Robust Testing for Causal Inference in Observational Studies

Md. Noor-E-Alam
Dept. of Mechanical and Industrial Engineering, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA
md.alam@neu.edu

Cynthia Rudin
MIT CSAIL and Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
rudin@mit.edu

A vast number of causal inference studies use matching techniques, where treatment cases are matched with similar control cases. For observational data in particular, we claim there is a major source of uncertainty that is essentially ignored in these tests, which is the way the assignments of matched pairs are constructed. It is entirely possible, for instance, that a study reporting an estimated treatment effect with P-value of 10^{-4} can be redone in almost the same way, with the same match quality, yet with a P-value well above 0.10, making the test result no longer significant. Experimenters often specifically choose *not* to consider the output in the design of the assignments; this allows for easier computation and clearer testing, but it does not consider possible biases in the way the assignments were constructed. What we would really like to be able to report is that *no matter* which assignment we choose, as long as the match is sufficiently good, then the hypothesis test result still holds. This will be called a *robust* matched pairs test, since its result is robust to the choice of the assignment. In this paper, we provide methodology based on discrete optimization to create these robust tests. This method explores the full variation of possible results one can obtain with all possible acceptable assignments. It demonstrates that one cannot always trust statistically significant results, even with a large number of matched pairs.

Key words: causal inference, observational studies, hypothesis test, matched pairs design, discrete optimization, integer programming.

History:

1. Introduction

As massive and varied amounts of observational data are accumulating in healthcare, internet marketing, and governance, these data are increasingly used for understanding important cause and effect relationships. We might want to know whether a policy causes people to use fewer public services, or we might want to know whether a particular drug causes a side effect, or whether a view of an internet advertisement results in an increased chance of purchase. Controlled trials with matching treatment and controlled populations are often small and expensive, and not often possible due to the ethics of offering the treatment (e.g., exposure to lead), or perhaps not possible due to the fact that the treatment happened only in the past (e.g., lived in Morristown in the 1960s). Typically, for large amounts of observational data, we can construct matched pairs of observations, where one received the treatment and the other did not. Observations are matched on a set of attributes (age, gender, race), propensity scores, etc. The quality of the matches can be very important – poor matching algorithms can lead to wrong conclusions, and potentially to harmful politics, laws, and medical decisions.

Classically, assignments of treatment and control units to matches are constructed using a fixed design (Rosenbaum 2010a), without regard to the outcome (Rubin 2007, 2008). From our perspective, this could be a major flaw in the current paradigm. Choosing a single fixed design ignores a major source of uncertainty, which is the design itself, or in other words, the uncertainty related to the choice of experimenter. What if there were two possible equally good matchings, one where the treatment effect estimate is very strong and one where it is nonexistent? When we report a result on a particular matching assignment, we thus ignore the possibility of the opposite result occurring on an equally good assignment. It is entirely possible that two separate researchers studying the same effect on the same data, using two different equally good sets of pairs, would get results that disagree.

Our goal is to create robust matched pairs hypothesis tests for observational data. These tests implicitly consider *all possible reasonably good assignments* and consider the *range of possible*

outcomes for tests on these data. This is a more computationally demanding approach to hypothesis testing than the standard approach where one considers just a single assignment, but the result would be robust to the choice of experimenter. It is computationally infeasible (and perhaps not very enlightening) to explicitly compute all possible assignments, but it is possible to look at the range of outcomes associated with them. In particular, our algorithms compute the maximum and minimum of quantities like the treatment effect estimate, P-value, and z-score among the set of reasonably good assignments.

Finding a set of matched pairs that obey certain conditions is purely a data mining problem – we aim to locate a particular pattern of data within a database. Similar subfields of data mining, where the goal is to locate optimal subsets of data, include association rule mining and event detection. For these types of problems, modern discrete optimization techniques have rarely been used, though there is one recent precedent in the literature for matched pairs, namely the line of work by Zubizarreta (2012), Zubizarreta et al. (2013, 2014). Optimization techniques have major advantages over other types of approaches: (i) they are extremely flexible and allow the experimenter to match on very complex conditions, such as quantiles of attributes, which network optimization methods for matching cannot handle; (ii) they can be computationally efficient, depending on the strength of the integer programming formulation – strong formulations have relaxations close to the set of feasible integer solutions, and (iii) mixed-integer linear programs have guarantees on the optimality of the solution – in fact, they produce upper and lower bounds on the value of the optimal solution.

In what follows, we provide three basic constructions for robust hypothesis tests, one for a matched pairs z-test on the difference between two means, the second one for a McNemar's test for proportion data (1:1 matched pair design), and the third one for a χ^2 test also for proportion data (1:m matched pair design). The z-test and χ^2 test create nonlinear optimization problems that (in our experiments) were not able to be solved with a MINLP (mixed integer nonlinear programming) solver within a reasonable amount of time. Instead, we propose algorithms that solve a series of integer linear programs (ILPs) along a coarse one-dimensional mesh and then at finer

one-dimensional meshes, until we achieve a solution with the desired precision. Computationally these methods are much more attractive, and scale to large datasets, as we demonstrate.

The remainder of this paper is organized as follows. Section 2 formalizes robust testing as our framework defines it. In Section 3, we propose an optimization-based approach to carry out the robust z-test for real valued outcome data. In Sections 4 and 5, we propose two optimization-based approaches for inference on a proportion, one for 1:1 pair matching and another one is for 1-to-many matching. Section 6 presents empirical results and analysis.

2. Matching for Robust Tests

This work concerns the potential outcomes framework (see Holland 1986, Rubin 1974). In our notation X represents a vector of covariates for an individual, and Y is an outcome variable that depends on X and whether the patient was treated, $Y(1, X)$ is the random variable for the outcome of a treated individual with covariates X, and $Y(0, X)$ is the outcome for an untreated individual. We make the classical SUTVA assumption (the treatment status of any unit does not affect the potential outcomes of the other units), and assume conditional ignorability, which is that Y is independent of treatment T given X. We also assume unconfoundedness. Our goal is to determine whether we can reject the claim that certain standard quantities are zero, such as:

$$\text{ATT}: \ \mathbb{E}_{(X|T=1),(Y|T=1,X)}[Y(1,X) - Y(0,X)|T=1]$$

$$\text{ATE}: \ \mathbb{E}_{(X,Y)}[Y(1,X) - Y(0,X)].$$

The distribution of X is different for the treatment and control groups, because there is a bias as to who receives the treatment. The distribution of $Y|0, X$ and $Y|1, X$ will be different if the treatment has an effect. To conduct our hypothesis test we have observations:

$$(\mathbf{x}_1^t, y_1^t), ..., (\mathbf{x}_T^t, y_T^t)$$

$$(\mathbf{x}_1^c, y_1^c), ..., (\mathbf{x}_C^c, y_C^c).$$

A matching operator determines which control is assigned to which treatment. For simplicity, we write an assignment operator that assigns at most one control to each treatment without

and wind speed are less or equal to 2, 5 and 5, respectively for treated unit i and control unit j, 0 otherwise.

Figures 1 through 3 show the upper and lower bounds for the maximum objective function value for different b_l with n=30. These figures illustrate the meshes at different scales within the algorithm for computing the optimal solution for the maximization problem. Figures 4 through 6 illustrate the same phenomenon for minimization of the objective function. To produce Figure 7, we used similar procedure for producing Figures 1-6 at different values of n, namely n=30, 50, 70, 90, 110. The problem of finding pairs becomes infeasible above $n = 110$. Then each z-score was translated into a P-value $1-\phi(z)$ for both the upper and lower bounds for each n. The P-value for the upper bounds are very close to 1 and the P-value for the lower bounds are close to 0, illustrating that there is a lot of uncertainty associated with the choice of experimenter – one experimenter choosing 90 matched pairs can find a P-value of ~ 0 and declare a statistically significant difference while another experimenter can find a P-value of ~ 1 and declare the opposite. In this case it is truly unclear whether or not mist has an effect on the total number of rental bikes.

A note of caution: if the experimenter simply chose the P-value for the largest set of matched pairs, this would have deterministically yielded a P-value of almost 1 for 110 matched pairs. This is potentially very misleading – if we asked for slightly fewer pairs the result becomes completely unclear. Essentially, this casts doubt on the robustness of this particular choice used in standard practice.

6.2. Case Study 2

In this case study we have used the data from a U.S. Department of Justice study regarding crime during the transition to adulthood, for youths raised in out-of-home care (see Courtney and Cusick 2010). Each observation represents a youth, and the outcome is whether he or she committed a violent crime over the 3 waves of the study.

The (binary) covariates are as follows: hispanic, white, black, other race, alcohol or drug dependency, mental health diagnosis, history of neglect, entry over age of 12, entry age under 12, in school

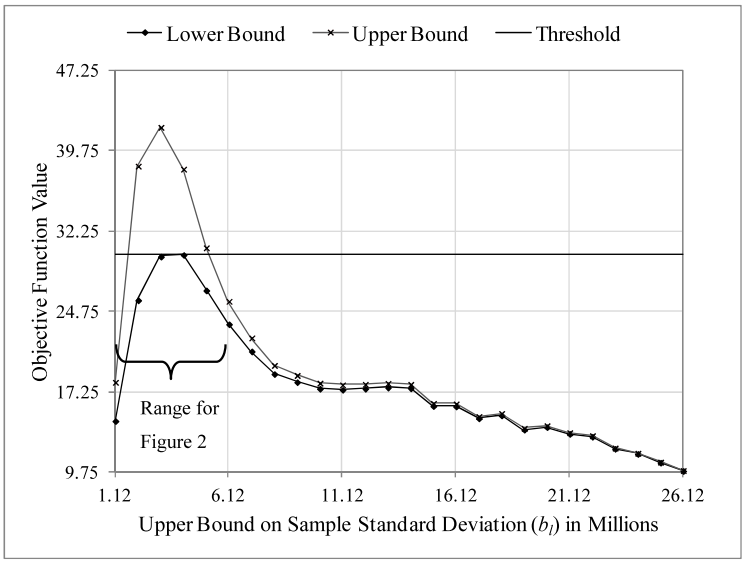

Figure 1 Upper and lower bounds for maximum z-test objective function value over a range of b_l (Case Study 1: n=30), illustrating the range for the finer mesh in the next step of the algorithm, which is found in Figure 2.

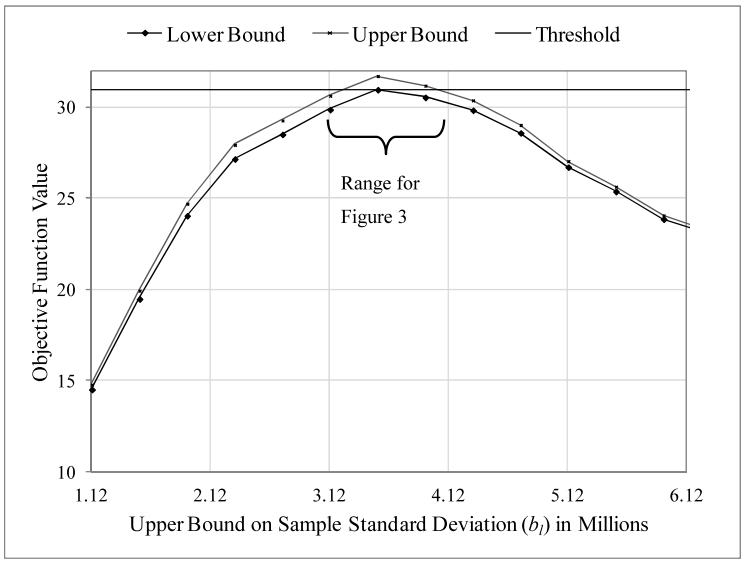

Figure 2 Upper and lower bounds for maximum z-test objective function value over a range of b_l (Case Study 1: n=30), illustrating the range for the finer mesh in the next step of the algorithm, which is found in Figure 3.

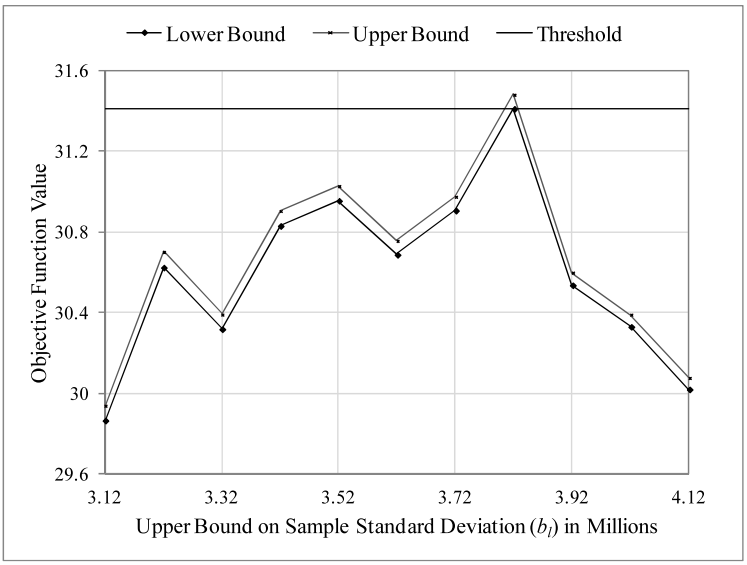

Figure 3 Upper and lower bounds for maximum z-test objective function value over a range of b_l (Case Study 1: n=30), at the finest mesh. The final value for the maximization problem is between 31.41 and 31.48.

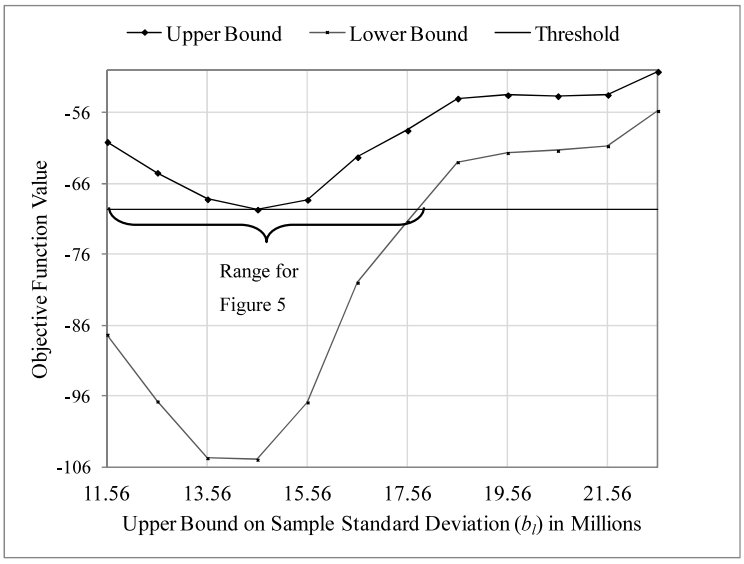

Figure 4 Upper and lower bounds for minimum z-test objective function value over a range of b_l (Case Study 1: n=30), illustrating the range for the finer mesh in the next step of the algorithm, which is found in Figure 5.

or employed, prior violent crime, any prior delinquency. The "treatment" variable is whether or not the individual is female; in particular we want to determine whether being female (controlling for race, criminal history, school/employment and relationship with parents) influences the probability

Figure 5 Upper and lower bounds for minimum z-test objective function value over a range of b_l (Case Study 1: n=30), illustrating the range for the finer mesh in the next step of the algorithm, which is found in Figure 6.

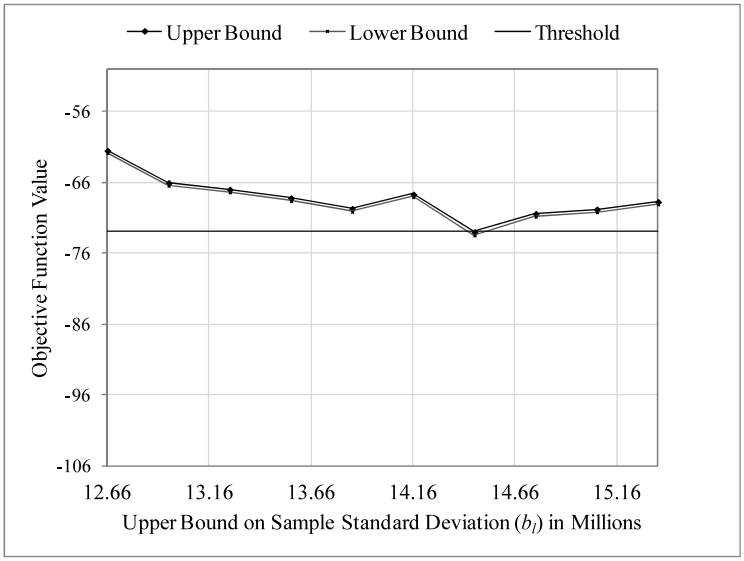

Figure 6 Upper and lower bounds for minimum z-test objective function value over a range of b_l (Case Study 1: n=30), at the finest mesh. The final value for the minimization problem is between -73.01 and -73.47.

of committing a violent crime. Here dist_{ij} =1 whenever all covariates of treated unit i are the same as those of the control unit j, 0 otherwise.

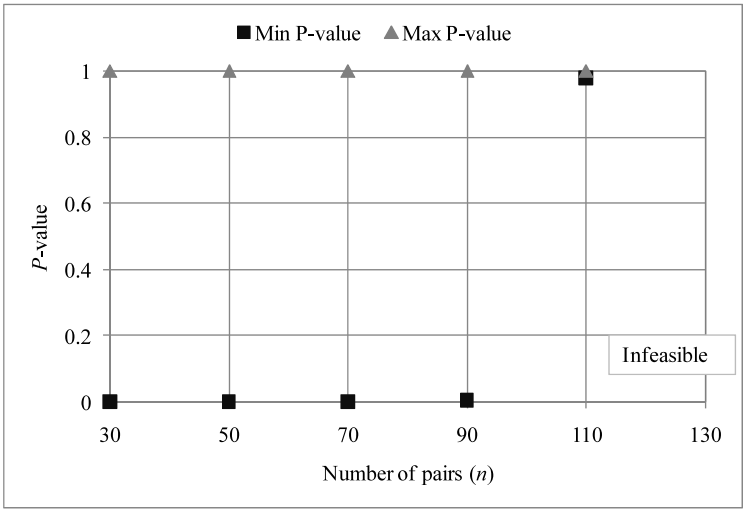

Figure 7 Variation of z-test optimum P-values for different n. (Case Study 1)

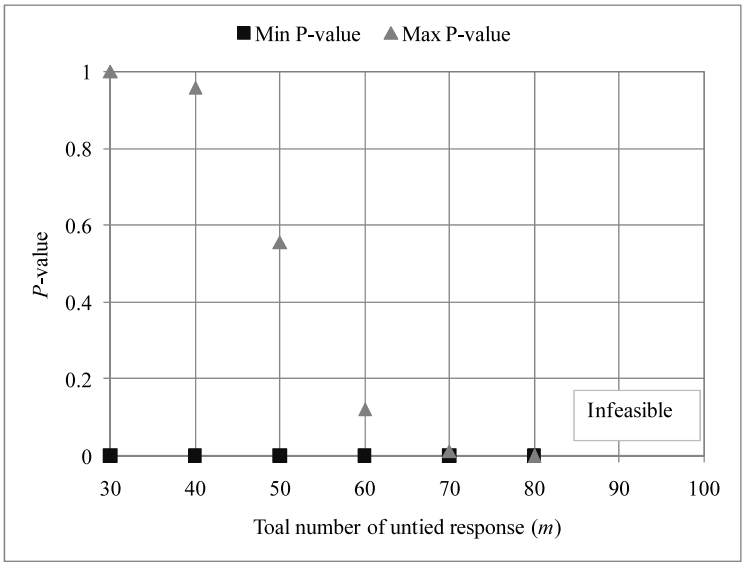

Figure 8 Variation of McNemar's test optimum P-value for different m. (Case Study 2)

Figure 7 is constructed in an analogous way to Figure 8 (using McNemar's test rather the z-test) showing the total number of discordant pairs along the x axis. Here, any matched pairs assignment would show a significant difference for the risks of violence between males and females. This difference becomes more pronounced as the number of pairs increases. Thus, the outcome is robust to the choice of experimenter.

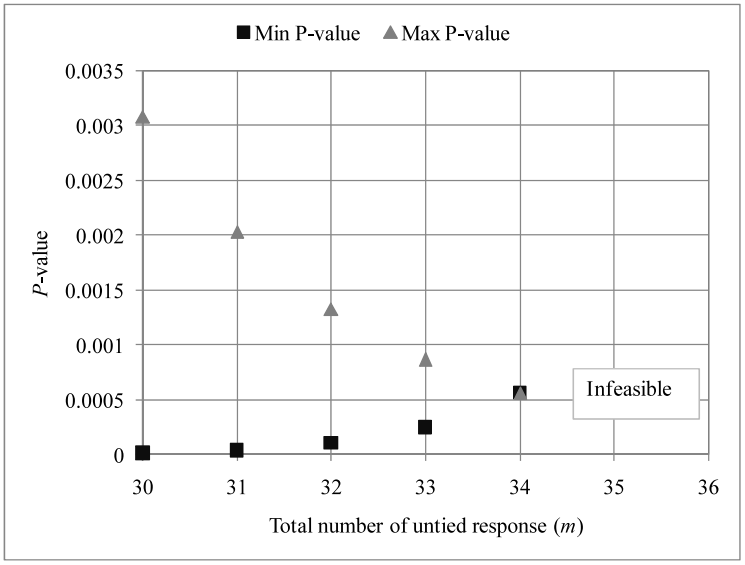

Figure 9 Variation of McNemar's test optimum P-value for different m. (Case Study 3)

6.3. Case Study 3

In this case study we used GLOW (*Global Longitudinal study of Osteoporosis in Women*) data used in the study of Hosmer et al. (2013). Each row of data represents a patient and the outcome is whether or not the person developed a bone fracture. The treatment is smoking. The covariates are: age, weight, height, and BMI. Here dist_{ij} =1 whenever the difference between covariates of treated unit i and control unit j are all less than or equal to 6, and dist_{ij} =0 otherwise.

Figure 9 indicates robustly that smoking causes fracture, no matter which experimenter conducts McNemar's test.

We also ran the robust χ^2 test allowing up to $M = 4$ control units per matched set. The minimum χ^2 value was 0, and the maximum was 5.2, where the solution for the maximum had several sets with $m > 1$. In particular, the count of pairs for the maximum solution is shown in Figure 10, illustrating that several treatment observations were matched with $m = 4$ control observations, which gives them more weight in the solution. This unevenness of m could result naturally, depending on the density of treatment and control observations within the space x.

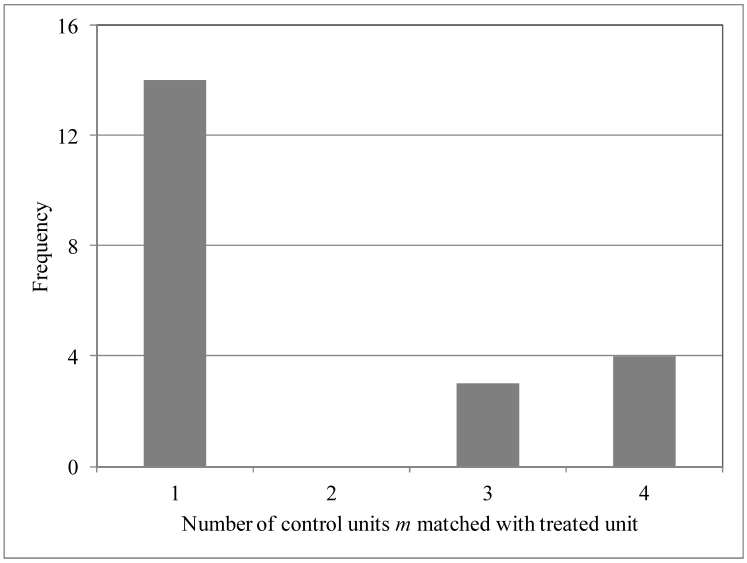

Figure 10 Frequency of number of control units m matched with treated unit obtained by maximizing χ^2 test statistic. (Case Study 3)

6.4. Case Study 4

In this case study, we used training program evaluation data described in Dehejia and Wahba (1999), and Dehejia and Wahba (2002), which were drawn from Lalonde (1986). This data set contains 15,992 control units and 297 treatment units. The covariates for matching are as follows: age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if married, 0 otherwise), nodegree (1 if no degree, 0 otherwise), and earnings in 1975. The outcome is the earnings in 1978. The treatment variable is whether an individual receives job training. We computed distance between units as follows: dist_{ij} =1 if covariates Black, Hispanic, married and nondegree were the same, and the differences in age, education and earnings in 1975 were less or equal to 5, 4 and 4000, respectively, for treated unit i and control unit j, 0 otherwise.

In the Figure 11, the P-value upper bounds are 1 and the P-value lower bounds are 0, illustrating that there is a lot of uncertainty associated with the choice of experimenter – one experimenter choosing 287 matched pairs can find a P-value of ~ 0 and declare a statistically significant difference while another experimenter can find a P-value of ~ 1 and declare the opposite. In this case it is truly unclear whether or not training has an effect on the earnings. To sanity check whether

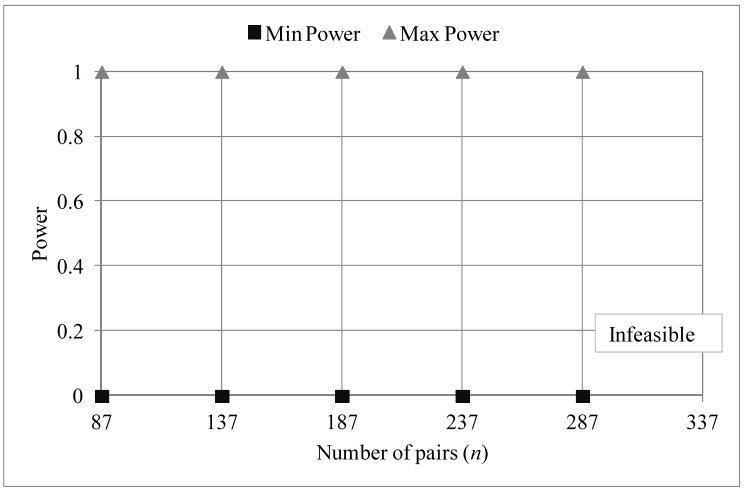

Figure 11 Variation of z-test optimum P-values for different n. (Case Study 4)

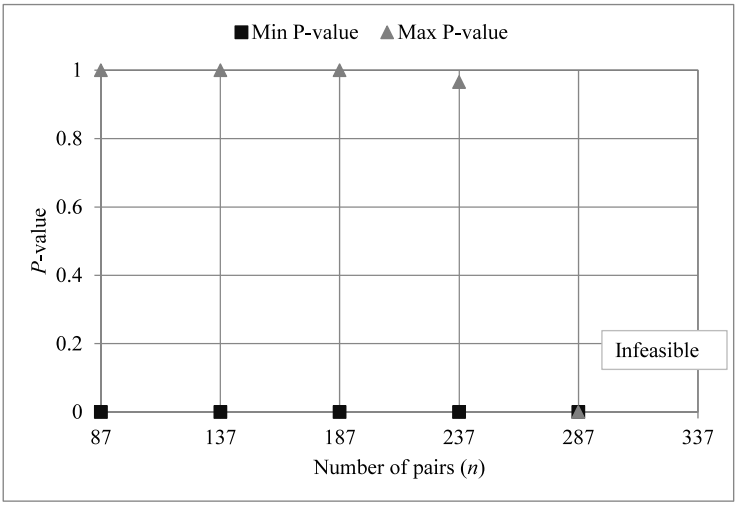

Figure 12 Variation of z-test optimum P-values for different n. (Case Study 4 with additional random noise on the treatment outcome)

a reasonably sized effect would have been detected had one been present, we injected synthetic random noise (with normal distribution of mean $\simeq$ \$10,000 and standard deviation $\simeq$ \$100) on the treatment outcome, and the z-test (see Figure 12) robustly detects the treatment effect before the solutions become infeasible.

7. Conclusions

Believing hypothesis test results conducted from matched pairs studies on observational data can be dangerous. These studies typically ignore the uncertainty associated with the choice of experimenter, and in particular, how the experimenter chooses the matched pairs. In one of our case studies, we showed that it is possible to construct matched pairs so that the treatment seems to be effective with high significance, and yet another set of matched pairs exists where the estimated treatment effect is completely insignificant. We want to know that for *any* reasonable choice of experimenter who chooses the assignments, the result of the test would be the same.

In this work, we considered the most extreme sets of assignments that can be constructed, in a computationally efficient way involving integer linear programs. These are the most extreme hypothesis test results from reasonable matched pairs. Across the sciences, medical informatics, politics, and in other areas, if we not only knew that the test was significant for *some* experimenter, but that it was significant for *all* reasonable experimenters, then the result would be much more trustworthy, and relevant to policy makers.

In the future, we have been extending this work to nonparametric hypothesis tests. Another avenue for further research is to consider tests applying to evidence factors, where units are given different levels of treatment assignment (see Rosenbaum 2010b, 2011).

Acknowledgments

The authors express their gratitude to the Natural Sciences and Engineering Research Council of Canada (NSERC) for their financial support of this research.

References

Breslow, N.E., N.E. Day. 1980. Statistical methods in cancer research. volume i - the analysis of case-control studies. *IARC scientific publications* **32** 5–338.

Chen, D.-S., R.G. Batson, Y. Dang. 2011. *Applied Integer Programming: Modeling and Solution*. Wiley.

Courtney, M.E., G.R. Cusick. 2010. Crime during the transition to adulthood: How youth fare as they leave out-of-home care in illinois, iowa, and wisconsin, 2002-2007. ICPSR27062-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Dehejia, R.H., S. Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**(448) 1053–1062.

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

Fanaee-T, Hadi, Joao Gama. 2014. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence* **2**(2-3) 113–127.

Fourer, R., D.M. Gay, B.W. Kernighan. 2002. Ampl: A modeling language for mathematical programming. Duxbury Press, Cole Publishing Co.

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

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

Ho, Daniel E., Kosuke Imai, Gary King, Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15**(3 (Summer)) 199–236.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* **81**(396) 945–960.

Hosmer, D.W., S. Lemeshow, R.X. Sturdivant. 2013. *Applied Logistic Regression: Third Edition.*. John Wiley and Sons Inc.

ILOG. 2007. Cplex 11.0 user's manual. ILOG, Inc.

Lalonde, R. 1986. Evaluating the econometric evaluations of training programs. *American Economic Review* **76**(4) 604–620.

Morgan, Stephen L., Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

Noor-E-Alam, M., A. Mah, J. Doucette. 2012. Integer linear programming models for grid-based light post location problem. *European Journal of Operational Research* **222**(1) 17–30.

Rosenbaum, Paul R. 2012. Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics* **21** 57–71.

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

Rosenbaum, P.R. 2010a. *Design of Observational Studies*. Springer, New York.

Rosenbaum, P.R. 2010b. Evidence factors in observational studies. *Biometrika* **97**(2) 333–345.

Rosenbaum, P.R. 2011. Some approximate evidence factors in observational studies. *Journal of the American Statistical Association* **106**(493) 285–295.

Rubin, D.B. 2007. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* **26**(1) 20–36.

Rubin, D.B. 2008. For objective causal inference, design trumps analysis. *Annals of Applied Statistics* **2**(3) 808–840.

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

Tamhane, A.C., D.D. Dunlop. 2000. *Statistics and Data Analysis.*. Prentice Hall, New Jersey.

Winston, W.L., M. Venkataramanan. 2003. *Introduction to Mathematical Programming,(4th ed.)*. Thomson (Chapter 9).

Wolsey, L.A. 1998. *Integer Programming*. Wiley-Interscience, Series in Discrete Mathematics and Optimization, Toronto.

Zubizarreta, J.R. 2012. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association* **107**(500) 1360–1371.

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

8. Supplement A

8.1. Integer Linear Programming Basics

ILP techniques have become practical for many large-scale problems over the past decade, due to a combination of increased processor speeds and better ILP solvers. Any type of logical condition can be encoded as linear constraints in an ILP formulation with binary or integer variables. Consider two binary variables $x \in \{0,1\}$ and $y \in \{0,1\}$. The logical condition "if $y = 0$ then $x = 0$" can be simply encoded as

$$x \leq y.$$

Note that this condition imposes no condition on x when $y = 1$. Translating if-then constraints into linear constraints can sometimes be more complicated; suppose, we would like to encode the logical condition that if a function $f(w)$ is greater than 0, then another function $g(w)$ is greater or equal to 0. We can use the following two linear equations to do this, where θ is a binary variable and M is a positive number that is larger than the maximum values of both f and g:

$$-g(w) \leq M\theta$$

$$f(w) \leq M(1-\theta).$$

In order for $f(w)$ to be positive, then θ must be 0, in which case, $g(w)$ is then restricted to be positive. If $f(w)$ is negative, θ must be 0, in which case no restriction is placed on the sign of $g(w)$. (See for instance the textbook of Winston and Venkataramanan (2003), for more examples of if-then constraints).

ILP can capture other types of logical statements as well. Suppose we would like to incorporate a restriction such that the integer variable S_i takes a value of K only if $i = t$, and 0 otherwise. The following four if-then constraints can be used to express this statement, where λ_1 and λ_2 are binary variables:

$$\lambda_1 = 1 \text{ if } i + 1 > t$$

$$\lambda_2 = 1 \text{ if } t+1 > i$$

$$S_i = k \text{ if } \lambda_1 + \lambda_2 > 1$$

$$S_i = 0 \text{ if } \lambda_1 + \lambda_2 < 2.$$

Each of these if-then constraints (4)-(7) can be converted to a set of equivalent linear equations, similar to what we described above. (See also Noor-E-Alam et al. (2012) and Winston and Venkataramanan (2003)).

There is no known polynomial-time algorithm for solving ILP problems as they are generally NP-hard, but they can be solved in practice by a number of well-known techniques (Wolsey (1998)). The LP relaxation of an ILP provides bounds on the optimal solution, where the LP relaxation of an ILP is where the integer constraints are relaxed and the variables are allowed to take non-integer (real) values. For instance, if we are solving a maximization problem, the solution of the LP relaxation can serve as an upper bound, since it solves a problem with a larger feasible region, and thus attains a value at least as high as that of the more restricted integer program. ILP solvers use branch-and-bound or cutting plane algorithms combined with other heuristics, and are useful for cases where the optimal integer optimal solution is not attained by the LP relaxation. The branch-and-bound algorithms often use LP relaxation and semi-relaxed problems as subroutines to obtain upper bounds and lower bounds, in order to determine how to traverse the branch-and-bound search tree (Chen et al. 2011, Wolsey 1998). The most popular ILP solvers such as CPLEX, Gurobi and MINTO each have different versions of branch-and-bound techniques with cutting plane algorithms and problem-specific heuristics.