathbf{x}}$; however, at the risk of losing this desirable property, one could alternatively adjust for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\overline{\mathbf{x}}$ using some form of covariance adjustment.

Suppose that two control groups are formed, perhaps overlapping, perhaps not, one matched for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, the other just for $\widetilde{\mathbf{x}}$. In the context of Proposition 10, if there are benefits to not matching for $\widetilde{\mathbf{x}}$, then we see such an analysis, but if the strong assumptions in Definitions 8 and 9 are false or doubtful, then we see an analysis that does not depend upon these assumptions. Moreover, we are able to compare these two analyses.

Strict use of Proposition 10 would perform two unrelated and therefore typically overlapping matches, one for $\overline{\mathbf{x}}$ alone, the other for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$. In this strict use, each match does not alter the other match: the match for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ does not alter the distribution of $\widetilde{\mathbf{x}}$ in the match for $\overline{\mathbf{x}}$ alone, so Proposition 10 speaks directly to the consequences of leaving $\widetilde{\mathbf{x}}$ unmatched. An alternative approach inspired by Proposition 10 but only informally linked to it would force the two matches to use different controls, thereby typically using more controls, with better matches for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ going to the match that controls $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ and worse matches for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ going to the match for $\overline{\mathbf{x}}$ alone, as happens in tapered matching (Daniel et al., 2008). Because this alternative approach forces the two matched control groups to be nonoverlapping, the two control groups compete for controls, so there is some distortion of the distribution of the unmatched prod $\widetilde{\mathbf{x}}$. Another alternative also inspired by Proposition 10 but even more informally linked to it would force the two matches to use different controls and additionally force the controls matched for $\overline{\mathbf{x}}$ alone to differ from the treated group in terms of $\widetilde{\mathbf{x}}$. The goal in this second alternative is to achieve greater attenuation

of bias from u by picking controls precisely because the prod $\widetilde{\mathbf{x}}$ pushed them into the control group; see §3.5.2.

As noted previously, the attenuation result in Proposition 10 holds whether or not $\widetilde{\mathbf{x}}$ is innocuous, but attenuation is useful in an observational study only if $\widetilde{\mathbf{x}}$ is innocuous given $(\overline{\mathbf{x}}, u)$ in the sense of Definition 8, for otherwise the attenuation of bias from u may be more than offset by bias from failure to control $\widetilde{\mathbf{x}}$. In the current paragraph, assume treatment assignment is ignorable given $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}}', u)$. Were it true that $\widetilde{\mathbf{x}}$ is innocuous given $(\overline{\mathbf{x}}, u)$, then

$$\Pr\left(r_T,\, r_C \mid \overline{\mathbf{x}},\, u\right) = \Pr\left(r_T,\, r_C \mid \overline{\mathbf{x}}, \widetilde{\mathbf{x}},\, u\right)$$

$$= \Pr\left(r_T,\, r_C \mid Z = z,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}},\, u\right) = \Pr\left(r_T,\, r_C \mid Z = z,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}',\, u\right) \text{ for all } \widetilde{\mathbf{x}},\, \widetilde{\mathbf{x}}'. \quad (3.11)$$

We observe treated response distributions from treated subjects, say $\Pr\left(R \mid Z = 1,\, \overline{\mathbf{x}}\right) = \Pr\left(r_T \mid Z = 1,\, \overline{\mathbf{x}}\right)$ or $\Pr\left(R \mid Z = 1,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}\right) = \Pr\left(r_T \mid Z = 1,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}\right)$, and control response distributions from control subjects, say $\Pr\left(R \mid Z = 0,\, \overline{\mathbf{x}}\right) = \Pr\left(r_C \mid Z = 0,\, \overline{\mathbf{x}}\right)$, or $\Pr\left(R \mid Z = 0,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}\right) = \Pr\left(r_C \mid Z = 0,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}\right)$. Treated response distributions may differ from control response distributions either because of a treatment effect or because of a bias. In contrast, if we compare two control response distributions, say $\Pr\left(r_C \mid Z = 0,\, \overline{\mathbf{x}}\right)$ versus $\Pr\left(r_C \mid Z = 0,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}\right)$, for controls matched to the same treated subject, then these differ when (3.11) holds only because of bias from the failure to control the unobserved covariate u. This is true of all three matches in the previous paragraph when (3.11) holds, and forcing $\widetilde{\mathbf{x}}$ to differ in the third match may provide a greater opportunity to check whether or not $\Pr\left(r_C \mid Z = 0,\, \overline{\mathbf{x}}\right)$ and $\Pr\left(r_C \mid Z = 0,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}\right)$ differ. If $\Pr\left(r_C \mid Z = 0,\, \overline{\mathbf{x}}\right)$ and $\Pr\left(r_C \mid Z = 0,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}\right)$ do differ, then this can indicate bias from u or it can indicate that $\widetilde{\mathbf{x}}$ is not innocuous given $(\overline{\mathbf{x}}, u)$ or both (so (3.11) does not hold), but it surely indicates that at least one control

group cannot be trusted.

3.5.2 Attenuation with forced separation

The magnitude of attenuation is now considered under a simple method for forcing separation on a prod, so treated and control groups are further apart on the prod than they would be if the prod were left unmatched. In brief summary, forcing separation increases attentuation when the initial bias is large, but the attenuated bias that remains is still large, even when treated and control groups are widely separated on the prod, as in the example in Figure 1. The logit-model formulation used here is similar to §4 except treated units are matched to controls whose prod $\tilde{x}$ is less than or equal to $c\sigma$ for some cutoff c. The smaller c is for a given σ, the greater the separation on the prod. By analogy with (3.3), we use Bayes theorem and measure attenuation by comparing odds ratios of $u = 1$ versus $u = 0$ in treated, $Z = 1$, and control, $Z = 0$, groups. Here we consider $\alpha = -1$ so that there are more control units than treated units, which is needed for matching to ensure separation on a prod; the results (not shown) were similar for $\alpha = 0$ and $\alpha = 1$. Table 3.2 shows the attenuation for different values c. The top half of Table 3.2 can be compared to the first line of Table 3.1 and the second half of Table 3.2 can be compared to the fourth line of Table 3.1. In Table 3.2 matching to ensure separation on a prod creates greater attenuation than leaving the prod unmatched. The differences are fairly small for moderate Γ: for $\Gamma = 1.5$ and $\sigma = 1/2$, even for $c = -1$, ensuring separation on the prod only increased the attenuation from 1.47 to 1.39. The differences are more substantial for larger Γ, e.g., for $\Gamma = 10$ and $\sigma = 1/2$, for $c = -1$, ensuring separation on the prod increases the attenuation from 8.88 to 6.42. Table 3.3 shows how much separation on the prod is created by matching to ensure separation on the prod for different values of c. The table reports the standardized difference on the prod when matching treated units to

Table 3.1: Degree of attenuation of bias Γ by not matching for a Normally distributed prod $\widetilde{x}$ with expectation 0 and standard deviation σ.

		Γ						
σ	α	1	1.5	2	3	4	5	10
1/2	−1	1.00	1.47	1.93	2.83	3.71	4.59	8.88
1/2	0	1.00	1.47	1.93	2.83	3.73	4.63	9.07
1/2	1	1.00	1.47	1.94	2.88	3.82	4.75	9.41
1	−1	1.00	1.40	1.78	2.49	3.16	3.81	6.82
1	0	1.00	1.40	1.78	2.50	3.19	3.86	7.11
1	1	1.00	1.41	1.81	2.58	3.34	4.08	7.75

Table 3.2: Degree of attenuation of bias Γ by matching treated units to control units with prod $\leq c\sigma$ for a normally distributed prod with expectation 0 and standard deviation σ, where the treatment assignment probabilities follow a logit model $\log\{\Pr(Z=1|\tilde{x},u)/\Pr(Z=0|\tilde{x},u)\} = \alpha + \tilde{x} + \gamma u$ with $\alpha = -1$ and $\Gamma = \exp(\gamma)$. The attenuation is measured by the odds ratio linking u and the group.

		Γ						
σ	c	1	1.5	2	3	4	5	10
1/2	-1	1.00	1.39	1.77	2.40	3.07	3.62	6.42
1/2	0	1.00	1.44	1.83	2.61	3.33	4.09	7.58
1/2	1	1.00	1.45	1.88	2.74	3.60	4.43	8.48
1	-1	1.00	1.30	1.54	1.99	2.35	2.67	4.14
1	0	1.00	1.36	1.66	2.20	2.73	3.22	5.42
1	1	1.00	1.38	1.74	2.42	3.04	3.68	6.49

control units with prod $\leq c\sigma$. For $\sigma = 1/2$, the standardized difference ranges from about $1.8 - 1.9$ (depending on Γ) with $c = -1$ to $0.6 - 0.7$ with $c = 1$.

3.5.3 An algorithm for matching to ensure separation on a prod

We now introduce an algorithm to create matches that exhibit balance on $\overline{\mathbf{x}}$ and force separation on $\widetilde{\mathbf{x}}$. The algorithm produced the match in Figure 1. This new algorithm slightly extends the balanced optimal matching technique of Pimentel et al. (2015); see Hansen and Klopfer (2006) and Stuart (2010) for other discussions of matching algorithms in observational studies. That approach used penalized network flows to select controls with a covariate distribution as similar as possible to the treated group for large numbers of nominal covariates and their interactions. The extension

Table 3.3: Standardized difference on the prod $\tilde{x}$ when matching treated units to control units with prod $\leq c\sigma$ for a normally distributed prod with expectation 0 and standard deviation σ, where the treatment assignment probabilities follow a logit model $\log\{\tilde{\omega}(\bar{x}, \tilde{x}, u)\} = \alpha + \tilde{x} + \gamma u$, with $\alpha = -1$ and $\Gamma = \exp(\gamma)$. The standardized difference is $\{E(\tilde{x}|Z=1) - E(\tilde{x}|Z=0, \tilde{x} \leq c\sigma)\} \cdot \left\{\sqrt{\frac{1}{2}\left[Var(\tilde{x}|Z=1) + Var(\tilde{x}|Z=0)\right]}\right\}^{-1}$.

		Γ						
σ	c	1	1.5	2	3	4	5	10
1/2	-1	1.93	1.91	1.89	1.86	1.83	1.81	1.75
1/2	0	1.21	1.19	1.16	1.14	1.12	1.10	1.05
1/2	1	0.72	0.71	0.70	0.68	0.65	0.63	0.58
1	-1	2.31	2.28	2.24	2.19	2.15	2.11	2.00
1	0	1.56	1.53	1.50	1.45	1.41	1.39	1.30
1	1	1.09	1.07	1.05	1.01	0.98	0.96	0.87

proposed here selects controls to be similar to treated subjects in some ways and as different as possible in others. The original algorithm has a target distribution for the covariates in the control group, and the extension simply changes the target distribution. In the example, this means that controls should resemble the treated group in terms of biological quantities, age, gender, BMI, but should be as high as possible in terms of education and income. A precise description is given in Appendix **??**. To create separation on a prod $\widetilde{\mathbf{x}}$ while balancing $\overline{\mathbf{x}}$, we first define a new covariate

$$\eta(\widetilde{\mathbf{x}}_i) = \begin{cases} 1 & \text{if } \widetilde{\mathbf{x}}_i \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{X}$ is a set of desired values for the prod. The target distribution for controls has the same distribution of $\overline{\mathbf{x}}_i$ as the treated group and has $\eta(\widetilde{\mathbf{x}}_i) = 1$. Running the algorithm for this target group and with balance constraints on $\overline{\mathbf{x}}$ and $\eta(\widetilde{\mathbf{x}})$ selects a control group with a distribution of $\overline{\mathbf{x}}$ very similar to that in the treated population, but also ensures that as many of the controls as possible are chosen with $\widetilde{\mathbf{x}}$ values in the region $\mathcal{X}$, thereby creating separation on the prod.

3.6 Inference with and without a prod

3.6.1 Sensitivity analysis with two control groups controlling the familywise error rate

Figure 2 shows homocysteine levels in blood plasma for the $I = 512$ matched triples in Figure 1; see §3.1.2. The current section is concerned with the simultaneous analysis of prodded and unprodded match sets of the type displayed in Figure 2.

Define the null hypothesis H'_Γ to be the conjunction of (i) Fisher's hypothesis of no effect, H_0, (ii) treatment assignment Z is ignorable given $\mathbf{v} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)$ and (iii) a bias in treatment assignment from u of at most $\Gamma \geq 1$ in pairs of individuals matched for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, so that (3.8) holds for all $\overline{\mathbf{x}}$, $\widetilde{\mathbf{x}}$, u, u'. Define the null hypothesis H^*_Υ to be the conjunction of (i) Fisher's hypothesis of no effect, H_0, (ii) treatment assignment Z is ignorable given $\mathbf{v} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)$, (iv) $\widetilde{\mathbf{x}}$ is innocuous given $(\overline{\mathbf{x}}, u)$, (v) a bias in treatment assignment from u of at most $\Upsilon \geq 1$ in pairs of individuals matched for $\overline{\mathbf{x}}$, so that (3.9) holds for all $\overline{\mathbf{x}}$, u, u'. Obviously, rejecting H'_Γ or H^*_Υ leaves open whether Fisher's H_0 is false or whether the additional assumptions are false. Notably, H'_Γ and H^*_Υ share (i) and (ii) but H'_Γ adds (iii) while H^*_Υ omits (iii) and adds (iv) and (v), although all of assumptions (i)-(v) could be jointly true. The data used to test H'_Γ and H^*_Υ are dependent because the same treated subjects are used in both tests, as in Figure 2, and also if the control groups are allowed to overlap or share some controls, as is not true in Figure 2.

If H'_Γ or H^*_Υ were both true, then Proposition 10 would lead us to anticipate modest attenuation of unmeasured biases. That is, Proposition 10 leads us to be interested in testing pairs $(H'_\Gamma, H^*_\Upsilon)$ with Υ modestly smaller than Γ, perhaps $\Upsilon = \omega\Gamma$ for $\omega = 0.9$, or 10% smaller based on Table 2.

We propose to use a multiple testing procedure to conduct two sensitivity analyses,

one for H'_Γ and one for H^*_Υ, correcting for multiple testing using the recycling method of Burman et al. (2009). The recycling procedure strongly controls the familywise error rate. Let $0 < \alpha' \leq \alpha < 1$ be two fixed numbers, conventionally $\alpha = 0.05$. Fix (Γ, Υ), say $(\Gamma, \Upsilon) = (\Gamma, \omega\Gamma)$, and compute the two upper bounds on P-values, say $p'_{\Gamma,\max}$ and $p^*_{\Upsilon,\max}$, from separate sensitivity analyses for H'_Γ and H^*_Υ, respectively. In the example, the method in Rosenbaum (2007) yields $p'_{\Gamma,\max}$ and $p^*_{\Upsilon,\max}$ using the `R` package `sensitivitymw`. The recycling steps are:

Recycling procedure:

1. **Test H'_Γ:** Reject H'_Γ at level α in the presence of a bias of at most Γ if $p'_{\Gamma,\max} \leq \alpha'$.

2. **Test H^*_Υ:** If H'_Γ was rejected in step 1, then reject H^*_Υ at level α in the presence of a bias of at most Υ if $p^*_{\Upsilon,\max} \leq \alpha$. Otherwise, if H'_Γ was not rejected in step 1, then reject H^*_Υ at level α in the presence of a bias of at most Υ if $p^*_{\Upsilon,\max} \leq \alpha - \alpha'$.

3. **Recycle to retest H'_Γ:** If H'_Γ was not rejected in step 1 but H^*_Υ was rejected in step 2, then reject H'_Γ at level α in the presence of a bias of at most Γ if $p'_{\Gamma,\max} \leq \alpha$.

For a fixed (Γ, Υ) with $\alpha' = \alpha/2$, then this recycling procedure is easily seen to be equivalent to the standard version of Holm's (1979) procedure, and if $0 < \alpha' < \alpha$, then it is equivalent to Holm's (1979) weighted procedure with $w' = \alpha'/\alpha$ and $w^* = (\alpha - \alpha')/\alpha$. These equivalences are seen by considering the four possible outcomes of steps 1-3. As noted by Benjamini and Hochberg (1997, p. 411), the weighted Holm procedure is superior to another weighting scheme with two hypotheses, as here. Taking $\alpha' = \alpha$ is fixed sequence testing, so rejection of H^*_Υ can occur only if H'_Γ is rejected in step 1, and step 3 is redundant. So in our case with two hypotheses, the recycling procedure reduces to one of two other methods, but is attractive in

unifying them. To reject both H'_Γ and H^*_Υ is to have $\max\left(p'_{\Gamma,\max}, p^*_{\Upsilon,\max}\right) \leq \alpha$ as for intersection-union testing (Berger, 1982; Laska and Meisner, 1989); however, intersection-union testing could reject when recycling does not if $\alpha' < \alpha$, and recycling could reject just one hypothesis, either H'_Γ and H^*_Υ, which intersection-union testing cannot.

Conventionally, $\alpha = 0.05$. How should α' be chosen? If an analysis that controlled $\overline{\mathbf{x}}$ but not $\widetilde{\mathbf{x}}$ would be implausible if it disagreed with an analysis that controls $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, then α' should be close to α, perhaps $\alpha' \in [0.8\alpha, \alpha]$. Arguably this is the case with $\widetilde{\mathbf{x}}$ recording income and education in the smoking example, so we take $\alpha' = 0.04 < \alpha = 0.05$, but taking $\alpha' = \alpha = 0.05$ would be reasonable also. In this way, little power is lost in the analysis that adjusts for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, yet both analyses are considered with strong control for testing two null hypotheses.

The discussion above considered a single fixed (Γ, Υ). In fact, we consider not a fixed (Γ, Υ) but rather a sequence $(\Gamma, \Upsilon) = \{\Gamma_n, \max(1, \omega\Gamma_n)\}$, $n = 1, 2, \ldots$, with $\Gamma_1 = 1$ and $\Gamma_n \to \infty$ as $n \to \infty$, where $\omega > 0$ is fixed. In practice, reasonable values of ω are $\omega = 0.9$, hoping for modest attenuation, or $\omega = 1$, preferring to handle the two control groups symmetrically. At step n, a total of $2n$ hypotheses have been tested using the recycling procedure.

Proposition 11 *For fixed $\omega > 0$, apply the recycling procedure to*

$$(\Gamma_n, \Upsilon_n) = \{\Gamma_n, \max(1, \omega\Gamma_n)\} \qquad \textit{for } n = 1, 2, \ldots.$$

The chance of falsely rejecting at least one true hypothesis, H'_{Γ_n} or $H^*_{\Upsilon_n}$, $n = 1, 2, \ldots$, is at most α.

Proof. Recall that $p'_{\Gamma,\max}$ is a valid P-value for testing H'_Γ alone and $p'_{\Gamma,\max}$ increases with Γ, whereas $p^*_{\Upsilon,\max}$ is a valid P-value for testing H^*_Υ alone and $p^*_{\Upsilon,\max}$ increases

with Υ. Also, the recycling procedure controls the familywise error when testing both H'_Γ and H^*_Υ with any one fixed (Γ, Υ). Let $\overline{\Gamma} = \inf\left\{\Gamma_n : H'_{\Gamma_n} \text{ is true}\right\}$ and $\overline{\Upsilon} = \inf\left\{\Upsilon_n : \ H^*_{\Upsilon_n} \text{ is true}\right\}$, where $\overline{\Gamma} = \infty$ and $\overline{\Upsilon} = \infty$ are possible values. To avoid a separate discussion of the infinite cases, define $p'_{\infty,\max} = p^*_{\infty,\max} = 1$. By definition of the hypotheses earlier in this section, H'_{Γ_n} is true for all $\Gamma_n \geq \overline{\Gamma}$ and $H^*_{\Upsilon_n}$ is true for all $\Upsilon_n \geq \overline{\Upsilon}$. Hence, the smallest $p'_{\Gamma_n,\max}$ for a true H'_{Γ_n} is $p'_{\overline{\Gamma},\max}$ and the smallest $p^*_{\Upsilon_n,\max}$ for a true $H^*_{\Upsilon_n}$ is $p^*_{\overline{\Upsilon},\max}$. We consider cases. If $\overline{\Gamma} = \overline{\Upsilon} = \infty$, then there is nothing to prove, because no true hypothesis is tested. If $\overline{\Upsilon} = \omega\overline{\Gamma} < \infty$, then to reject any true hypothesis, one must have $\left(p'_{\overline{\Gamma},\max} \leq \alpha'\right) \vee \left(p^*_{\overline{\Upsilon},\max} \leq \alpha - \alpha'\right)$ and the chance of this is as most α. If $\overline{\Upsilon} < \omega\overline{\Gamma}$, then a false rejection for (Γ_n, Υ_n) with $\Gamma_n < \overline{\Gamma}$ and $\overline{\Upsilon} \leq \Upsilon_n < \omega\overline{\Gamma}$ requires rejection of the true $H^*_{\overline{\Upsilon}}$ with $p^*_{\overline{\Upsilon},\max} \leq \alpha$ which occurs with probability at most α, whereas false rejection for (Γ_n, Υ_n) with $\Gamma \geq \overline{\Gamma}$ and $\Upsilon \geq \omega\overline{\Gamma}$ requires $\left(p'_{\Gamma_n,\max} \leq \alpha'\right) \vee \left(p^*_{\Upsilon_n,\max} \leq \alpha - \alpha'\right)$, which implies $\left(p'_{\overline{\Gamma},\max} \leq \alpha'\right) \vee \left(p^*_{\overline{\Upsilon},\max} \leq \alpha - \alpha'\right)$ which has probability at most α. The case $\overline{\Upsilon} > \omega\overline{\Gamma}$ is analogous. ■

3.6.2 Example: Using wealthy, educated nonsmokers as a second control group

For the data in §3.1.2, Figure 2 compares homocysteine levels among smokers to two control groups, one (M) matched to controls for all measured covariates, the other (P) separated from the smokers on the prod $\widetilde{\mathbf{x}}$ of education and income; see, again, Figure 1 for the difference in education and income among these groups. The smokers in Figure 2 appear to have somewhat higher homocysteine levels than both control groups, whereas control groups M and P appear similar. We now conduct a sensitivity analysis using the procedure in §3.6.

With I treatment-control matched pairs, Maritz (1979) used the null randomization distribution of Huber's one-sample M-statistic $T = \sum_{i=1}^{I} \psi\left(Y_i/s\right)$ to test

Fisher's null hypothesis of no effect, where Y_i is a treated-minus-control matched pair difference in responses R, s is the median $|Y_i|$, and $\psi(\cdot)$ is an odd function, $\psi(y) = -\psi(-y)$. Taking $\psi_{\rm t}(y) = y$ makes T into a constant multiple of the sample mean, and then Maritz's method is equivalent to the randomization distribution of the mean, that is, the permutational t-test; see Pitman (1937) and Welch (1937). Huber's $\psi_{\rm hu}(\cdot)$ has $\psi_{\rm hu}(y) = {\rm sign}(y) \cdot \min(|y|, \kappa)$ for some $\kappa > 0$, where ${\rm sign}(y) = 1$, 0, -1 as $y > 0$, $y = 0$, $y < 0$, so $\psi_{\rm hu}(\cdot)$ has the same influence function as a trimmed mean. A sensitivity analysis for T when used in observational studies was proposed in Rosenbaum (2007), its power and large sample properties in sensitivity analysis with various choices of $\psi(\cdot)$ were examined in Rosenbaum (2013), and the method was implemented in the `sensitivitymv` and `sensitivitymw` packages in `R`; see Rosenbaum (2015c). In particular, taking $\psi_{\rm in}(y) = \{\kappa/(\kappa-\iota)\} \cdot {\rm sign}(y) \cdot \max\{0, \min(|y|, \kappa) - \iota\}$ for some $\kappa > \iota \geq 0$ entails inner trimming and means $\psi_{\rm in}(y)$ is zero for $|y| \in [0, \iota]$, is ${\rm sign}(y) \cdot \kappa$ for $|y| \geq \kappa$, and rises linearly from 0 to κ on $[\iota, \kappa]$. For many distributions of Y_i, the M-statistic $T = \sum_{i=1}^{I} \psi_{\rm in}(Y_i/s)$ reports greater insensitivity to unmeasured biases than does $\psi_{\rm hu}(\cdot)$. Here, we set $\kappa = 2$ and $\iota = 1/2$; see Rosenbaum (2013, Table 3) and `method="p"` in the `senmw` function of the `sensitivitymw` package in `R`.

Table 3.4 performs a sensitivity analysis with $(\Gamma, \Upsilon) = (\Gamma,\ 0.9 \times \Gamma)$ for an increasing sequence of values of Γ, as discussed in §3.6.1, reporting the upper bounds, $p^{'}_{\Gamma,\max}$ and $p^{*}_{\Upsilon,\max}$, on the marginal P-values testing H'_Γ and H^*_Υ, respectively. Table 3.4 does not control for testing two hypotheses. Using the method in §3.6.1 to control for testing two hypotheses and testing in a fixed sequence with $\alpha = \alpha' = 0.05$ leads to rejection of H'_Γ and H^*_Υ for $(\Gamma, \Upsilon) = (1.640,\ 1.476)$, and no rejections at $(\Gamma, \Upsilon) = (1.650,\ 1.485)$. Recycling with $\alpha = 0.05$ and $\alpha' = 0.04$ rejects H'_Γ and H^*_Υ for $(\Gamma, \Upsilon) = (1.640,\ 1.476)$, tests H'_Γ at the 0.05 level for $\Gamma = 1.650$ but fails to reject, and barely rejects H^*_Υ for $\Upsilon = 1.575$. To put this in context, $\Gamma = 5/3 = 1.667$ cor-

Table 3.4: Sensitivity analysis in the example with two control groups, one matched (M) for $\widetilde{\mathbf{x}}$ with sensitivity parameter Γ, the other using $\widetilde{\mathbf{x}}$ as a prod (P) with sensitivity parameter $\Upsilon = 0.9 \times \Gamma$. The tabled values are upper bounds on marginal P-values using M-statistics with inner trimming, ψ_{in} with $\iota = 0.5$, $\kappa = 2$.

	Sensitivity parameters						
Γ	1.000	1.250	1.500	1.600	1.640	1.650	1.750
$\Upsilon = 0.9 \times \Gamma$	1.000	1.125	1.350	1.440	1.476	1.485	1.575
	Upper bounds on P-values testing no effect						
M-controls ($p'_{\Gamma,\max}$)	0.000	0.000	0.009	0.031	0.047	0.051	0.119
P-controls ($p^*_{\Upsilon,\max}$)	0.000	0.000	0.000	0.001	0.002	0.002	0.009

responds with an unobserved covariate that triples the odds of treatment and triples the odds of a positive pair difference in outcomes, while $\Gamma = 1.5$ corresponds with an unobserved covariate that doubles the odds of treatment and doubles the odds of positive pair difference in outcomes; see Rosenbaum and Silber (2009a) and the `amplify` function in the `sensitivitymv` package in `R`.

If, as may be, it is important to adjust for socioeconomic factors $\widetilde{\mathbf{x}}$, then Table 3.4 does this, finding that the results are not sensitive to small unmeasured biases. If, as may be, socioeconomic factors introduce a biologically irrelevant source of variation in smoking behavior, so that comparing people differing in $\widetilde{\mathbf{x}}$ attenuates bias from unmeasured covariates u, then Table 3.4 does this also, finding again that the results are not sensitive to small unmeasured biases. Whether you compare people with the same or different education and income, smokers tend to have higher homocysteine levels than nonsmokers. These two analyses bracket the one analysis in Bazzano et al. (2003), where $\widetilde{\mathbf{x}}$ was neither controlled nor separated.

Arguably, Table 3.4 allows us to see more in an observational study than we would have seen with either comparison alone, yet it avoids committing us to one or another set of assumptions about unmeasured covariates, assumptions that are easy enough to state but difficult if not impossible to justify. Moreover, in this example, the M-controls were tested at level 0.05, yet the familywise error rate for two tests was also controlled at $\alpha = 0.05$, so the addition of the P-controls came without cost. The

simulation in §3.6.3 asks whether this pattern is expected in general.

3.6.3 Simulation: power of the recycling procedure in a sensitivity analysis

Ideally, a sensitivity analysis would reject the null hypothesis of no effect when there is no unmeasured bias and there is a treatment effect, and the power of a sensitivity analysis is the probability that this will happen; see Rosenbaum (2004, 2013). More precisely, the power of an α-level sensitivity analysis allowing for bias Γ is the probability that the upper bound on the P-value leads to rejection when computed with this Γ. The simulation contrasts testing in a fixed sequence, $\alpha = \alpha' = 0.05$, and recycling with $\alpha = 0.05$ and $\alpha' = 0.04$. The simulation also contrasts exploring the (Γ, Υ) sequence along $(\Gamma, \Upsilon) = (\Gamma, \Gamma)$ with equal sensitivity parameters and along $(\Gamma, \Upsilon) = \{\Gamma, \max(1, 0.9 \times \Gamma)\}$. The latter sequence makes sense if the investigator included the prodded controls anticipating moderate attenuation of unmeasured biases.

Table 3.5 simulates a simple situation in which all treated-minus-control pair differences in both control groups are Normal with expectation τ and variance 1. The correlation between the pair-differences in the two control groups is 1/2 because the same treated subject is matched to two different controls, as in §3.1.2. The effect size is either $\tau = 1/4$ or $\tau = 1/2$. Of course, the results are less sensitive with a larger effect, and Γ is adjusted accordingly, $\Gamma = 1.5$ for $\tau = 1/4$, $\Gamma = 2.8$ for $\tau = 1/2$. Columns a and b, labeled $\alpha' = 0.05$, refer to testing in a fixed sequence. Columns c and d, labeled $\alpha' = 0.04$, refer to recycling. The final column is for comparison only: column e gives the power if H^*_Υ were tested at the $\alpha = 0.05$ level with no correction for testing two hypotheses. Although a part of fixed sequence testing, column a for $\alpha' = 0.05$ and H'_Γ analogously gives the power when testing H'_Γ without correction

Table 3.5: Simulated power of an $\alpha = 0.05$ level sensitivity analysis with $I = 500$ matched triples, Normal errors, and an additive constant treatment effect that is τ standard deviations of a treated-minus-control matched pair difference. For $\alpha' = 0.05$, there is fixed-sequence testing, whereas for $\alpha' = 0.04$ there is recycling or equivalently a weighted Holm procedure. Either $\Upsilon = 0.9 \times \Gamma$ or $\Upsilon = \Gamma$. Two ψ-functions are compared. Estimated from 50000 independent replicates.

Γ	Υ	$\alpha' = 0.05$		$\alpha' = 0.04$		
Column Label		a	b	c	d	e
		H'_Γ	H^*_Υ	H'_Γ	H^*_Υ	H^*_Υ Alone
		$\tau = 1/4$ with ψ_{hu}				
1.50	1.35	0.47	0.44	0.46	0.66	0.82
1.50	1.50	0.47	0.30	0.44	0.35	0.47
		$\tau = 1/4$ with ψ_{in}				
1.50	1.35	0.66	0.63	0.65	0.80	0.90
1.50	1.50	0.66	0.50	0.63	0.56	0.66
		$\tau = 1/2$ with ψ_{hu}				
2.80	2.52	0.32	0.28	0.31	0.47	0.68
2.80	2.80	0.32	0.17	0.29	0.20	0.32
		$\tau = 1/2$ with ψ_{in}				
2.80	2.52	0.77	0.75	0.77	0.87	0.94
2.80	2.80	0.77	0.65	0.75	0.69	0.77

for multiple testing, because in fixed sequence testing the first hypothesis in the sequence is tested without correction. Table 3.5 also compares the power when using $\psi_{\text{hu}}(\cdot)$ and $\psi_{\text{in}}(\cdot)$ with $\kappa = 2$ and $\iota = 1/2$ and, as expected from Rosenbaum (2013), the power is greater with $\psi_{\text{in}}(\cdot)$. Each situation is replicated 50,000 times, so the standard error of an estimated power is at most $\sqrt{0.25/50000} = 0.0022$.

With fixed sequence testing, adding a second comparison does not reduce the power of the first comparison, but it affects the power of subsequent comparisons. For instance, in Table 3.5 with $(\Gamma, \Upsilon) = (1.5, 1.5)$, $\tau = 1/4$, $\psi_{\text{hu}}(\cdot)$, the power is 0.47 for H'_Γ alone, for H'_Γ as first in sequence, and for H^*_Υ alone in the last column, but H^*_Υ tested in fixed sequence after testing H'_Γ has power of only 0.30. In contrast, with $(\Gamma, \Upsilon) = (\Gamma, \Gamma)$, recycling with $\alpha' = 0.04$ slightly reduces the power for H'_Γ and somewhat increases the power for H^*_Υ.

There is some attraction to conducting the sensitivity analysis through a sequence of the form $(\Gamma, \Upsilon) = \{\Gamma, \max(1, 0.9 \times \Gamma)\}$, as was done in the example in Table 3.4.

That sequence is interesting because the prod, if it actually works, is intended to attenuate bias, as in Proposition 10, so values of Υ somewhat below Γ are not without interest. At the same time, in the simulated situation, recycling of unused α is much more likely to occur when $\Upsilon = 0.9 \times \Gamma$, so the power loss is smaller. Specifically, when $\Upsilon = 0.9 \times \Gamma$ in Table 3.4, the power for H'_Γ is only slightly lower in column c than in column a, typically about 1% lower, whereas the power for H^*_Υ is much higher in column d than in column b. The combination of $\alpha' = 0.04$ and $\Upsilon = 0.9 \times \Gamma$ is, therefore, attractive: despite correction for performing two tests, the M-controls are tested at nearly the power of a single test, while the smaller value of $\Upsilon = 0.9 \times \Gamma$ for the P-controls means the second comparison also has high power.

3.7 Summary: Prefer additional analyses to additional assumptions

It has been argued in the literature that leaving a measured covariate $\widetilde{\mathbf{x}}$ uncontrolled, say unmatched, may attenuate biases from an unmeasured covariate u. Although this is formally true, the argument requires very strong, typically doubtful, assumptions about both observed and unobserved covariates, and even when those assumptions are true the magnitude of the attenuation is modest. We suggest that one should not conduct a single analysis that presumes these doubtful assumptions are true. Rather, we suggest building two control groups, with two analyses, one that controls for $\widetilde{\mathbf{x}}$ and one that leaves $\widetilde{\mathbf{x}}$ uncontrolled. Often, the second control group uses individuals who would otherwise be excluded from the analysis because they are so different from treated subjects in terms of $\widetilde{\mathbf{x}}$. A second control group entails a second hypothesis test, hence a correction for testing two hypotheses; however, by careful organization of the analyses, there is only a slight loss of power in the primary comparison controlling

$\widetilde{\mathbf{x}}$, so the second control group is nearly without cost.

4

An Exact Test of Fit for the Gaussian Linear Model using Optimal Nonbipartite Matching

4.1 Notation and review

4.1.1 The Gaussian linear model

The familiar Gaussian linear model assumes that an n-dimensional stochastic outcome $\mathbf{y}$ and an $n \times p$ dimensional fixed matrix $\mathbf{X}$, with $p < n$, are related by

$$\mathrm{E}\left(\mathbf{y} \middle| \mathbf{X}\right) = \mathbf{X}\boldsymbol{\beta}, \quad \boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \sim N_n\left(\mathbf{0},\, \sigma^2\mathbf{I}\right), \tag{4.1}$$

where $\boldsymbol{\beta}$ and σ^2 are unknown parameters, $\mathbf{0}$ and $\mathbf{I}$ are, respectively, the n-dimensional zero vector and identity matrix, and $N_n\left(\cdot,\, \cdot\right)$ is the n-dimensional multivariate Normal distribution. A test of fit of (4.1) is a test of the null hypothesis H_0 that (4.1) is true, and such a test is said to be exact, as opposed to asymptotic — that it, the test has exact level α — if the probability that the test rejects H_0 when it is true is $\leq \alpha$. Generally, we assume that $\mathbf{X}$ has full column rank p, so the least squares estimate of $\boldsymbol{\beta}$ under (4.1) is $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$, the fitted values are $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} =$

$\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$ where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$, the residuals are $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$, and the unbiased estimate of σ^2 is $\widehat{\sigma}^2 = \mathbf{e}^T\mathbf{e}/\left(n-p\right)$.

4.1.2 Tests of fit based on replicates and near-replicates

Fisher (1922) proposed testing the Gaussian linear model in experiments by including replicates of design points, thereby providing an estimate of pure error unaffected by misspecification of the linear model, yielding an exact F-test of H_0. This device is commonly used in central composite designs in response surface experiments, with the center-point replicated several times, and the factorial and axial points appearing only once in isolation; see Box and Draper (1982) and Draper (1982).

Outside of designed experiments, exact replicates occur sporadically if at all. Several investigators have proposed an analogous test based on near-replicates; see, for instance, Christensen (1989; 1991; 2011, §6.6.2), Daniel and Wood (1971, §7.5), Green (1971), Joglekar et al. (1989), Neill and Johnson (1985), Shillington (1979), and Su and Yang (2006). This work emphasizes certain options in the choice of test statistic, whereas our contribution emphasizes the construction of the near-replicates. In particular, we use optimal nonbipartite matching, reviewed in §4.1.3, and the device that Tukey (1949) introduced in constructing his "one degree of freedom for nonadditivity" test in the unreplicated row-by-column design. For different approaches to constructing near-replicates, see Miller et al. (1998, 1999) and Miller and Neill (2008).

4.1.3 Optimal nonbipartite matching

Given L points with L even and an $L \times L$ symmetric matrix of nonnegative distances between pairs of points, an optimal nonbipartite match divides the L points into $L/2$ nonoverlapping pairs of two points so that the total of the $L/2$ within-pair distances

is minimized. This combinatorial optimization problem may be solved in polynomial time by a suitable algorithm; see Jungnickel (2013, §14.4). In `R`, the `nbpMatching` package of Lu et al. (2011) makes available the algorithm of Derigs (1988), and we used it in the current paper.

Nonbipartite matching has been used to solve various problems in observational studies, including matching with time-dependent propensity scores (Lu, 2005) and strengthening instrumental variables (Baiocchi et al., 2010; Zubizarreta et al., 2013). See Lu et al. (2011) for a survey of statistical applications of nonbipartite matching. Here, we use optimal nonbipartite matching as one aspect of constructing near replicates. For general discussion of optimal matching in observational studies, see Rosenbaum (2010, Part II) and Stuart (2010).

In the statistical applications described above, it is common to form pairs using only some of the available observations, with the algorithm itself deciding which observations to leave unpaired. This is done using so-called "sinks". Suppose that there are n observations with an $n \times n$ distance matrix and we want m pairs, with specified $m \leq n/2$. Then $n - 2m$ observations are not paired. If the $n \times n$ distance matrix contains any zeros off the diagonal, then we add a constant, say 1, to all of the off-diagonal entries, so they are all strictly positive. Introduce $n - 2m$ sinks that are at 0 distance to all observations and at infinite distance to one another. That is, expand the distance matrix with 3 blocks, a block of extra columns of 0's of dimension $n \times (n - 2m)$, a block of extra rows of 0's of dimension $(n - 2m) \times n$, and a square lower-right-corner block of ∞'s of dimension $(n - 2m) \times (n - 2m)$. One then calculates an optimal nonbipartite match with this expanded distance matrix, regarding any observation paired with a sink as unpaired. This strategy forms m pairs of observations in such a way that the total of the m within pair distances is minimized over two choices: (i) which $n - 2m$ observations to leave unpaired, and (ii)

how to best pair the $2m$ observations that are paired.

In central composite experimental designs, only central points are replicated. In nonexperimental data, there are often many points that have no near-replicate. Motivated by these considerations and some preliminary simulations, we leave approximately $(n-p)/3$ observations unpaired, pairing the rest. When n is large compared to p, nearly a third of the observations are unpaired and two thirds are paired, leaving nearly $n/3$ degrees of freedom within-pairs to estimate error from near replicates. Stated precisely, we form m pairs where m is $n/2-(n-p)/6$ rounded to the nearest integer, and we leave exactly $n-2m$ observations unpaired using $n-2m$ sinks. Here, $n-2m$ is approximately $(n-p)/3$.

4.1.4 Tukey's device and its extensions

A well-known problem with techniques that rely on near-neighbors or near-replicates is that, unless the number of predictors is very small, we will rarely see two individuals who are nearly the same on all of the predictors. In light of this, we need to define the distance with some guidance from the data about which predictors actually matter for prediction. At the same time, we need to prevent this double use of the $\mathbf{y}$'s from invalidating the test. For this purpose, a device introduced by Tukey (1949) is helpful.

Tukey (1949) proposed a test for interaction in the unreplicated row-by-column design using the following clever device. The device has been generalized several times, and we describe the generalized form for Gaussian linear models here; see, for instance, Mandel (1959), Scheffé (1959, Problem 4.19), Milliken and Graybill (1970), Andrews (1971), Rao (1973, §4e.1), St. Laurent (1990), Christensen and Utts (1992) and Christensen (2011, §9.5). A basic fact about the distribution of $\boldsymbol{\epsilon}$ in (4.1) is that projections of $\boldsymbol{\epsilon}$ onto orthogonal subspaces are independent; this fact is the key

element in the Fisher-Cochran theorem. In particular, the fitted values $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ and residuals $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\,\mathbf{y}$ are independent in (4.1). Write $\rho(\mathbf{M})$ for the rank of a matrix $\mathbf{M}$. Let $\mathbf{L}$ be any matrix with n rows that is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$ such that $\rho([\mathbf{X}, \mathbf{L}]) < n$. It is easily seen that

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T H\mathbf{y} \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\mathbf{y}}, \qquad (4.2)
\end{aligned}$$

so that $\widehat{\boldsymbol{\beta}}$ is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$, and in particular, $\mathbf{L}$ can be a function of $\widehat{\boldsymbol{\beta}}$, as in Tukey (1949). Milliken and Graybill (1970, §2) observe that if (4.1) is true and $\mathbf{y}$ is regressed on $[\mathbf{X}, \mathbf{L}]$, then the usual F-test of the hypothesis that the coefficients of $\mathbf{L}$ are simultaneously zero has a central F-distribution with degrees of freedom $\rho([\mathbf{X}, \mathbf{L}]) - p$ and $n - \rho([\mathbf{X}, \mathbf{L}])$. Here, $[\mathbf{X}, \mathbf{L}]$ need not have full column rank, but must have rank less than n. As discussed by Milliken and Graybill (1970), the distribution of this F-statistic under the alternative that (4.1) is false is not, in general, a noncentral F-distribution and is typically intractable.

4.2 An exact test of fit for the Gaussian linear model

4.2.1 General procedure

Starting with a suitable distance matrix, we round $n/2 - (n-p)/6$ to the nearest integer to obtain m, as in §4.1.3, and we use optimal nonbipartite matching in §4.1.3

to build m pairs of two observations and $n-2m$ unpaired observations so that the total distance within the m pairs is minimized. This match is intended to find the closest m pairs of near replicates and $n-2m$ individuals who are further away, yielding roughly m degrees of freedom from near replicates to estimate an error variance less affected by any misspecification of model (4.1). Define $\mathbf{L}$ to be a matrix with $m+n-2m=n-m$ columns, where the first m columns of $\mathbf{L}$ each contain exactly two ones and $n-2$ zeros, the two ones in column k indicating the two individuals paired in pair k, $k=1,\ldots,m$. The last $n-2m$ columns of $\mathbf{L}$ each contain a one and $n-1$ zeros, the 1 indicating the ℓth individual who was not paired, $\ell=1,\ldots,n-2m$. Notice that the $n-2m$ unpaired individuals each have their own column and will be fitted exactly, somewhat in parallel with the proposal of Utts (1982); see also Christensen (2011, p. 153). The $n-m$ columns of $\mathbf{L}$ have rank $\rho(\mathbf{L})=n-m-1$ because each row of $\mathbf{L}$ sums to 1. In general, the rank of $\rho([\mathbf{X},\,\mathbf{L}])$ will depend on $\mathbf{X}$.

The test of fit of (4.1) is simply an F-test of $H_0:\boldsymbol{\gamma}=\mathbf{0}$ in the Gaussian linear model

$$\mathbf{y}=[\mathbf{X},\,\mathbf{L}]\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{bmatrix}+\boldsymbol{\zeta},\quad \boldsymbol{\zeta}\sim N_n\left(\mathbf{0},\,\omega^2\mathbf{I}\right) \tag{4.3}$$

with degrees of freedom $\rho([\mathbf{X},\,\mathbf{L}])-p$ and $n-\rho([\mathbf{X},\,\mathbf{L}])$. In our proposed test, the residual degrees of freedom, $n-\rho([\mathbf{X},\,\mathbf{L}])$, will approach $n/3$ as $n\to\infty$ with p fixed. Because $[\mathbf{X},\,\mathbf{L}]$ is not of full column rank, a little care, of a conventional sort, is needed in computing the F-test.

Christensen (1991) proposed an alternative modified test, no longer the standard F-test of $H_0:\boldsymbol{\gamma}=\mathbf{0}$, with a view to gains in power. Our limited simulation (not shown) comparing the standard F-test to this modified test suggests that the dimension reduction devices we describe later have large effects on power, while the

choice of test statistic has a smaller effect, so we adhere to the standard F-test in our discussion here, rather than add an extra dimension to our simulated comparisons.

If the distances were based on $\mathbf{X}$ alone, then [$\mathbf{X}$, $\mathbf{L}$] would be a function or transformation of $\mathbf{X}$, and the test of (4.1) against (4.3) is simply a comparison of two nested Gaussian linear models. If the null hypothesis, namely (4.1), were true, then $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ is true in (4.3) and, in the standard way, the corresponding F-statistic has a central F-distribution; see, for instance, Rao (1973, §4b.2) or Christensen (2011, §3.2). In §4.2.2, we permit $\mathbf{L}$ to depend upon both $\mathbf{X}$ and $\widehat{\mathbf{y}}$; then, the corresponding F-statistic is no longer a standard test of a general linear hypothesis, but it still has a central F-distribution when the null hypothesis (4.1) is true using the generalization of Tukey's device; see Milliken and Graybill (1970), Rao (1973, pp. 251-252) or Christensen (2011, §9.5).

For instance, the distance matrix could be the Mahalanobis distance between pairs of rows of $\mathbf{X}$. The usual Mahalanobis distance can perform oddly when a column of $\mathbf{X}$ is either long tailed or a rare binary variable. An alternative robust Mahalanobis distance addresses both issues: it replaces the columns of $\mathbf{X}$ by column ranks before computing the distances, with average ranks for ties; however, it uses untied variances and covariance of ranks, thereby reducing the role of rare binary variables; see Rosenbaum (2010, §8). As is commonly done, we speak of the quadratic form as the Mahalanobis distance (or the robust Mahalanobis distance), whereas technically it is its square root that is a norm.

The estimate of ω^2 in (4.3) may be smaller than the estimate of σ^2 in (4.1) for two reasons. First, the estimate of ω^2 only reflects differences in y's between paired individuals, and paired individuals are as close as possible on the predictors. Second, the $n - 2m$ individuals who were not paired do not contribute to the estimate of ω^2 because they are fitted exactly in (4.3). These $n - 2m$ unpaired individuals are

each far from all other individuals. If some of these $n - 2m$ unpaired individuals are poorly fit by (4.1), eliminating them from the estimate of ω^2 may aid in recognizing this lack of fit.

Model (4.3) creates an estimate $\widehat{\omega}^2$ from neighbors that may be less affected by model misspecification than $\widehat{\sigma}^2$ obtained from fitting model (4.1). The parameter $\boldsymbol{\gamma}$ is of high dimension and is not typically of interest, so one may use computational simplifications with the highly structured matrix $\mathbf{L}$ to obtain the F-test without estimating $\boldsymbol{\gamma}$.

When the number of predictors is not small, close matches on all predictors will be rare. An alternative distance matrix is discussed in §4.2.2: it emphasizes the predictors that appear to matter in the fit of model (4.1), but avoids double use of the $\mathbf{y}$ by employing Tukey's device from §4.1.4.

4.2.2 Using y in the construction of the distance matrix

In principle, failures of model (4.1) could involve any of the predictors in the model. With just a few predictors, all of them could be used to define the distance. In other cases, it will often seem reasonable to bet that failures of model (4.1) involve predictors that exhibit some predictive power in the fit of model (4.1). For instance, this might be true if either $\mathbf{y}$ or a predictor requires a monotone increasing transformation, or if two important predictors require inclusion of their interaction.

Tukey's method in §4.1.4 permits the matrix $\mathbf{L}$ to be any function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$. In particular, Tukey's method yields a central F-distribution for the test statistic if $\mathbf{L}$ is built from an optimal nonbipartite match using a distance matrix that is itself a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$. One such very simple distance matrix has as a distance between individuals i and i' the absolute difference in their predicted values, $|\widehat{y}_i - \widehat{y}_{i'}|$. However, in addition to matching for $\widehat{\mathbf{y}}$, it makes sense to also match for several of

the most important predictors.

Define d_j to be the square root of the jth diagonal element of $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$. The usual t-statistic testing the hypothesis that the jth coordinate β_j of $\boldsymbol{\beta}$ in (4.1) is 0 is $\widehat{\beta}_j / \left(d_j \, \widehat{\sigma}\right)$ where $\widehat{\beta}_j$ is the jth coordinate of $\widehat{\boldsymbol{\beta}}$. This t-statistic depends on $\widehat{\sigma}$, so it depends on the residuals $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$ and not just on the fitted values, $\widehat{\mathbf{y}}$, so the generalization of Tukey's method in §4.1.4 does not permit the use of this t-statistic. In contrast, the quantity $\widehat{\beta}_j / d_j$ is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$ because of (4.2). Because $\widehat{\sigma}$ in the t-statistic $\widehat{\beta}_j / \left(d_j \, \widehat{\sigma}\right)$ is the same for all predictors, we may identify the r predictors with the largest absolute t-statistics as the r predictors with the largest $\left|\widehat{\beta}_j / d_j\right|$, which is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$. To emphasize, we can use the generalization of Tukey's method if we select a fixed number, r, of variables with the largest t-statistics because we can identify those variables using $\mathbf{X}$ and $\widehat{\mathbf{y}}$, but we cannot select all variables with, say, $\left|\widehat{\beta}_j\right| / \left(d_j \, \widehat{\sigma}\right) \geq 2$, because that makes use of $\widehat{\sigma}$.

The proposed test computes the robust Mahalanobis distance from $\widehat{\mathbf{y}}$ and the r predictors with the largest $\left|\widehat{\beta}_j / d_j\right|$. Here, $\widehat{\mathbf{y}}$ depends upon all predictors. Because this distance is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$, as noted above, $\mathbf{L}$ too is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$, so the generalization of Tukey's (1949) method yields a null F-distribution for the F-statistic comparing models (4.1) and (4.3).

4.3 Simulation study of the power of the test

Tables 4.1 and 4.2 report simulated power of a 0.05-level test for five nonlinear functions with Gaussian errors. In additional simulations not shown in Tables 4.1 and 4.2, we found that a 0.05-level test did indeed reject a linear model in close to 5% of simulated samples. In the simulation, the model (4.1) is fit with a constant term included in $\mathbf{X}$, so there are $p' = p - 1$ predictors aside from the constant term. There

are $p' = 10$, 30 or 50 predictors and $n = 100$ or 500 observations; however, we do not consider the combination of $p' = 50$ predictors and $n = 100$ observations. In all cases, a linear model with p' predictors is mistakenly fit to various nonlinear surfaces, and the question is whether the test can recognize this mistake.

Each sampling situation is replicated 3000 times, so the standard error of a simulated power is at most $\sqrt{0.25/3000} < 0.01$. For $p' = 30$ or $p' = 50$, many of the predictors x_j, $j = 1, \ldots, p'$, do not affect the response surface, but the investigator does not know this, so the fitted model mistakenly uses all p' predictors. When $p' = 30$ or $p' = 50$, the test is looking for genuine model failures involving a few predictors amid distraction from many irrelevant predictors.

The test is performed in nine variations, $9 = 2 \times 5 - 1$. In five of the nine variations, the optimal nonbiparitite matching paired for $\widehat{y}$, and in four variations it did not. The optimal nonbiparitite matching paired for the r predictors with the largest absolute t-statistics, for $r = 0$, 3, 5, 10 and p'. One needs to pair for something, so the case of not pairing for $\widehat{y}$ and pairing for $r = 0$ predictors does not occur, making 9 variations in total. When $p' = 10$, the last two columns of Tables 4.1 and 4.2 are identical for $r = 10$ and $r = p'$. Two consecutive rows of Tables 4.1 and 4.2 — the first with 5 estimated powers, the second with 4 estimated powers — constitute one sampling situation in which 9 methods are competing to produce the largest power. In each sampling situation, the largest power or powers are in **bold**.

In matching, we use the robust Mahalanobis distance described in §4.2.1. As a consequence in Tables 4.1 and 4.2, matching for all predictors, $r = p'$, and matching for all predictors plus $\widehat{y}$ are slightly different. With the conventional Mahalanobis distance, $\widehat{y}$ would be linearly dependent on the constant plus p'-predictors and hence redundant, not affecting the distance.

The five nonlinear response surfaces will now be described. In Table 4.1, the

true nonlinear regression is $y = x_1 + x_2 + x_3 + x_4 + x_3x_4 + x_4x_5 + x_5^2 + \epsilon$ and the predictors are multivariate Normal with covariances indicated in the table. Three of the response surfaces in Table 4.2 were discussed and depicted by Friedman (1991) and have been used in the literature before and after 1991 as test cases of nonlinear regression surfaces. In Table 4.2, the predictors are independent uniform random variables, with standard Normal errors, and response surfaces given by:

Exponential: $y = \exp\left(\sum_{j=1}^{10} x_j\right) + \epsilon$

Friedman (1991), equation (56): $y = 0.1e^{4x_1} + 4/\left\{1 + e^{-20(x_2 - 0.5)}\right\} + 3x_3 + 2x_4 + x_5 + \epsilon$

Friedman (1991), equation (61): $y = 10\sin\left(()\,\pi x_1 x_2\right) + 20\left(x_3 - 0.5\right)^2 + 10x_4 + 5x_5 + \epsilon$

Friedman (1991), equation (66):

$$y = 40\frac{\exp\left\{8\left[(x_1 - 0.5)^2 + (x_2 - 0.5)^2\right]\right\}}{\exp\left\{8\left[(x_1 - 0.2)^2 + (x_2 - 0.7)^2\right]\right\}} + e^{8\left[(x_1 - 0.7)^2 + (x_2 - 0.2)^2\right]} + \epsilon.$$

All five models involve 10 or fewer variables, so that, when $p' > 10$, many of the variables are simply distractions, as noted above.

Consider, now, the estimated powers. The strongest pattern in Tables 4.1 and 4.2 is the least interesting: the power is higher when the sample size n is larger. Setting that aside, within a sampling situation or pair of rows, the power varies dramatically among nine methods.

Should one try to match for all p' variables? When $p' = 30$ or $p' = 50$, trying to match for all p' variables usually reduces power: it is better to use the mistaken linear fit to reduce the number of variables employed in the matching, even though the mistaken fit need not be a reliable guide to the importance or role of particular

Table 4.1: Simulated power of a 0.05-level test with $p' = p-1$ predictors and n observations, $y = x_1 + x_2 + x_3 + x_4 + x_3x_4 + x_4x_5 + x_5^2 + \epsilon$, where $(x_1, \ldots, x_p, \epsilon)$ is multivariate Normal, with $\mathrm{E}(x_j) = 0$, $\mathrm{var}(x_j) = 1$ and, except as noted below, $\mathrm{cov}(x_j, x_{j'}) = 0$, and with $\mathrm{E}(\epsilon) = 0$ $\mathrm{var}(\epsilon) = 1$, $\mathrm{cov}(\epsilon, x_j) = 0$. The matching either matched for $\widehat{y}$, case 1, or did not, case 0, and it matched for $r = 0$, 3, 5, 10, or all p' predictors with the largest absolute t-statistics. Each situation was replicated 3000 times. A sampling situation is two consecutive rows, and the highest power in a sampling situation is in **bold**.

Nonzero				Matched for r predictors				
Covariances	p'	n	Matched for $\widehat{y}$	0	3	5	10	p'
None	10	100	1	0.10	0.25	0.48	**0.50**	**0.50**
	10	100	0		0.19	0.48	0.48	0.48
Predictors are	10	500	1	0.13	0.64	0.87	**1.00**	**1.00**
independent	10	500	0		0.40	0.81	**1.00**	**1.00**
	30	100	1	0.07	0.11	0.13	**0.14**	0.10
	30	100	0		0.09	**0.14**	**0.14**	0.08
	30	500	1	0.12	0.56	0.79	**0.81**	0.74
	30	500	0		0.40	0.72	0.75	0.68
	50	500	1	0.13	0.50	**0.74**	0.71	0.49
	50	500	0		0.38	0.65	0.68	0.42
$\mathrm{cov}(x_1, x_5) = 0.8$	10	100	1	0.20	0.49	**0.63**	0.55	0.55
	10	100	0		0.34	0.62	0.45	0.45
Nonlinear x_5	10	500	1	0.47	0.99	**1.00**	**1.00**	**1.00**
is highly correlated	10	500	0		0.73	**1.00**	**1.00**	**1.00**
with linear	30	100	1	0.09	0.16	**0.18**	0.14	0.10
predictor x_1	30	100	0		0.12	**0.18**	0.13	0.08
	30	500	1	0.42	0.96	**1.00**	0.98	0.83
	30	500	0		0.68	1.00	0.96	0.62
	50	500	1	0.36	0.93	**1.00**	0.95	0.55
	50	500	0		0.64	0.99	0.93	0.35
$\mathrm{cov}(x_5, x_6) = 0.8$	10	100	1	0.11	0.28	0.47	**0.53**	**0.53**
	10	100	0		0.23	0.41	0.51	0.51
Nonlinear x_5	10	500	1	0.17	0.75	0.99	**1.00**	**1.00**
is highly correlated	10	500	0		0.47	0.97	**1.00**	**1.00**
with irrelevant x_6	30	100	1	0.08	0.12	**0.15**	0.14	0.10
	30	100	0		0.10	**0.15**	0.13	0.09
	30	500	1	0.16	0.68	**0.98**	0.95	0.88
	30	500	0		0.44	0.96	0.91	0.79
	50	500	1	0.14	0.62	**0.96**	0.91	0.60
	50	500	0		0.41	0.94	0.87	0.45
$\mathrm{cov}(x_j, x_{j'}) = 0.5$	10	100	1	0.57	0.52	**0.59**	0.57	0.57
for all $j \neq j'$	10	100	0		0.52	0.57	0.49	0.49
	10	500	1	0.99	0.96	**1.00**	**1.00**	**1.00**
All predictors	10	500	0		0.93	0.99	**1.00**	**1.00**
are correlated.	30	100	1	0.20	**0.22**	0.21	0.16	0.12
	30	100	0		0.20	0.18	0.15	0.10
	30	500	1	0.97	0.94	**0.99**	0.95	0.91
	30	500	0		0.92	**0.99**	0.93	0.81
	50	500	1	0.94	0.93	**0.99**	0.93	0.70
	50	500	0		0.91	0.98	0.90	0.55

Table 4.2: Simulated power of a 0.05-level test with $p' = p-1$ predictors and n observations, for four nonlinear functions. The matching either matched for $\widehat{y}$, case 1, or did not, case 0, and it matched for $r = 0$, 3, 5, 10, or p' predictors with the largest absolute t-statistics. Covariates are independent uniform random variables. Each situation was replicated 3000 times. A sampling situation is two consecutive rows, and the highest power in a sampling situation is in **bold**.

Function				Matched for r predictors				
	p'	n	Matched for $\widehat{y}$	0	3	5	10	p'
Exponential	10	100	1	0.60	0.71	0.65	**0.99**	**0.99**
	10	100	0		0.36	0.53	0.69	0.69
10 active	10	500	1	0.90	0.99	0.97	**1.00**	**1.00**
predictors	10	500	0		0.41	0.74	**1.00**	**1.00**
	30	100	1	0.38	0.48	0.40	**0.67**	0.56
	30	100	0		0.18	0.25	0.40	0.16
	30	500	1	0.93	**1.00**	0.99	**1.00**	**1.00**
	30	500	0		0.38	0.74	**1.00**	0.78
	50	500	1	0.93	**1.00**	0.99	**1.00**	**1.00**
	50	500	0		0.36	0.70	**1.00**	0.56
Friedman (1991, 56)	10	100	1	0.05	0.31	0.21	0.09	0.09
	10	100	0		**0.38**	0.22	0.09	0.09
5 active	10	500	1	0.05	**0.99**	0.95	0.56	0.56
predictors	10	500	0		0.98	0.96	0.48	0.48
	30	100	1	0.05	0.13	0.09	0.07	0.05
	30	100	0		**0.15**	0.10	0.07	0.05
	30	500	1	0.05	**0.98**	0.89	0.47	0.11
	30	500	0		0.97	0.94	0.42	0.09
	50	500	1	0.05	**0.97**	0.82	0.38	0.08
	50	500	0		0.94	0.89	0.37	0.07
Friedman (1991, 61)	10	100	1	0.06	0.46	0.47	0.34	0.34
	10	100	0		**0.50**	0.48	0.32	0.32
5 active	10	500	1	0.08	**1.00**	**1.00**	**1.00**	**1.00**
predictors	10	500	0		0.99	**1.00**	**1.00**	**1.00**
	30	100	1	0.06	0.16	0.15	0.10	0.06
	30	100	0		**0.18**	0.15	0.11	0.06
	30	500	1	0.06	**1.00**	**1.00**	0.90	0.51
	30	500	0		0.99	**1.00**	0.88	0.43
	50	500	1	0.07	**1.00**	0.99	0.82	0.25
	50	500	0		0.99	0.99	0.81	0.21
Friedman (1991, 66)	10	100	1	0.41	0.91	**0.93**	**0.93**	**0.93**
	10	100	0		0.80	0.74	0.52	0.52
2 active	10	500	1	0.54	**1.00**	**1.00**	**1.00**	**1.00**
predictors	10	500	0		0.97	0.96	0.85	0.85
	30	100	1	0.41	0.77	0.69	0.75	**0.79**
	30	100	0		0.62	0.59	0.60	0.22
	30	500	1	0.60	**1.00**	**1.00**	**1.00**	**1.00**
	30	500	0		0.97	0.97	0.91	0.61
	50	500	1	0.71	**1.00**	**1.00**	**1.00**	**1.00**
	50	500	0		0.97	0.97	0.93	0.50

variables. Also, matching for $\widehat{y}$ alone with $r = 0$ is not the best procedure in any sampling situation, and it often has low power.

How many variables should be used in the matching? In Table 4.1, only 5 variables affect the response surface, and good power often occurs when matching for either the $r = 5$ variables or $r = 10$ variables with the largest t-statistics. The situation is slightly more complicated in Table 4.2, where the first response surface involves 10 variables and $r = 10$ is best, while the last response surface involves just 2 variables and $r = 3$ or $r = 5$ is better than $r = 10$.

Should one match just for the r variables with the largest t-statistics or should one additionally match for $\widehat{y}$? In each sampling situation, this is the comparison of two adjacent rows in the same column. There is no uniform winner here, but including $\widehat{y}$ in a match for $r = 5$ or $r = 10$ variables rarely does much harm and sometimes greatly increases power. For instance, consider in Table 4.2 the Exponential model with $r = p' = 10$ where including $\widehat{y}$ yields power 0.99 and excluding $\widehat{y}$ yields power 0.69. Also in Table 4.2, consider Friedman's equation (66) model with $r = 10$, where again including $\widehat{y}$ in the matching distance increases power. In Table 4.1, five variables affect the response surface, so matching for $r = 3$ variables must omit relevant variables: in the $r = 3$ column of Table 4.1, matching also for $\widehat{y}$ often yields meaningful gains in power.

The Exponential response surface in Table 4.2 is interesting. In this case, $\mathrm{E}(y|\mathbf{x})$ is an increasing but nonlinear function of $x_1, \ldots, x_{10}$, whereas $x_{11}, \ldots, x_p$ are irrelevant. As might be expected, the highest powers occur with $r = 10$ including $\widehat{y}$ in the match. Even when $p' = 10$, so there are no irrelevant variables, it is still helpful to include $\widehat{y}$, presumably because a very high $\widehat{y}$ means most of $x_1, \ldots, x_{10}$ are high. For many values of r, omitting $\widehat{y}$ from the match for the Exponential surface can ruin the power. Matching for $\widehat{y}$ and $r = 5$ variables has lower power for the Exponential

surface than matching for $\widehat{y}$ and $r = 3$ variables for reasons that are not completely clear, but perhaps because $\widehat{y}$ gets more attention in the Mahalanobis distance with $r = 3$ variables than with $r = 5$ variables.

How close are the "near replicates" produced by matching? Before matching, there were $\binom{n}{2}$ distances in the distance matrix, whereas after matching there were $m \doteq n/2 - (n-p)/6$ distances within m pairs. For $n = 500$ observations with $p' = 50$ predictors, $p = p' + 1$, there were initially $\binom{500}{2} = 124,750$ pairwise distances and $m = 175$ within pair distances. How does the average distance within m pairs compare to the average of $\binom{n}{2}$ distances before matching? We computed the two averages, averaging also over 3000 simulations, and took the ratio. If we match for $\widehat{y}$ and r predictors, then the distance is computed among n points in $r+1$ dimensional space. Not surprisingly, if r is larger, the average distance after matching is a larger fraction of the average distance before matching: it is hard to find similar observations in high dimensions. Consider the case of $n = 500$ observations with $p' = 50$ predictors, matching for $\widehat{y}$ and r predictors in Tables 4.1 and 4.2. Among the eight such situations in Tables 4.1 and 4.2, the average distance within $m = 175$ pairs was never more than 2% of the average distance with $\binom{500}{2} = 124,750$ pairs if $r = 3$, was never more than 7% if $r = 5$, was never more than 19% if $r = 10$, and ranged from 54% to 57% for $r = p' = 50$. In other words, when trying to match for $r + 1 = 51$ variables, the matched pairs were closer than two observations picked at random, but the distance was reduced by less than half. This may partly explain why the power in Tables 4.1 and 4.2 is often higher when matching for $\widehat{y}$ and $r = 5$ predictors than when matching for $\widehat{y}$ and $r = 50$ predictors.

In brief, there is no uniformly best choice among our nine methods. We must choose a test in ignorance of the true response surface. For the admittedly limited situations we have considered, matching for $\widehat{y}$ plus the $r = 5$ variables with the largest

t-statistics would have been a tolerable choice in most cases given our ignorance of the true response surface, but $\widehat{y}$ plus the $r = 10$ variables is competitive, winning in many cases.

4.4 Example: testing fit without replicates in an experiment

Nelson (1981) discusses an experiment involving degradation of electrical insulation measured as y = dialectic strength in kV. There are two factors, duration of aging x_1 as 1, 2, 4, 8, 16, 32, 48, or 64 weeks, and the temperature x_2 as 180, 225, 250, 275 degrees Celsius. Nelson makes a physical argument for a particular nonlinear relationship, but for the purpose of illustrating our test of fit, we assume the investigator is unaware of this argument and ask whether our test will help the investigator discover this mistake. Each of the 8×4 factor combinations was replicated 4 times, making $8 \times 4 \times 4 = 128$ observations. Although the relationship between y and (x_1, x_2) is highly nonlinear, this is only very slightly apparent in the two bivariate plots of y versus x_1 and y versus x_2, so a careless investigator could fail to notice a serious problem. If one fits a Gaussian linear model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, and uses the four replicates to perform Fisher's test with exact replicates, then the linear model is rejected with a very small P-value. We adapted this example for illustration in two ways.

First, we created a smaller unreplicated design by randomly picking one replicate from each condition. This meant that each of the (x_1, x_2) combinations occurred once in an unreplicated design with $n = 8 \times 4 = 32$ observations. Here $\mathbf{X}$ has $n = 32$ rows and $p = 3$ columns, namely a constant and $p' = 2$ predictors . We then used an optimal nonbipartite matching based on $(\widehat{y}, x_1, x_2)$ to form $m = \lfloor n/2 - (n-p)/6 \rfloor =$

$\lfloor 32/2 - (32-3)/6 \rfloor = 11$ pairs and $n - 2m = 10$ isolated observations, so that $21 = 11+10$ predictors in $\mathbf{L}$ were added to the linear model, and $[\mathbf{X}, \mathbf{L}]$ had $24 = 3+11+10$ columns. We did this 10 times, randomly picking one replicate from the 4 available each time. In 8 of the 10 tests, the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ was rejected at the 0.05 level, despite the reduction in sample size from 128 to 32 and the absence of exactly replicated observations.

Second, we added 10 independent Gaussian noise predictors to the original $n = 128$ observation design so that the revised design was now unreplicated in terms of all 12 predictors, and $\mathbf{X}$ had $n = 128$ rows and $p = 13$ columns. We then used an optimal nonbipartite matching based on $\widehat{y}$ and the five predictors with the largest t-statistics to create 45 pairs and 38 isolated observations, adding $83 = 45+38$ predictors in $\mathbf{L}$ to the model, so $[\mathbf{X}, \mathbf{L}]$ had $96 = 13+83$ columns. Again, we did this 10 times, creating 10 different sets of noise predictors. All ten tests rejected $y = \beta_0+\beta_1 x_1+\ldots+\beta_{12} x_{12}+\epsilon$ at the 0.05 level.

4.5 Discussion: Summary; Alternative methods for selecting variables

We have been testing the fit of the Gaussian linear model (4.1) with n observations and $p' = p - 1$ predictors by: (i) determining $\widehat{y}$ and the r predictors x_j with the largest $\left|\widehat{\beta}_j/d_j\right|$, (ii) creating a distance matrix using these variables, (iii) using optimal nonbipartite matching to form roughly $n/3$ pairs and $n/3$ isolated observations, and (iv) determining whether these $2n/3$ additional predictors enhance the fit of model (4.1). Here, the r predictors x_j with the largest $\left|\widehat{\beta}_j/d_j\right|$ are also the r predictors x_j with the largest t-statistics, but $\left|\widehat{\beta}_j/d_j\right|$ does not yield the numerical value of the t-statistic; that is, one cannot select the predictors with absolute t-statistics above 2.

This is an exact test: when (4.1) is true, the probability of false rejection at nominal level α is $\leq \alpha$ (in fact, the test has size α). A key element is the generalization of Tukey's device in which functions of $\widehat{\mathbf{y}}$ and $\mathbf{X}$ may be used in test of fit of (4.1). We have been using functions of $\widehat{\mathbf{y}}$ and $\mathbf{X}$ to create near-replicate observations, yielding an estimate of error based on paired observations less affected by model misspecification. In simulations, matching for both $\widehat{\mathbf{y}}$ and $r = 5$ or $r = 10$ predictors with large $\left|\widehat{\beta}_j/d_j\right|$ gave good results for several nonlinear response surfaces with either few or many irrelevant predictors.

The proposed test of fit is not a substitute for other diagnostic checks of (4.1). In particular, one should check for outliers and for non-Gaussian errors. A single outlier, if sufficiently severe, can greatly reduce the power of an F-test, including specifically the test of (4.1) against (4.3).

There are many related methods that might be considered. For instance, if the p' predictors in (4.1) are highly correlated, it might not be wise to select r predictors x_j for the distance using $\left|\widehat{\beta}_j/d_j\right|$ from the full p' variable model, because an important predictor might have a small value of $\left|\widehat{\beta}_j/d_j\right|$ due to its high correlation with other predictors. Could we, instead, use Mallows' C_P to select r variables x_j for the distance? As with t-statistics, the numerical value of C_P depends on $\widehat{\sigma}^2$, so one cannot use the numerical value of C_P, as one cannot use the numerical value of the t-statistic, if one is going to employ the Tukey-Milliken-Graybill device to obtain an exact test. Consider the $\binom{p'}{r}$ submodels $P \subseteq \{1, \ldots, p'\}$ of (4.1) that involve exactly r of the p' predictors. Write $\mathbf{X}_P$ for the $n \times (r+1)$ matrix obtained from $\mathbf{X}$ by retaining the constant and the r columns in P, and write $\mathbf{H}_P = \mathbf{X}_P \left(\mathbf{X}_P^T \mathbf{X}_P\right)^{-1} \mathbf{X}_P$ so the predicted values from model P are $\widehat{\mathbf{y}}_P = \mathbf{H}_P \mathbf{y}$. It is readily checked that $\widehat{\mathbf{y}}_P$ is a function of $\mathbf{X}$ and the predicted values $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ from the full model (4.1); specifically, $\widehat{\mathbf{y}}_P = \mathbf{H}_P \widehat{\mathbf{y}}$.

Also, the usual C_P may be rewritten (Mallows, 1973, §1) as:

$$C_P = \frac{\left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)^T \left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)}{\widehat{\sigma}^2} - p + 2\left(r+1\right). \tag{4.4}$$

Now, (4.4) depends in a fundamental way upon $\widehat{\sigma}^2$ to compare models with different numbers of variables. However, if one restricts attention to models with exactly r predictors, then the model P with the smallest C_P and r predictors is the model with the smallest $\left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)^T \left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)$ and r predictors, so this model can be determined from $\widehat{\mathbf{y}}$ and $\mathbf{X}$ alone, so the Tukey-Milliken-Graybill device may be used. In brief, instead of selecting for the distance the r variables with the largest $\left|\widehat{\beta}_j / d_j\right|$, we may select the r variables in the r-variable model with the smallest $\left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)^T \left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)$, or equivalently in the r-variable model with the smallest C_P. The identification of this model may be based on the algorithm of Furnival and Wilson (1974), as implemented in the `R` package `leaps`.

Tables 1 and 2 consider matching for $\widehat{\mathbf{y}}$ and/or r predictors. To avoid focusing on r individual predictors, one might match for a few functions of $\widehat{\mathbf{y}}$. For instance, let $P \subseteq \{1, \ldots, p'\}$ be the model in the previous paragraph determined using C_P for some fixed $r < p'$, and let $\overline{P} = \{1, \ldots, p'\} - P$ be the $p' - r$ variables left out of this model. One could match for three variables, namely $\widehat{\mathbf{y}}$, $\widehat{\mathbf{y}}_P = \mathbf{H}_P \widehat{\mathbf{y}}$, and $\widehat{\mathbf{y}}_{\overline{P}} = \mathbf{H}_{\overline{P}} \widehat{\mathbf{y}}$. This would avoid the impossible task of matching in high dimensions while permitting the ostensibly less important variables in $\overline{P}$ to contribute meaningfully to the match distance.

The proposed method forms roughly $n/3$ pairs and $n/3$ isolated observations, yielding roughly $n/3$ degrees of freedom for the within-pair estimate of error. With this structure, the simulation found good power when matching for $r = 5$ or $r = 10$ predictors with $n = 100$ or $n = 500$ observations. If n were much larger than 500, say $n = 30,000$ for data from an administrative database, then one might reconsider these

choices. In particular, one might prefer fewer than $n/3$ pairs that match for more than $r = 10$ predictors. There is little value in having $30,000/3 = 10,000$ degrees of freedom for error, rather than a much smaller number of degrees of freedom; that is, we might be happy with fewer than $n/3$ pairs. On the other hand, think about cutting $r = 20$ predictors each at their median to form two categories per variable; then, there would be 2^{20} or about a million very coarse categories, and 30,000 people would be thinly spread among a million categories. In brief, having $n = 30,000$ rather than $n = 500$ has a big impact on degrees of freedom, but only a small impact on our ability to match for $r = 20$ predictors, so we might wish to have far fewer than $n/3$ pairs that are more closely matched for additional predictors. For instance, for large n, one might set a requirement for the distance, letting that requirement determine the number of pairs. If two multivariate observations are drawn independently from the same r-dimensional multivariate Gaussian distribution, then the Mahalanobis distance between them is distributed as two-times a chi-square random variable with r degrees of freedom; hence, the expected distance is $2r$. If κ is the ζ-quantile of the chi-square distribution on r degrees of freedom, then with probability ζ this Mahalanobis distance is less than 2κ. For $r = 20$ predictors and $\zeta = 0.05$, the expected Mahalanobis distance is $2r = 40$ and the 5% point is $2\kappa = 21.7$. One strategy for very large n would be to solve the maximum cardinality matching problem (Korte et al., 2008, §10.5): find the maximum number of disjoint pairs such that the Mahalanobis distance within every matched pair is at most 2κ.

It is virtually impossible to find many observations that are very close on many predictors; see Giraud (2015, §1.2). In light of this, the proposed procedure bets that lack of fit will involve predictors that appear to matter when the possibly mistaken model (4.1) is fitted, and, as the discussion above indicates, there are several if not many options for identifying these predictors. In all cases, the test has its nominal

level α, because the level is computed assuming model (4.1) is true. However, the power of the test is affected by whether the bet is correct. It is easy to construct examples in which the bet is mistaken, and the power is low, because subtle nonlinearity in some predictor gives the false impression that the predictor is unimportant, hence not included in the matching algorithm. In light of this, the test should be seen as a test of fit of (4.1) against alternatives in which the ostensibly active predictors enter the model in a misspecified form. This is a practical and interesting class of alternatives to (4.1), but it is far from exhaustive.

5 Conclusion

The three papers composing this thesis each highlight a different set of design goals and resulting strategies for forming matched sets from observational data. The first paper considers a setting in which it is possible to match individuals exactly on only a few of the many measured categorical covariates. However, as demonstrated in this chapter, when matching exactly within a small set of variables it may still possible to achieve a high level of marginal balance on the remaining measured covariates and their interactions, meaning that matched groups have similar overall distributions even when individual pairs do not share identical values. A new matching algorithm is presented which allows investigators to request marginal balance on variables in a prioritized manner, giving more attention to the most important variables first before addressing those of secondary importance. This is achieved by representing pair matching as a minimum-cost network flow optimization problem and defining a new set of constraints which can be imposed on the problem. In addition, the algorithm enjoys very attractive computational properties, due to the sparsity of the underlying network, and scales well to large problems. In the large-scale performance comparison between new and experienced surgeons that motivated the work, the matching algorithm succeeds admirably well in forming comparable matched samples of Medicare patients, achieving levels of balance on high-order interactions of categorical vari-

ables. In this case relaxing the impossible goal of matching exactly on all variables to an appropriate combination of some exact matching with marginal balance allows for efficient construction of attractive matches. Future methodological development of these matching techniques could include techniques to assess possible conflicts or tradeoffs between the two objectives (exact matching and balance), as well as an asymptotic analysis to clarify the effectiveness of marginal balance for categorical variables as sparse matches grow larger and larger.

The second paper considers the question of bias due to unmeasured confounding variables. In particular, it identifies a class of observed variables, called "prods" to receive treatment, which explain variance in treatment assignment but obey certain conditional independence relationships with unobserved confounders, and demonstrates that choosing not to match on these variables strictly reduces unmeasured bias when it is present. However, applying this result in practice is non-trivial, since it is often difficult to assess whether a particular variable satisfies the necessary conditions to be a prod. If an observed confounding variable is incorrectly labeled as a prod and ignored, the net result may actually be an increase in bias. This uncertainty introduces a tension between two different design goals: should an investigator aggressively ignore potential prods in the hope of reducing unmeasured biases, or should she instead match on any variable that is not absolutely known to be a prod to avoid biases due to observed discrepancies? A way of doing both, by forming two matched control groups with different criteria, is suggested. One control group is selected for similarity to the treated group on all observed variables, while the other is constructed with similarity on certain variables but in a way that either ignores or explicitly prioritizes substantial differences on potential prod variables. A multiple testing strategy is used in assessing evidence about the presence or absence of effects of treatment from the two treated-control comparisons, producing stronger evidence than would be derived

from either of the two individual comparisons alone. Here the use of multiple control groups offers a way to work with multiple design goals without needing to make an ad hoc a priori decision about which is most important. Further research could include the derivation of formulae for the degree of attenuation under specific models for treatment and outcome, and generalizations of the multiple-testing strategy to larger numbers of treated-control comparisons.

The third and final paper uses matching in a very different setting. Here matches are not treated-control pairs to be used for measuring effects of a treatment, but groups of similar observations to be treated as near-replicates in a lack-of-fit test for regression. Matched pairs are formed using an optimal nonbipartite matching routine which takes in a matrix of distances between all observations in the problem and minimizes the sum of within-pair distances. The main design challenge in this problem is selecting an appropriate covariate distance for the matching routine to optimize. The lack-of-fit test tends to perform best when observations are paired so that their expected outcomes under the true model are near-identical, but since the true model is not in general known this is a non-trivial objective to optimize. While distances such as the Mahalanobis distance attempt to pair individuals similar on all covariates with minimal modeling assumptions, they tend to perform poorly when many covariates are present, not forming pairs close on any individual covariate. To address these challenges, a guess is made that the important predictors in the true model are also important predictors in the null model, and that similarity in expected mean outcome under the null model predicts similarity in expected mean outcome under the true model. Guided by this assumption, variable selection is conducted before forming matching distances, selecting the covariates most predictive in the null model, and the fitted values from the null model are also incorporated into the matching distance. The resulting test is exact despite the use of the old fit in constructing the test

statistics, and simulations bear out its effective performance under a wide variety of model misspecification settings. Principled choices about which variables to use for matching in this setting result in important improvements in test performance, much as judicious focus and careful design in the previous two chapters enhanced the extraction of evidence about treatment effects from observational data. Extensions to this work could include the consideration of more general nonbipartite grouping algorithms that form not just pairs but potentially larger clusters of observations, or the use of more sophisticated dimension reduction techniques on the covariates prior to matching.

A

Appendices

A.1 Proofs of main results in Chapter 2

Proof of Lemma 4: Suppose there is a feasible flow f for the network $(\mathcal{N}, \mathcal{E})$ and define $\mathcal{M} = \{(\tau_t, \kappa_c) \in \mathcal{E} : f\{(\tau_t, \kappa_c)\} = 1\}$. By the definition of $\mathcal{E}$ in §2.4.4, if $(\tau_t, \kappa_c) \in \mathcal{E}$ then $(\tau_t, \kappa_c) \in \mathcal{A}$. There is only one edge exiting from control $\kappa_c \in \mathcal{C} \subset \mathcal{N}$, namely $(\kappa_c, \lambda_{K\ell})$ for the category $K\ell$ to which κ_c belongs, and because f is feasible we have $0 \leq f(\kappa_c, \lambda_{K\ell}) \leq \texttt{cap}\{(\kappa_c, \lambda_{K\ell})\} = 1$, so either $f(\kappa_c, \lambda_{K\ell}) = 0$ or $f(\kappa_c, \lambda_{K\ell}) = 1$. If $f(\kappa_c, \lambda_{K\ell}) = 1$ then κ_c received its one unit of flow from a unique treated node $\tau_t \in \mathcal{T} \subset \mathcal{N}$. Moreover, because f is feasible and $\texttt{demand}(\tau_t) = -m$, it follows that $m = \sum_{\kappa_c \in \mathcal{C}} f(\tau_t, \kappa_c)$ for each $\tau_t \in \mathcal{T}$, so $\mathcal{M}$ is indeed an acceptable 1-to-m match $\mathcal{M}$ such that $(\tau_t, \kappa_c) \in \mathcal{M}$ implies $(\tau_t, \kappa_c) \in \mathcal{A}$. Conversely, suppose there is an acceptable 1-to-m match $\mathcal{M}$. Then, by the definition in §2.4.1 of an acceptable 1-to-m match, $(\tau_t, \kappa_c) \in \mathcal{M}$ implies $(\tau_t, \kappa_c) \in \mathcal{A}$. For $\tau_t \in \mathcal{T}$ and $\kappa_c \in \mathcal{C}$ define $f(\tau_t, \kappa_c) = 1$ if $(\tau_t, \kappa_c) \in \mathcal{M}$ and $f(\tau_t, \kappa_c) = 0$ otherwise. By the definition of an acceptable 1-to-m match, each treated unit $\tau_t \in \mathcal{T}$ issues m units of flow, $m = \sum_{\kappa_c \in \mathcal{C}} f(\tau_t, \kappa_c)$, so (2.2) is satisfied for $n = \tau_t$. By the definition of an acceptable 1-to-m match, each control κ_c is matched to at most one treated unit τ_t, so $1 \geq \sum_{\tau_t \in \mathcal{T}} f(\tau_t, \kappa_c)$ for each $\kappa_c \in \mathcal{C}$,

and the zero or one unit of flow leaving κ_c may be passed through $(\kappa_c, \lambda_{K\ell}) \in \mathcal{E}$ with its capacity of $\mathtt{cap}\{(\kappa_c, \lambda_{K\ell})\} = 1$. The indirect paths in triangles, $\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right)$ and $\left(\lambda_{k\ell}^{'}, \lambda_{k\ell}^{''}\right)$, have infinite capacity, so all of the flow reaching $\lambda_{k\ell}$ may feasibly be passed on to the corresponding $\lambda_{k-1,\ell'}$ and on to the sink ω, so a feasible flow f may be completed by passing flow along indirect paths. ■

Proof of Lemma 5: Compute $\beta_{k\ell}$ in (2.1) for match $\mathcal{M}$ recalling that $0 = \sum_{\ell=1}^{L_k} \beta_{k\ell}$ for $k = 1, \ldots, K$. Write $\beta_{k\ell}^{+} = \max(0, \beta_{k\ell}) \geq 0$ and $\beta_{k\ell}^{-} = \max(0, -\beta_{k\ell}) \geq 0$ so that $\sum_{\ell=1}^{L_k} \beta_{k\ell}^{+} = \sum_{\ell=1}^{L_k} \beta_{k\ell}^{-}$ and $\sum_{\ell=1}^{L_k} |\beta_{k\ell}| = \sum_{\ell=1}^{L_k} \beta_{k\ell}^{+} + \sum_{\ell=1}^{L_k} \beta_{k\ell}^{-} = 2\sum_{\ell=1}^{L_k} \beta_{k\ell}^{-}$ or equivalently $\sum_{\ell=1}^{L_k} \beta_{k\ell}^{-} = \sum_{\ell=1}^{L_k} |\beta_{k\ell}| /2$. The total cost of f is the sum of the costs in two disjoint subsets of edges of $(\mathcal{N}, \mathcal{E})$; namely,

$$\sum_{e\in\mathcal{E}} f(e)\ \mathtt{cost}(e) = \sum_{e\in\mathcal{A}} f(e)\ \mathtt{cost}(e) + \sum_{e\in\mathcal{E}-\mathcal{A}} f(e)\ \mathtt{cost}(e).$$

The total cost of f over $\mathcal{A} \subset \mathcal{E}$, namely $\sum_{(\tau_t,\kappa_c)\in\mathcal{A}} f\{(\tau_t, \kappa_c)\}\ \mathtt{cost}\{(\tau_t, \kappa_c)\}$ is precisely $\sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc}$ by the definition of f. The remaining cost of the flow f is $\sum_{e\in\mathcal{E}-\mathcal{A}} f(e)\ \mathtt{cost}(e)$, and in $\mathcal{E} - \mathcal{A}$ there is nonzero cost only from edges of the form $\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right)$ in the indirect paths in triangles because $\mathtt{cost}\left(\lambda_{k\ell},\ \lambda_{k\ell}^{''}\right) = \mathtt{cost}\left(\lambda_{k\ell}^{'},\ \lambda_{k\ell}^{''}\right) = 0$. The number of units of flow entering the triangle defined by $\lambda_{k\ell}$, $\lambda_{k\ell}^{'}$, and $\lambda_{k\ell}^{''}$ (through node $\lambda_{k\ell}$), is given by $|\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_k(\kappa_c) = \lambda_{k\ell}\}|$. Since $\mathtt{cap}\left\{\left(\lambda_{k\ell}, \lambda_{k\ell}^{''}\right)\right\} = m \times d_{k\ell}$ also, we know from (2.1) that at least $\beta_{k\ell}^{-}$ units of flow pass through $\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right)$ with total cost $f\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right)\ \mathtt{cost}\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right) = f\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right)\ \Upsilon^{K-k+1}$. This yields the inequality (2.3). In a minimum cost feasible flow, $f\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right) = \beta_{k\ell}^{-}$ as $f\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right) > \beta_{k\ell}^{-}$ pointlessly increases the cost. This proves the case of equality in (2.1) for a minimum cost flow. ■

Proof of Theorem 6: Because the specific value of $\Upsilon > 1$ is not relevant for feasibility, the parts of the proposition that discuss existence merely restate Lemma 4. Fix $\Upsilon > mTK + \sum_{(\tau_t,\kappa_c)\in\mathcal{A}} \delta_{tc}$. With this Υ, let f be a minimum cost feasible flow in

$(\mathcal{N}, \mathcal{E})$, and let $\mathcal{M} = \{(\tau_t, \kappa_c) \in \mathcal{A} : f\{(\tau_t, \kappa_c)\} = 1\}$ be the corresponding acceptable 1-to-m match. Let $\beta_{k\ell}$ be the imbalances (2.1) for the match $\mathcal{M}$. The triangle defined by $\lambda_{k\ell}$, $\lambda_{k\ell}^{'}$, and $\lambda_{k\ell}^{''}$ receives $|\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_k(\kappa_c) = \lambda_{k\ell}\}|$ units of flow entering $\lambda_{k\ell}$, and so from the proof of Lemma 5, $f\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right) = \beta_{k\ell}^{-} = \max(0, -\beta_{k\ell}) \geq 0$ with a cost of $f\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right) \texttt{cost}\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right) = f\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right) \Upsilon^{K-k+1} = \beta_{k\ell}^{-} \Upsilon^{K-k+1}$. The cost of f is $\sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc} + \sum_{k=1}^{K} \Upsilon^{K-k+1} \sum_{\ell=1}^{L_k} |\beta_{k\ell}|/2$ by Lemma 5. Because the total flow is only mT, for each k we have $\sum_{\ell=1}^{L_k} f\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right) \leq mT$. Because $\mathcal{M} \subset \mathcal{A}$, we have $\sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc} \leq \sum_{(\tau_t,\kappa_c)\in\mathcal{A}} \delta_{tc}$. We now use these to bound the total cost of f strictly before all of the triangles defined by $\lambda_{k\ell}$, $\lambda_{k\ell}^{'}$, and $\lambda_{k\ell}^{''}$, that is,

$$\sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc} + \sum_{j=k+1}^{K} \Upsilon^{K-j+1} \sum_{\ell=1}^{L_j} |\beta_{j\ell}|/2$$

$$= \sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc} + \sum_{j=k+1}^{K} \Upsilon^{K-j+1} \sum_{\ell=1}^{L_j} f\left(\lambda_{j\ell}, \lambda_{j\ell}^{'}\right) \leq \sum_{(\tau_t,\kappa_c)\in\mathcal{A}} \delta_{tc} + mTK\,\Upsilon^{K-k} \quad \text{(A.1)}$$

$$\leq \left\{ \sum_{(\tau_t,\kappa_c)\in\mathcal{A}} \delta_{tc} + mTK \right\} \Upsilon^{K-k} < \Upsilon \times \Upsilon^{K-k} = \Upsilon^{K-k+1}, \quad \text{(A.2)}$$

where (A.1) uses the two upper bounds, the first inequality in (A.2) simply uses $\Upsilon^{K-k} \geq 1$, and the second inequality in (A.2) uses $\Upsilon > mTK + \sum_{(\tau_t,\kappa_c)\in\mathcal{A}} \delta_{tc}$. The cost of each single unit of flow passing through any edge $\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right)$ is Υ^{K-k+1}, and from (A.2) it exceeds the total cost of everything before $\left(\lambda_{k\ell}, \lambda_{k\ell}^{'}\right)$ in $(\mathcal{N}, \mathcal{E})$. Using (A.1)-(A.2) with $k = 1$ shows that it is not possible to further reduce $\sum_{\ell=1}^{L_1} |\beta_{1\ell}|$, because if any feasible flow f' had a lower value of $\sum_{\ell=1}^{L_1} |\beta_{1\ell}|$ then f' would have a lower total cost than f, and this is not possible because f is a minimum cost flow. Similarly, it is not possible to further reduce $\sum_{\ell=1}^{L_1} |\beta_{1\ell}|, \ldots, \sum_{\ell=1}^{L_k} |\beta_{k\ell}|$ for the same reason: even a 1 unit reduction in any of these quantities would reduce the cost by at least Υ^{K-k+1}, and this is greater than the total cost of all flow routing decisions

made before the $\lambda_{k\ell} \in \mathcal{N}$, so this would (impossibly) reduce the cost of a minimum cost flow. In short, the match $\mathcal{M}$ from a feasible minimum cost flow f exhibits refined balance in the sense of Definition 2. A match achieving refined balance in Definition 2 must, by virtue of this definition, have achieved the smallest possible value of $\sum_{k=1}^{K} \Upsilon^{K-k+1} \sum_{\ell=1}^{L_k} |\beta_{k\ell}| /2$, and in particular $\mathcal{M}$ has done this; moreover, $\mathcal{M}$ has minimized $\sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc} + \sum_{k=1}^{K} \Upsilon^{K-k+1} \sum_{\ell=1}^{L_k} |\beta_{k\ell}| /2$, so it has minimized $\sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc}$ among all 1-to-m acceptable matches with refined balance. ■

A.2 Formal description of matching algorithm in Chapter 3

A.2.1 Introduction

In Section 3.5.3 above, we briefly describe an algorithm that can balance treated and control groups closely on certain covariates $\overline{\mathbf{x}}$ while separating them on others $\widetilde{\mathbf{x}}$. This supplement provides a full technical specification of this algorithm. In Section A.2.2 below, the general algorithm is described and its optimality is proven, using concepts and notation from Chapter 2. In Section A.2.3, further detail is given about how the general algorithm can be fine-tuned to create better separation on a prod. This section also gives specifics about how the second control group was created in the NHANES example given in Chapter 3.

A.2.2 Matching to a different target distribution

The large, sparse matching algorithm of Chapter 2 requires that balance covariates $\nu_1, \nu_2, \ldots, \nu_K$ (given in decreasing order of importance) be nested within each other, i.e. all categories of ν_j are finer subdivisions of the categories of ν_{j-1}. In practice the covariates ν_i are often interactions of many nominal covariates measured in the dataset. The algorithm computes an optimal match by formulating the task as a network flow problem. Flow constraints in certain edges of the network are set based on the empirical covariate distribution of the treated units, and require the covariate distribution of the controls to be as close as possible to this distribution. In this technical sense the treated group provides the "target distribution" to which the selected control will be made similar. We wish to modify this algorithm so that a different target distribution can be used.

Formally, we transform the algorithm as follows. Here we adopt the notation of Section 2.4.1. In the original algorithm there was a treated group $\mathcal{T}$ and a control group $\mathcal{C}$. Define a third group $\mathcal{T}' = \{\tau'_1, \tau'_2, \ldots, \tau'_T\}$ where $T = |\mathcal{T}|$ and call it the target group. We also extend the domain of each nested covariate ν_k to include $\mathcal{T}'$ so now $\nu_k : \mathcal{T} \cup \mathcal{C} \cup \mathcal{T}' \longrightarrow \mathcal{K}_k$; in other words, the units in the target group take values for each of the nested covariates. We now alter the algorithm by changing the definition of the quantities $d_{k\ell}$ for $\ell = 1, \ldots, L_k$ and $k = 1, \ldots, K$. In the original algorithm, these are defined as:

$$d_{k\ell} = |\{\tau_t \in \mathcal{T} \ : \ \nu_k(\tau_t) = \lambda_{k\ell}\}|$$

In short, $d_{k\ell}$ counts the number of individuals in category ℓ of covariate k in the treated group $\mathcal{T}$. We change the definition so that instead $d_{k\ell}$ is equal to the number of individuals in category ℓ of covariate k in the target group $\mathcal{T}'$:

$$d_{k\ell} = |\{\tau'_t \in \mathcal{T}' \ : \ \nu_k(\tau'_t) = \lambda_{k\ell}\}|$$

This mainly affects the algorithm through the quantities $\beta_{k\ell}$, which give the covariate imbalance at a particular category and are defined as follows:

$$\beta_{k\ell} = m \times d_{k\ell} - |\{(\tau_t, \kappa_c) \in \mathcal{M} \ : \ \nu_k(\kappa_c) = \lambda_{k\ell}\}|$$

These $\beta_{k\ell}$ terms are used in Definition 2 of Section 2.4.2 to define refined covariate balance. So in changing the $d_{k\ell}$ values we not only transform the algorithm but broaden the definition of refined covariate balance, so that balance is now with respect to a particular target distribution $\mathcal{T}'$. This leads to the following proposition;

Proposition 12 *Given a target group $\mathcal{T}'$, if we alter the $d_{k\ell}$ values and associated*

$\beta_{k\ell}$ values as outlined above to obtain a new algorithm and a new definition of refined covariate balance, then the new algorithm produces an optimal match with refined covariate balance with respect to $\mathcal{T}'$.

Proof. The proof is identical to the optimality proof for the original algorithm, except that we use the new definition for $d_{k\ell}$ and the resulting new definition of $\beta_{k\ell}$. ■

Notice that the new version of the proof includes the old version as the special case when $\mathcal{T} = \mathcal{T}'$. However, it also shows the optimality of the modified algorithm for matching under refined covariate balance with respect to any empirical covariate distribution generated by T observations on $\nu_1 \times \nu_2 \times \ldots \times \nu_K$.

A.2.3 Creating better separation on a prod

The balance constraints for the large, sparse optimal matching algorithm are described by the decreasingly-important, increasingly-fine nominal covariates $\nu_1, \nu_2, \ldots, \nu_K$. When matching to create separation, these covariates could be formed by relevant functions and interactions of $\overline{\mathbf{x}}$ and $\eta(\widetilde{\mathbf{x}})$ (where η is defined as in Section 5.3 of the main paper). As in the original version of large, sparse matching with refined covariate balance, the best choice of K and of the nested covariates $\nu_1, \ldots, \nu_K$ is highly application- and data-dependent, and researchers may need to experiment with several different configurations to obtain acceptable balance results.

To improve observed separation on $\widetilde{\mathbf{x}}$, the researcher may find it useful to define balance constraints not just in terms of the single function of $\widetilde{\mathbf{x}}$ described by η but in terms of a series of such constraints $\eta_1, \ldots, \eta_J$. For example, one might define a series of J sets $\mathcal{X}'_j \subset \mathcal{X}$ such that $\mathcal{X}'_1 \subset \mathcal{X}'_2 \subset \ldots \subset \mathcal{X}'_J$ where $\mathcal{X}'_1$ is the region from which the researcher would most like controls to be selected and $\mathcal{X}'_2, \ldots, \mathcal{X}'_J$ are regions from which to select the controls if this is not possible, in decreasing order of preference.

Then one could define new covariates

$$\eta_j(\widetilde{\mathbf{x}}_i) = \begin{cases} 1 & \text{if } \widetilde{\mathbf{x}}_i \in \mathcal{X}_j \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1, \ldots, J$ and set the values of each $\eta_j(\widetilde{\mathbf{x}}_i)$ to 1 in the target distribution. These covariates and their interactions with $\overline{\mathbf{x}}$ would grant the researcher greater flexibility in defining the balance constraints $\nu_1, \ldots, \nu_K$ and might lead to better combinations of $\widetilde{\mathbf{x}}$-separation and $\overline{\mathbf{x}}$-balance.

In the NHANES example of Section 3.1.2, we used the 5-level ordinal measure of education and the continuous measure of socioeconomic status to define the following desirable regions from which to draw controls:

$$\mathcal{X}_1 = \text{income-to-poverty ratio above 2}$$

$$\mathcal{X}_2 = \text{income-to-poverty ratio above 2, high school graduate}$$

$$\mathcal{X}_3 = \text{income-to-poverty ratio above 4, some college}$$

$$\mathcal{X}_4 = \text{income-to-poverty ratio above 4, college graduate}$$

We then enforced balance on a series of interactions of the resulting variables $\eta_1(\widetilde{\mathbf{x}})$, $\ldots, \eta_4(\widetilde{\mathbf{x}})$ with the balance covariates $\overline{\mathbf{x}}$. We controlled for η_1 at an early, coarse level in the balance hierarchy (to ensure most controls had at least a moderate level of income) and added the other more stringent variables η_j at finer, less-prioritized levels in the hierarchy (to ensure better-educated and wealthier controls were chosen when available).

Bibliography

Ali, M. S., Groenwold, R. H., and Klungel, O. H. (2014). Propensity score methods and unobserved covariate imbalance. *Health services research*, 49(3):1074–1082.

Andrews, D. (1971). A note on the selection of data transformations. *Biometrika*, 58:249–254.

Arias, E. (2012). Us life tables, 2008. *National Vital Statistics Reports*, 61(3).

Arratia, R., Goldstein, L., and Gordon, L. (1990). Poisson approximation and the chen-stein method. *Statistical Science*, 5:403–434.

Baiocchi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492):1285–1296.

Baiocchi, M., Small, D. S., Yang, L., Polsky, D., and Groeneveld, P. W. (2012). Near/far matching: a study design approach to instrumental variables. *Health Services and Outcomes Research Methodology*, 12(4):237–253.

Bauer, P. and Kieser, M. (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika*, pages 934–937.

Bazzano, L. A., He, J., Muntner, P., Vupputuri, S., and Whelton, P. K. (2003). Relationship between cigarette smoking and novel risk factors for cardiovascular disease in the united states. *Annals of Internal Medicine*, 138(11):891–897.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300.

Bertsekas, D. P. (1991). *Linear network optimization: algorithms and codes*. MIT Press, Cambridge, MA.

Bertsekas, D. P., Tseng, P., et al. (1994). *RELAX-IV: A faster version of the RELAX code for solving minimum cost flow problems*. Massachusetts Institute of Technology, Laboratory for Information and Decision Systems Cambridge, MA.

Bhattacharya, J. and Vogt, W. B. (2012). Do instrumental variables belong in propensity scores?

Box, G. E. and Draper, N. (1982). Measures of lack of fit for response surface designs and predictor variable transformations. *Technometrics*, 24(1):1–8.

Brooks, J. M. and Ohsfeldt, R. L. (2013). Squeezing the balloon: propensity scores and unmeasured covariate balance. *Health services research*, 48(4):1487–1507.

Burman, C.-F., Sonesson, C., and Guilbaud, O. (2009). A recycling framework for the construction of bonferroni-based multiple tests. *Statistics in medicine*, 28(5):739–761.

Campbell, D. T. (1969). Prospective: Artifact and control. In Rosenthal, R. and Rosnow, R. L., editors, *Artifacts in Behavioral Research*, pages 351–382. Academic Press, New York, NY.

Christensen, R. (1989). Lack-of-fit tests based on near or exact replicates. *The Annals of Statistics*, 17:673–683.

Christensen, R. (1991). Small-sample characterizations of near replicate lack-of-fit tests. *Journal of the American Statistical Association*, 86(415):752–756.

Christensen, R. (2011). *Plane answers to complex questions: the theory of linear models*. Springer Science & Business Media, New York, NY.

Christensen, R. and Utts, J. (1992). Testing for nonadditivity in log-linear and logit models. *Journal of statistical planning and inference*, 33(3):333–343.

Cook, W. J., Cunningham, W. H., Pulleyblank, W. R., and Schrijver, A. (1998). *Combinatorial optimization*, volume 605. Wiley, New York, NY.

Daniel, C. and Wood, F. S. (1971). *Fitting Equations to Data: Computer Analysis of Multifactor Data for Scientists and Engineers*. John Wiley & Sons, Inc., New York, NY.

Daniel, S. R., Armstrong, K., Silber, J. H., and Rosenbaum, P. R. (2008). An algorithm for optimal tapered matching, with application to disparities in survival. *Journal of Computational and Graphical Statistics*, 17(4):914–924.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 4:1–31.

Derigs, U. (1988). Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research*, 13(1):225–261.

Diestel, R. (2010). *Graph Theory*. Graduate Texts in Mathematics, Volume 173. Springer, New York, NY, 4th edition.

Draper, N. R. (1982). Center points in second?order response surface designs. *Technometrics*, 24(2):127–133.

Fisher, R. A. (1922). *Statistical methods for research workers*. Oliver & Boyd, Edinburgh.

Fisher, R. A. (1935). The design of experiments. *The design of experiments*.

Freedman, D. A. (2008). Randomization does not justify logistic regression. *Statistical Science*, 23(2):237–249.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67.

Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics*, 16(4):499–511.

Giraud, C. (2015). *Introduction to High-Dimensional Statistics*, volume 138. CRC Press, Boca Raton, FL.

Goeman, J. J., Solari, A., and Stijnen, T. (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in medicine*, 29(20):2117–2125.

Green, J. (1971). Testing departure from a regression, without using replication. *Technometrics*, 13(3):609–615.

Hansen, B. (2007). Optmatch (r package optmatch). *R News*, 7:18–24.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, pages 481–488.

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

Heller, R., Manduchi, E., and Small, D. S. (2009). Matching methods for observational microarray studies. *Bioinformatics*, 25(7):904–909.

Heller, R., Rosenbaum, P. R., and Small, D. S. (2010). Using the cross-match test to appraise covariate balance in matched pairs. *The American Statistician*, 64(4):299–309.

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *Sociological Methodology*, 18:449–484.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Joglekar, G., Schuenemeyer, J. H., and LaRiccia, V. (1989). Lack-of-fit testing when replicates are not available. *The American Statistician*, 43(3):135–143.

Jungnickel, D. (2013). *Graphs, networks and algorithms*. Springer.

Keele, L., Titiunik, R., and Zubizarreta, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):223–239.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906.

Korte, B., Vygen, J., Korte, B., and Vygen, J. (2008). *Combinatorial optimization*. Springer-Verlag, Heidelberg.

Lange, K. (2003). *Applied probability*. Springer, New York, NY.

Laska, E. M. and Meisner, M. J. (1989). Testing whether an identified treatment is best. *Biometrics*, 45:1139–1151.

Lu, B. (2005). Propensity score matching with time-dependent covariates. *Biometrics*, 61(3):721–728.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011). Optimal nonbipartite matching and its statistical applications. *The American Statistician*, 65(1):21–30.

Mallows, C. L. (1973). Some comments on cp. *Technometrics*, 15:661–665.

Mandel, J. (1959). The measuring process. *Technometrics*, 1(3):251–267.

Maritz, J. (1979). A note on exact robust confidence intervals for location. *Biometrika*, 66:163–166.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of business & economic statistics*, 13(2):151–161.

Miller, F. R. and Neill, J. W. (2008). General lack of fit tests based on families of groupings. *Journal of Statistical Planning and Inference*, 138(8):2433–2449.

Miller, F. R., Neill, J. W., and Sherfey, B. W. (1999). Implementation of a maximin power clustering criterion to select near replicates for regression lack-of-fit tests. *Journal of the American Statistical Association*, 94(446):610–620.

Miller, F. R., Neill, J. W., Sherfey, B. W., et al. (1998). Maximin clusters for near-replicate regression lack of fit tests. *The Annals of Statistics*, 26(4):1411–1433.

Milliken, G. A. and Graybill, F. A. (1970). Extensions of the general linear hypothesis model. *Journal of the American Statistical Association*, 65(330):797–807.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., and Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, page kwr364.

Neill, J. W. and Johnson, D. E. (1985). Testing linear regression function adequacy without replication. *The Annals of Statistics*, 13:1482–1489.

Nelson, W. (1981). Analysis of performance-degradation data from accelerated tests. *IEEE Transactions on Reliability*, 30(2):149–155.

Neuman, M. D., Rosenbaum, P. R., Ludwig, J. M., Zubizarreta, J. R., and Silber, J. H. (2014). Anesthesia technique, mortality, and length of stay after hip fracture surgery. *Jama*, 311(24):2508–2517.

Neyman, J., Dabrowska, D. M., Speed, T. P., et al. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):463–480.

Pearl, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In Grunwald, P. and Spirtes, P., editors, *Proceedings of UAI*, pages 417–424.

Pearl, J. (2011). Understanding bias amplification. *American journal of epidemiology*, 174(11):1223–1227.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association*, 110(510):515–527.

Rao, C. R. (1973). *Linear statistical inference and its applications*. John Wiley & Sons, New York, NY.

Rosenbaum, P. R. (1984). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, 147:656–666.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74:13–26.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

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

Rosenbaum, P. R. (2002b). *Observational Studies*. Springer, New York, NY.

Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika*, 91(1):153–164.

Rosenbaum, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics*, 63(2):456–464.

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

Rosenbaum, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics*, 69(1):118–127.

Rosenbaum, P. R. (2015a). How to see more in observational studies: Some new quasi-experimental devices. *Annual Review of Statistics and Its Application*, 2:21–48.

Rosenbaum, P. R. (2015b). Some counterclaims undermine themselves in observational studies. *Journal of the American Statistical Association*, 110(512):1389–1398.

Rosenbaum, P. R. (2015c). Two r packages for sensitivity analysis in observational studies. *Observational Studies*, 1:1–17.

Rosenbaum, P. R. et al. (1987). The role of a second control group in an observational study. *Statistical Science*, 2(3):292–306.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, 102(477):75–83.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rosenbaum, P. R. and Silber, J. H. (2009a). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, 104(488):1398–1405.

Rosenbaum, P. R. and Silber, J. H. (2009b). Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *Journal of the American Statistical Association*, 104(486):501–511.

Rosenbaum, P. R. and Silber, J. H. (2013). Using the exterior match to compare two entwined matched control groups. *The American Statistician*, 67(2):67–75.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Scheffé, H. (1959). *The Analysis of Variance*. John Wiley, New York, NY.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin, Boston, MA.

Shillington, E. R. (1979). Testing lack of fit in regression without replication. *Canadian Journal of Statistics*, 7(2):137–146.

Silber, J. H., Rosenbaum, P. R., Clark, A. S., Giantonio, B. J., Ross, R. N., Teng, Y., Wang, M., Niknam, B. A., Ludwig, J. M., Wang, W., et al. (2013). Characteristics associated with differences in survival among black and white women with breast cancer. *Jama*, 310(4):389–397.

Silber, J. H., Rosenbaum, P. R., Kelz, R. R., Reinke, C. E., Neuman, M. D., Ross, R. N., Even-Shoshan, O., David, G., Saynisch, P. A., Kyle, F. A., et al. (2012). Medical and financial risks associated with surgery in the elderly obese. *Annals of Surgery*, 256(1):79–86.

St. Laurent, R. T. (1990). The equivalence of the milliken?graybill procedure and the score test. *The American Statistician*, 44(1):36–37.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Stuart, E. A. and Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, 33(3):279–306.

Su, Z. and Yang, S.-S. (2006). A note on lack-of-fit tests for linear models without replication. *Journal of the American Statistical Association*, 101(473):205–210.

Traskin, M. and Small, D. S. (2011). Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Statistics in Biosciences*, 3(1):94–118.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, 5(3):232–242.

Utts, J. M. (1982). The rainbow test for lack of fit in regression. *Communications in Statistics-Theory and Methods*, 11(24):2801–2815.

Walker, A. M. (2013). Matching on provider is risky. *Journal of clinical epidemiology*, 66(8):565–568.

Welch, B. L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika*, 29:21–52.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., et al. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, 98(8):1359–1366.

Wolfe, D. A. (1974). A characterization of population weighted-symmetry and related results. *Journal of the American Statistical Association*, 69(347):819–822.

Wooldridge, J. (2009). Should instrumental variables be used as matching variables? Technical report, Michigan State University, East Lansing, MI.

Yang, D., Small, D. S., Silber, J. H., and Rosenbaum, P. R. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics*, 68(2):628–636.

Zubizarreta, J. R., Neuman, M., Silber, J. H., and Rosenbaum, P. R. (2012). Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. *Journal of the American Statistical Association*, 107(499):901–915.

Zubizarreta, J. R., Paredes, R. D., Rosenbaum, P. R., et al. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics*, 8(1):204–231.

Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2011). Matching for several sparse nominal variables in a case-control study of readmission following surgery. *The American Statistician*, 65(4):229–238.

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S., Rosenbaum, P. R., et al. (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *The Annals of Applied Statistics*, 7(1):25–50.