Inferring optimal matching from experimental data

Debopam Bhattacharya,[*][†]

Dartmouth College.

February 15, 2006.

Abstract

This paper concerns the problem of dividing a set of individuals from a target population into peer-groups based on observed discrete-valued covariates so as to maximize the expected outcome of the entire set, e.g. allocation of freshmen to dorm rooms to maximize mean GPA. The decision is based on the outcome of random grouping of a pilot sample drawn from the target population. Computationally, the allocation problem reduces to a standard linear program whose objective function involves a first-step nonparametric estimate of the production function. I show that either the population problem has a unique solution or all feasible allocations are optimal which can be tested using a pivotal statistic. Under uniqueness, asymptotic distributions can be established using Cramer's theorem for large deviations. Under nonuniqueness the estimated maximum value has an asymptotically biased non-normal distribution. This bias can be analytically corrected to produce pretest confidence intervals for the maximum value with uniformly good coverage. This work complements recent work by Graham et al (2005) who compare a finite set of allocation rules rather than finding the optimal one. It differs from the "treatment choice" literature as it concerns the efficient way to simultaneously allocate a large group of individuals under aggregate resource constraints and focuses on the asymptotic analysis of estimated optimal solutions for a fixed decision rule. I illustrate the methods using data from Dartmouth's random assignment of freshmen to dorm rooms, where Sacerdote (2001) detected significant contextual peer effects. Segregation by previous academic achievement and by race are seen to minimize mean enrolment into sororities and maximize mean enrolment into fraternities. Segregation appears to have no effect on aggregate freshman year GPA.

[*]Address for correspondence: 327, Rockefeller Hall, Dartmouth College, Hanover, NH 03755. e-mail: debopam.bhattacharya@dartmouth.edu. This work is preliminary, please do not quote without permission.

[†]I am grateful to Geert Ridder for his encouragement in the early stages of the project. I thank Victor Chernozhukov, Whitney Newey, Jay Shambaugh, Doug Staiger, Bruce Sacerdote and participants of lunch seminars at Dartmouth College and MIT for helpful comments. All errors are mine.

1 Introduction

Analysis of peer effects has played an important role in both economic theory and econometrics. Conceptual issues involved in defining what constitutes "peer effects" and in the identification of these effects from nonrandomized field data were pioneered by Manski (1995). Subsequent research has addressed various methods of identifying peer effects from experimental as well as observational studies and tested the applicability of alternative models of peer interaction. This paper addresses a complementary question- namely, given the evidence on the magnitude and nature of peer effects, what is the socially "optimal" way for an outside planner to divide individuals into peer-groups? Obviously, the optimal grouping depends on what social criterion is to be optimized. But given a social welfare function, whether alternative grouping can affect aggregate social outcome will also depend on how peer effects interact with own effect in producing individual outcomes. Here, by "peer effects", I will mean "contextual effects" in Manski's terminology- the effect of peers' background on own outcome, controlling for own background.

To fix ideas, let us begin with the following allocation problem. Suppose that a college authority wants to improve average (and thus total) freshman year GPA of the incoming class, using dorm allocation as a policy instrument. The underlying behavioral assumption is that sharing a room with a "better" peer can potentially improve one's own outcome, where "better" could mean a high ability student, a student who is similar to her roommate etc. Scope for improvement exists if peer effects are nonlinear- i.e. the composite effect of own background and roommate's background on own outcome are not additively separable into an effect of own background plus an effect of roommate's background. Otherwise, all assignments should yield the same total, and thus average, outcome.

Assume that every dorm room can accommodate two students and the college can assign individuals to dorms based on an index of their previous academic achievement, say, SAT scores. For simplicity, assume that SAT score can take 3 distinct values- low, medium and high abbreviated by l,m,h. Denote the expected total score of a dorm room with each of 6 types of couples, denoted by $\mathbf{g} = (g_{hh}, g_{mm}, g_{ll}, g_{hl}, g_{hm}, g_{ml})'$. For instance, g_{ml} is the mean *per person* GPA score across all rooms which have one m-type and one l-type student. Also, denote the marginal distribution of SAT score for the current class by $\boldsymbol{\pi} = \pi_l, \pi_m, \pi_h$. Then an allocation is a vector $\mathbf{p}(\boldsymbol{\pi}) = (p_{hh}, p_{mm}, p_{ll}, p_{hl}, p_{hm}, p_{ml})'$, satisfying $p_{ij} \geq 0$ and $\mathbf{p'1} = 1$. Here p_{ij} (which equals p_{ji} by definition) denotes the fraction of dorm rooms that have one student of type i and one of type j, with $i, j \in \{h, m, l\}$. Then the authority's problem is defined by the following LP problem.

$$\max_{\{p_{ij}\}} \left[g_{hh}p_{hh} + g_{mm}p_{mm} + g_{ll}p_{ll} + g_{hl}p_{hl} + g_{hm}p_{hm} + g_{ml}p_{ml}\right]$$

s.t.

$$\begin{aligned} 2p_{hh} + p_{hl} + p_{hm} &= 2\pi_h \\ 2p_{mm} + p_{hm} + p_{ml} &= 2\pi_m \\ 2p_{ll} + p_{hl} + p_{ml} &= 2\pi_l = 2\left(1 - \pi_h - \pi_m\right) \\ p_{ij} &\geq 0,\ i, j \in \{h, m, l\}\,. \end{aligned}$$

The first set of linear constraints are just stating the budget constraint. For example, the first linear constraint simply says that the total number of students of h type in the dorm rooms (in every hh type room there are two h type students and hence the multiplier 2 appear before p_{hh}) should equal the total number of h type students that year. The first of these quantities is $N/2 \times (2p_{hh} + p_{hl} + p_{hm})$ if there are N students and hence $N/2$ dorm rooms. The second is $N \times \pi_h$. Thus one can view the g's as the preliminary parameters of interest and the solution to the LP problem and the resulting maximum value as functions of g's which constitute the ultimate parameters of interest. Note that the solution (the p_{ij}'s that solve the problem) may not always be unique but the maximum value is, provided the g's are bounded.

In general $\mathbf{g}$ will be unknown and so the above problem is infeasible. Now suppose, a sample was drawn from the same population from which the incoming freshmen are drawn (e.g. the freshmen class in the previous year). Further assume that this "pilot" sample was randomly grouped into rooms and the planner has access to freshman year GPA data for each member of this sample. Then the planner can simply calculate mean total score for this sample across dorm rooms, say, with one h and one l type to estimate $\hat{g}_{hl}$ which will be a good estimate of the unknown g_{hl} if the sample size is large. This assumes that peer interaction is unaffected by whether allocations are made through general randomization (as with the pilot sample) or by randomization within covariate categories (as will be done by the planner). This assumption can fail to hold if, for instance, students are more antagonistic to roommates who are different from them if they know that this allocation was, at least partly, a result of conscious choice of the planner.

Replacing unknown g's by the sample counterparts, the planner now solves

$$\max_{\mathbf{p}} \hat{\mathbf{g}}'\mathbf{p} \text{ s.t. } A\mathbf{p} = \boldsymbol{\pi},\ \mathbf{p} \geq \mathbf{0}$$

where

$$\begin{aligned}
\hat{\mathbf{g}} &= (\hat{g}_{hh}, \hat{g}_{mm}, \hat{g}_{ll}, \hat{g}_{hl}, \hat{g}_{hm}, \hat{g}_{ml})' \\
\mathbf{p} &= (p_{hh}, p_{mm}, p_{ll}, p_{hl}, p_{hm}, p_{ml})' \\
A &= \begin{pmatrix} 2 & 0 & 0 & 1 & 1 & 0 \\ 0 & 2 & 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 1 & 0 & 1 \end{pmatrix} \\
\boldsymbol{\pi} &= (\pi_h, \pi_m, \pi_l)'.
\end{aligned}$$

The purpose of this paper is to analyze the population problem and derive statistical properties of the estimated optimal solutions when the size of the pilot sample is large. The paper establishes necessary and sufficient conditions for uniqueness of the population solution which are testable. Further, it shows that asymptotic properties of the estimated solution and maximum value are completely different when the population solution is nonunique. In particular, the estimated maximum value is asymptotically zero-mean normal under uniqueness and is both asymptotically biased and nonnormal under nonuniqueness. The paper combines these findings to propose a valid method of inference of the pretest type. The key ingredient in the analysis is the fundamental theorem of linear programming which shows that the optimum of an LP problem can be attained only at a finite set of points, called the extreme points, within the feasible set. These points can be calculated a priori and their countable finiteness both makes the computation trivial and the analysis elegant.

Several other problems have similar (though not exactly the same) structure. For instance, consider a scenario where one observes student outcomes when teachers were randomly assigned to classes as in the well-known Project Star experiment. If one observes (mean) background characteristics of each class and characteristics of the teacher, one can estimate the mean outcome of each "type" of match. Using these estimates, one can devise the matching rule between teacher and class characteristics that should produce the largest expected test-score. The problem reduces mathematically to a finite LP problem where the RHS of the constraints now correspond to marginal distributions of teacher characteristics and class characteristics, respectively.

Randomized trials are becoming more prevalent in the study of economic phenomena (see, ...). The methods presented in this paper shed light on how to use the results of these trials to design effective policies. Although we present a simple (but probably the most relevant) case where the objective of the policy-maker is to maximize the mean value of the outcome, this analysis can be extended to other kinds of objectives like minimizing dispersion or maximizing quantiles.[1]

The plan of the paper is as follows. Section 2 discusses related papers and how they relate to the current work. Section 3 sets out the general problem and discusses characterizations and tests of uniqueness. Section 4 discusses asymptotic properties of the estimators under the assumption of uniqueness. Section 5 discusses the asymptotic behavior under nonuniqueness and proposes a pretest type method of inference. Section 6 applies these methods to data from Dartmouth College and discusses the findings. Section 7 concludes.

2 Relation to existing literature

The work presented here complements recent work by Graham, Imbens and Ridder (2005) who compare a fixed set of assignment rules (like positive and negative assortative matching) to see which one yields the higher expected pay-off. They are not concerned with calculating the optimal allocation. Having a method for calculating the optima, as presented in the current paper, also enables one to compute the efficiency of a given allocation rule by comparing it to the maximum attainable. This is discussed in the remarks at the end of section 4 of this paper. In contrast to Graham et al (2005), however, this paper is concerned exclusively with discrete covariates. Since most real-life allocation problems will concern discrete covariates, continuous ones can be easily converted into discrete ones and even the truly continuous cases will have to be "solved" by discrete approximations, this is not a large sacrifice in generality.[2]

Previous work on peer effects like Sacerdote (2001) in the experimental context and Cooley (2006) in a nonexperimental one have also considered comparing a set of specific allocations in terms of their expected outcome rather than trying to find the optimal one. Hoxby et al (2006) differs in focus in that they try to find evidence for and against alternative models of peer interaction. The current paper is different from these works in two fundamental ways. First, it develops methods for finding the (most) optimal allocation corresponding to any pre-specified social objective. Second, by utilizing the experimental set-up, it solves the relevant *policy question* without any behavioral assumption on what the agents actually do or any assumption about the structure of the "production function". The analysis, in other words, is completely nonparametric. However, unlike e.g. Cooley (2006), it does not attempt to recover the structural relationship between individual responses

[1] The present author is currently working on extending the analysis to quantiles.

[2] The nonstochastic version of optimization in the continuous case can be related to the Monge-Kantorovich mass transportation problem which is well-known to be analytically extremely difficult.

and that of one's peers. In that sense, the analysis is "reduced form" but provides the optimal policy prescription which is free from the classic Lucas Critique owing to the underlying randomized set-up. Note however that I continue to assume that nature of peer interaction is unaffected by whether allocations are made through general randomization or by randomization within covariate categories (as will be done by the planner). I am not aware of any experimental or other evidence suggesting that this assumption must fail in practice. Who the roommate is seems to be a far more important issue to a student than the precise *method* of roommate allocation that led to it.

The work presented here is also somewhat similar in spirit to the relatively new literature on treatment choice for heterogeneous populations , e.g. Manski (2004), Dahejia (2003) and Hirano and Porter (2005)- in that it concerns designing optimal policies based on results of a random experiment. However, there are several substantive differences between the current paper and the ones cited above. First, the current problem involves optimal matching between individuals rather than optimal assignment of individuals to treatments and is concerned with the outcome of both the assigned and the assignee. Two, resource constraints play an important role in the current paper. In our example above, not every student can potentially stay with a high type because the aggregate proportion of high types is fixed. The constraints make the current problem applicable to a different set of situations where a large number of individuals have to be allocated simultaneously and not everyone can be assigned what is the "first best" for them. Third, presence of the constraints makes the problem analytically different and ties in the solution to a stochastic version of linear programming problems. Fourth, this paper shows that Cramer type theorems for large deviations can be fruitfully used in such scenarios to derive the *asymptotic* distribution of the sample-based optimal allocation and the optimal values, *for a given choice of conditioning covariates*. This is in sharp contrast to Manski (2004) who is concerned with finding the relationship between the sample size and the choice of which covariates to condition on for minimizing maximum regret. Manski's analysis is based on derivation of bounds on the finite sample expected values of the criterion function. The main reason for this approach was that "the treatment selection probabilities are probabilities that one sample average exceeds another. Such probabilities generically do not have closed form expressions and are difficult to compute numerically" (Manski, 2004, page 1233, first paragraph). The analysis presented in section 4.3 of this paper shows that one can analyze the *asymptotic* behavior of such probabilities using Cramer type results under the assumption that all moments of the outcome variable are finite. When the outcome has a bounded support, as assumed in Manski (2004) for using Hoeffding's probability bounds, and relevant in many situations including the applications studied in this paper, this assumption will hold trivially. To summarize, this paper differs substantively from Manski (2004) in that it analyzes a constrained choice problem and is concerned with asymptotic properties of *a specific* sample-based decision rule for a given choice of covariates, rather than a finite sample-based analysis of finding the optimal rule.[3] However, in section 4 below, I provide a brief discussion on the choice of covariates that link it with Manski (2004).

3 The general problem and uniqueness

I first state the general form of the problem. Assume that there are M possible points of support of the covariate of interest and therefore a total of $K = M(M+1)/2$ possible types of room, indexed by the pair (j,k), $j = 1,..M$, $k = j,...M$. Let the vector of conditional mean of GPA obtained from a random assignment of the entire population be denoted by $g = (g_{jk})_{j=1,..M,k=j,...M}$. Let the proportion of incoming individuals (who are to be assigned to rooms)- called the "target" from now on- with value of covariate equal to m be denoted by π_m, $m = 1,...M$. Then the planner's problem, if she knew g, is referred to as the "population" problem and is given by

$$
\begin{aligned}
& \max_{\mathbf{p}=(p_{jk})_{j=1,..M,k=j,...M}} \quad \mathbf{p}'g \\
& \text{s.t.} \\
2p_{mm} + \sum_{j=1}^{m-1} p_{jm} + \sum_{k=m+1}^{M} p_{mk} &= 2\pi_m, \;\; m = 1,...M \\
p_{jk} &\geq 0, \, j = 1,..M, k = j,...M.
\end{aligned} \tag{1}
$$

The constraint set will be denoted by $\mathcal{P}$. Typically, the π's will be known and g's will not be known. A random sample is assumed to have been drawn from the same population from which the target comes and one observes the outcomes resulting from random assignment of this sample to rooms. Conditional means calculated on the basis of this sample are denoted by $\hat{g} = (\hat{g}_{jk})$, $j = 1,..M, k = j,...M$. The planner solves the problem in the previous display with g replaced by $\hat{g}$, which I will call the "sample problem". Whether this is the "optimal" action by the planner in the decision theoretic sense is an interesting and relevant question but is outside the scope of the current paper, which focusses on the asymptotic properties of *a* specific but "natural" action. For the asymptotic analysis of the decision theoretic optima in the treatment choice case see Hirano and

[3] Manski also restricts attention to "conditional empirical success" (CES) rules but considers the problem of choosing the set of covariates on which CES is conditioned as the decision problem.

Porter (2005). Note in passing that in my formulation, the constraint set for the sample problem is identical to that of the population problem.

The rest of this section presents results about uniqueness of solution to the *population* problem. I state and prove three propositions. The first proposition says that either there is a unique solution to the population problem or all feasible solutions are optimal (i.e. the entire parameter space is the "identified set") with no other intermediate possibilities. In other words, we cannot have nontrivial "set-identified" situations. The second proposition states a necessary and sufficient condition for nonuniqueness which can be tested. The third proposition sets out the asymptotic distribution theory for the statistic used to test nonuniqueness. Together, these three propositions completely characterize uniqueness of the population solution and a consistent test for detecting it.

I start by showing that we cannot have a situation where the population maxima (minima) is nonunique but the maximum (minimum) value is different from the minimum (maximum) value.

Proposition 1 *If all* $g_j > 0$*, then either* $\min_{p\in\mathcal{P}}(g'p) = \max_{p\in\mathcal{P}}(g'p)$ *or* $\{p^* : p^{*\prime}g = \min_{p\in\mathcal{P}}(g'p)\}$ *and* $\{p^* : p^{*\prime}g = \max_{p\in\mathcal{P}}(g'p)\}$ *are singletons.*

Proof. Suppose there exist $p \neq q \neq r \in \mathcal{P}$ such that $p, q \in \{p^* : p^{*\prime}g = \max_{p\in\mathcal{P}}(g'p)\}$ and $g'p = g'q > g'r$. Then

$$\mathcal{R} = \{\lambda p + (1-\lambda) q : \lambda \in [0,1]\} \subset \left\{p^* : p^{*\prime}g = \max_{p\in\mathcal{P}}(g'p)\right\} \text{ and } r \notin \mathcal{R}.$$

Choose $r_1, ...r_m, r_{m+1} \in \mathcal{R}$, where $m = M(M+1)/2 = \dim(g)$. Define $s_j = r_j - r$. Then the matrix S with columns $s_1, ...s_{m+1}$ has rank at most m and at least 1 since $r \notin \mathcal{R}$. So there exist constants $c_1, ...c_{m+1}$ such that $\sum_{j=1}^{m+1} c_j s_j > \mathbf{0}_{m\times 1}$. To see this, note that since the rank of S is at least 1, there will always exist at least one solution to the equation $S_{m\times(m+1)}c_{(m+1)\times 1} = a_{m\times 1}$ where all entries of a are nonnegative and at least one is strictly positive since this represents $m+1$ equations in m unknowns. In particular, if rank of S is m, then choose $a = 1_{m\times 1}$, if rank of S equals $k < m$, arrange the rows of S such that the first k rows are independent and choose $a = \left(1'_k, 0'_{m-k}\right)$. This implies that $t = q + \sum_{j=1}^{m+1} c_j s_j$ satisfies that $g't = g'q + g'\sum_{j=1}^{m+1} c_j s_j > g'q$ since all $g_j > 0$. Moreover, $t \in \mathcal{P}$ since $At = \pi$ and $t \geq \mathbf{0}_{m\times 1}$. This violates that $q \in \{p^* : p^{*\prime}g = \max_{p\in\mathcal{P}}(g'p)\}$ since t is feasible, not equal to q and gives a larger value of the objective function. An exactly analogous proof will work for the minimum. ■

Nonuniqueness of the type discussed above comes from what might be called "additivity" of peer effects. Additive peer effects means that for some functions $\gamma_1(.)$ and $\gamma_2(.)$,

$$E(Score|OB, RB) = \gamma_1(OB) + \gamma_2(RB) \tag{2}$$

where *score* is one's own outcome, OB and RB are one's own background covariate value and one's roommate's covariate value respectively. Additivity implies that all allocations would yield the same average effect overall and so the population version of problem (1) will not have a unique solution. However, the maximum (or minimum) value will still be unique. In other words, if our parameter of interest is the maximum value, then that parameter is a well-defined (many-to-one or one-to-one) functional of g (given A, π) in that for every g, the mapping gives one number. However, the solution, i.e, the argmax or the argmin is not a well-defined function of g for every g in that it is a one-to-many mapping.

A natural statistic for testing additivity would be the difference between the sample maximum and the sample minimum. But the asymptotic distribution of these quantities is far from obvious under the null of additivity. I will return to the distribution theory under nonuniqueness in section 5 of the paper where I will show that these distributions are not asymptotically centered at the population values. The following proposition shows that it is possible to bypass these complications altogether by establishing that additivity is equivalent to an easily testable condition. It also establishes an interesting equivalence result which fully characterizes uniqueness of the population solution.

Proposition 2 *Let* $A_{M\times\frac{M(M+1)}{2}}$ *be the matrix of constraints with* $M = rank\,(A)$ *and let* $B' = \begin{pmatrix} g_{\frac{M(M+1)}{2}\times 1} & A'_{\frac{M(M+1)}{2}\times M} \end{pmatrix}$. *Then the following statements are equivalent (i) (2) holds, (ii)* $rank\,(B) = M$ *and (iii)* $\min_{p\in\mathcal{P}}(g'p) = \max_{p\in\mathcal{P}}(g'p)$.

Proof. Let wlog, let the points of support of the covariate be $1, 2, ...M$ with aggregate probabilities $\pi_1, ...\pi_M$. Then

$$\begin{aligned}
g_1 &= E\,(Score|OB = 1, RB = 1) = 2\,(\gamma_1\,(1) + \gamma_2\,(1)) \\
g_2 &= E\,(Score|OB = 1, RB = 2) = (\gamma_1\,(1) + \gamma_2\,(2) + \gamma_1\,(2) + \gamma_2\,(1)) \\
&\quad .... \\
g_M &= E\,(Score|OB = 1, RB = M) = (\gamma_1\,(1) + \gamma_2\,(M) + \gamma_1\,(M) + \gamma_2\,(1)) \\
&\quad ..... \\
g_{\frac{M(M+1)}{2}} &= E\,(Score|OB = M, RB = M) = 2\,(\gamma_1\,(M) + \gamma_2\,(M))\,.
\end{aligned}$$

Then for any $p_1, ...p_{\frac{M(M+1)}{2}}$, it can be seen by writing out that

$$\begin{aligned} & g_1 p_1 + g_2 p_2 + ... + g_{\frac{M(M+1)}{2}} p_{\frac{M(M+1)}{2}} \\ = \; & (\gamma_1(1) + \gamma_2(1)) \times 2\pi_1 + (\gamma_1(2) + \gamma_2(2)) \times 2\pi_2 + ... (\gamma_1(M) + \gamma_2(M)) \times 2\pi_M \end{aligned}$$

which does not depend on $p_1, ...p_{\frac{M(M+1)}{2}}$. This shows that (i)$\Longrightarrow$(iii). Moreover, since the rows of A sum to $2\pi_1, ...2\pi_M$ respectively, it also follows that g is a linear combination of the rows of A and hence (i)$\Longrightarrow$(ii). Now, (ii) implies that g is a linear combination of the rows of A, say, $g = \sum_{j=1}^{M} \beta_j a_j$ where $a_1, ...a_M$ are the rows of A. Thus

$$\begin{aligned} g_1 &= g(1,1) = 2\beta_1, \\ g_2 &= g(1,2) = \beta_1 + \beta_2, \\ & ... \\ g_M &= g(1,M) = \beta_1 + \beta_M \end{aligned}$$

and so on. Therefore, $g(i,j) = \beta_i + \beta_j$ which shows that (ii) implies (i) (equate β_i with $\gamma_1(i) + \gamma_2(i)$ in the previous notation). Next, we show that (ii)$\Leftrightarrow$(iii). (ii) implies that $g = \sum_{j=1}^{M} \beta_j a_j$; so $g'p = \sum_{j=1}^{M} \beta_j a_j' p = \sum_{j=1}^{M} \beta_j \pi_j$ which does not depend on p, implying (iii). To show that (iii) implies (ii), we will show that "not (ii)" implies "not (iii)". Let $m = M(M+1)/2$. (ii) implies that there exists δ such that $A\delta = 0$ and $g'\delta \neq 0$. Let $p \in \mathcal{P}$. We will construct a $q \in \mathcal{P}$ such that $g'q \neq g'p$. Now, for some $c \neq 0$, we will have $q = p + c\delta > 0$ (in particular, choose $c \geq \max\{-p_1/\delta_1, ... - p_m/\delta_m\}$. For this c, we have that $q > 0$, $Aq = Ap + cA\delta = \pi$ and $g'q = g'p + cg'\delta \neq g'p$ since $g'\delta \neq 0$. This contradicts (iii). This last proof assumes that $\mathcal{P}$ is not a singleton, which is easy to check. ∎

Thus, one can test for linear separability or uniqueness by testing whether $rank(B) = M$. This is what is described now. Previous research on testing rank of a matrix exists (see e.g. Cragg and Donald (1997)). The current situation is a little different from those discussed in previous research in that the matrix B has only one row- the first one corresponding to g- which is unknown and estimated. Our test is based on the following statistic

$$T_n = n \times \min_b \left(\hat{g} - A'b\right) \hat{\Sigma}^{-1} \left(\hat{g} - A'b\right),$$

where $\hat{\Sigma}$ is a consistent estimate of Σ.

Under the null that rank of B is M, the value of T_n will be close to 0 and under the alternative of B being full-rank, its value should diverge to infinity. The next proposition describes the asymptotic distribution of T_n.

Proposition 3 *Assume that* $\sqrt{n}\left(\hat{g}-g\right)\xrightarrow{d}N\left(0,\Sigma\right)$.[4] *Then under the null hypothesis that* $rank\left(B\right)=M$, $T_n\xrightarrow{d}\varkappa^2$ *with* $df=\frac{M(M+1)}{2}-M$. *Under the alternative, i.e.* $rank\left(B\right)=M+1$, $T_n\xrightarrow{P}+\infty$. *So the test is consistent.*

Proof. By usual first order conditions, it is easy to check that

$$\arg\min_b \left(\hat{g}-A'b\right)\hat{\Sigma}^{-1}\left(\hat{g}-A'b\right)=\left[A\hat{\Sigma}^{-1}A'\right]^{-1}A\hat{\Sigma}^{-1}\hat{g}$$

whence

$$T_n=n\times\hat{g}'\left[\hat{\Sigma}^{-1}-\hat{\Sigma}^{-1}A'\left[A\hat{\Sigma}^{-1}A'\right]^{-1}A\hat{\Sigma}^{-1}\right]\hat{g}=n\hat{g}'\hat{V}\hat{g}$$

where

$$\hat{V}=\left[\hat{\Sigma}^{-1}-\hat{\Sigma}^{-1}A'\left[A\hat{\Sigma}^{-1}A'\right]^{-1}A\hat{\Sigma}^{-1}\right].$$

Let

$$V=\left[\Sigma^{-1}-\Sigma^{-1}A'\left[A\Sigma^{-1}A'\right]^{-1}A\Sigma^{-1}\right].$$

Observe that we can rewrite

$$\begin{aligned}T_n &= n\left(\hat{g}-g\right)'V\left(\hat{g}-g\right)+2ng'\hat{V}\hat{g}-ng'\hat{V}g\\ &\quad +n\left(\hat{g}-g\right)'\left(\hat{V}-V\right)\left(\hat{g}-g\right).\end{aligned}\tag{3}$$

Under the null, $g=A'\delta$ for some δ and hence

$$\begin{aligned}g'\hat{V} &= \delta'A\left[\hat{\Sigma}^{-1}-\hat{\Sigma}^{-1}A'\left[A\hat{\Sigma}^{-1}A'\right]^{-1}A\hat{\Sigma}^{-1}\right]\\ &= \delta'\left[A\hat{\Sigma}^{-1}-A\hat{\Sigma}^{-1}A'\left[A\hat{\Sigma}^{-1}A'\right]^{-1}A\hat{\Sigma}^{-1}\right]\\ &= 0.\end{aligned}$$

Further, since $\hat{V}-V\xrightarrow{P}0$ and $\sqrt{n}\left(\hat{g}-g\right)=O_p\left(1\right)$, it follows that

$$n\left(\hat{g}-g\right)'\left(\hat{V}-V\right)\left(\hat{g}-g\right)=o_p\left(1\right),$$

whence

$$T_n=n\left(\hat{g}-g\right)'V\left(\hat{g}-g\right)+o_p\left(1\right).$$

Using the breakup $\Sigma^{-1}=P'P$ with $PP'=I$, we have that under the null,

$$T_n=\left(\sqrt{n}P\left(\hat{g}-g\right)\right)'\left(\left[I-PA'\left[AP'PA'\right]^{-1}AP'\right]\right)\sqrt{n}P\left(\hat{g}-g\right).$$

[4] Sufficient conditions for this are that the random variable Score has finite second moments and that the probability of each type of room is bounded away from zero- discussed in greater details in the following section.

Since $\sqrt{n}P\left(\hat{g}-g\right) \xrightarrow{d} N\left(0, I_{\frac{M(M+1)}{2}}\right)$ and $\left[I - PA'\left[AP'PA'\right]^{-1}AP'\right]$ is idempotent, we have by standard arguments that $T_n \xrightarrow{d} \varkappa^2$ with *df* equal to

$$\begin{aligned} & trace\left[I - PA'\left[AP'PA'\right]^{-1}AP'\right] \\ = & \frac{M(M+1)}{2} - tr\left(PA'\left[AP'PA'\right]^{-1}AP'\right) \\ = & \frac{M(M+1)}{2} - tr\left(PA'\left[AP'PA'\right]^{-1}AP'\right) \\ = & \frac{M(M+1)}{2} - tr\left(\left[AP'PA'\right]^{-1}AP'PA'\right) \\ = & \frac{M(M+1)}{2} - tr\left(I_M\right) = \frac{M(M+1)}{2} - M. \end{aligned}$$

Under the alternative,

$$g'\hat{V}g = \min_b \left(g - A'b\right)'\hat{\Sigma}^{-1}\left(g - A'b\right) > 0$$

w.p. 1. So under the alternative, continuing from (3),

$$\begin{aligned} T_n &= n\left(\hat{g}-g\right)'V\left(\hat{g}-g\right) + 2\sqrt{n}g'\hat{V}\sqrt{n}\left(\hat{g}-g\right) + ng'\hat{V}g \\ &\quad + n\left(\hat{g}-g\right)'\left(\hat{V}-V\right)\left(\hat{g}-g\right), \end{aligned}$$

which, for large enough n, will be dominated by $ng'\hat{V}g \to +\infty$, since $\sqrt{n}\left(\hat{g}-g\right) = O_p(1)$, $n\left(\hat{g}-g\right)'\left(\hat{V}-V\right)\left(\hat{g}-g\right) = o_p(1)$ and $n\left(\hat{g}-g\right)'V\left(\hat{g}-g\right) = O_p(1)$. ■

4 Properties of the estimator under uniqueness

I now consider properties of the estimator, i.e. solution to the sample problem corresponding to (1) above, under the *assumption* that the population problem (1) has a unique solution. Recall that the constraint set $\mathcal{P}$, defined by (1), is a convex polytope with extreme points corresponding to the set of basic solutions (e.g. Luenberger (1984) page 19). These extreme points or basic solutions are obtained by taking the independent columns of the constraint matrix and inverting the resultant matrix. Corresponding to the $M \times M(M+1)/2$ constraint matrix defined in (1), there are at most $\begin{pmatrix} M(M+1)/2 \\ M \end{pmatrix}$ feasible basic solutions. Call the set of basic solutions S with

$|S| \leq \begin{pmatrix} M(M+1)/2 \\ M \end{pmatrix}$ with $\frac{|S|}{n} \to 0$ as $n \to \infty$. Let $S = \{z_1, ...z_{|S|}\}$. Rewrite

$$\begin{aligned}
\hat{g}_j &= \frac{\frac{1}{n}\sum_{i=1}^{n} D_{ij} score_i}{\frac{1}{n}\sum_{i=1}^{n} D_{ij}} \equiv \frac{\bar{y}_j}{\bar{d}_j} \text{ and } g_j = \frac{E(\bar{y}_j)}{E(\bar{d}_j)} = \frac{\mu_j}{\delta_j}, \\
w_{ij} &= D_{ij} score_i - \mu_j - \frac{\mu_j}{\delta_j} \times (D_{ij} - \delta_j), \omega_j = E(w_{\cdot j})
\end{aligned}$$

where D_{ij} is a dummy which equals 1 if the ith sampled individual is in room type j, $i = 1, ...n$ and $j = 1, ... |S|$. The expectation terms in the above display correspond to the combined experiment of drawing one random sample and making one random allocation of this drawn sample.

It should be obvious that the solutions to both the sample problem $\hat{p}(\boldsymbol{\pi})$ and the population problem $p(\boldsymbol{\pi})$ (with $\hat{g}$ replaced by the true g above) will be members of S.

4.1 Consistency

Proposition 4 *Assume that (i) $score_1, ...score_n$ and $D_{1j}, ...D_{nj}$ for each j are i.i.d. with finite mean, (ii) The population problem (1) has a unique solution. Then the solution to the sample problem, $\hat{p}(\boldsymbol{\pi})$ converges in probability to the solution of the population problem $p(\pi)$ as $n \to \infty$.*

Proof. Suppose that for a given $\boldsymbol{\pi}$, $\hat{p}(\boldsymbol{\pi}), p_0(\boldsymbol{\pi})$ solve the problem under $\hat{g}$ and g respectively. We want to show that for that fixed $\boldsymbol{\pi}$, $p\lim_{n\to\infty}(\hat{p}(\boldsymbol{\pi}) - p_0(\boldsymbol{\pi})) = 0$. From now on, we drop the qualifier π from our notation.

Suppose $p\lim_{n\to\infty}(\hat{p} - p_0) \neq 0$. Then, we must have

$$p \lim_{n\to\infty} (\hat{p} - p_0)' g < 0. \tag{4}$$

Observe that since p_0 solves the problem uniquely for g, we cannot have $p\lim_{n\to\infty}(\hat{p} - p_0)' g = (p\lim_{n\to\infty}\hat{p} - p_0)' g = 0$ Otherwise, $p\lim_{n\to\infty}\hat{p} \neq p_0$ will be another solution because $p\lim_{n\to\infty}\hat{p}$ will belong to the parameter space (note that for each n, $\hat{p}$ satisfies the constraints of (1)). Also, since by definition of p_0, $g'p_0 > g'\hat{p}$ for all n, we cannot have $p\lim_{n\to\infty}(\hat{p} - p_0)' g > 0$. Now,

$$(\hat{p} - p_0)' \hat{g} = (\hat{p} - p_0)' g + (\hat{p} - p_0)' (\hat{g} - g). \tag{5}$$

For n large enough, $(\hat{p} - p_0)' (\hat{g} - g)$ can be made arbitrarily close to 0 since the entries of $\hat{p}, p_0$ lie within the unit cube and $\hat{g}$ is consistent for g (which follows from assumption (i) by WLLN and the continuous mapping theorem). In view of this and (4), for large enough n, the LHS of (5) can be made strictly negative. But then for this n, we have that $\hat{p}'\hat{g} < p_0'\hat{g}$ which contradicts the definition of $\hat{p}$ since p belongs to the parameter space and gives a strictly larger objective function. ■

Corollary 1 *The optimal sample value $\hat{p}'\hat{g}$ converges in probability to $p_0'g$ as $n \to \infty$.*

4.2 Rate of convergence

Proposition 5 *In addition to the assumptions of proposition 4 above, assume that (iii) $\delta_j = E\left(\bar{d}_j\right) > m > 0$ for all j, (iv) for each j, $E\left(e^{tw_{1j}}\right) < \infty$ for all $t \in \mathbb{R}$. Then, $\Pr\left(\hat{p} \neq p_0\right) = O\left(e^{-\rho n}\right)$ as $n \to \infty$ for some $\rho > 0$.*

Proof. We first show that if $z_1'g < z_2'g$, then $\Pr\left(z_1'\hat{g} > z_2'\hat{g}\right) = O\left(e^{-n}\right)$. So assume that $z_1'g < z_2'g$.

Then

$$
\begin{aligned}
& \Pr\left(z_1'\hat{g} > z_2'\hat{g}\right) = \Pr\left[\left(z_1 - z_2\right)'\left(\hat{g} - g\right) \geq \left(z_2 - z_1\right)' g\right] \\
= \; & \Pr\left[\sum_{j=1}^{|S|}\left(z_{1j} - z_{2j}\right)\left(\hat{g}_j - g_j\right) \geq \left(z_2 - z_1\right)' g\right] \\
= \; & \Pr\left[\sum_{j=1}^{|S|}\left(z_{1j} - z_{2j}\right)\left(\frac{\bar{y}_j - E\left(\bar{y}_j\right)}{\bar{d}_j} - \frac{E\left(\bar{y}_j\right) \times \left(\bar{d}_j - E\left(\bar{d}_j\right)\right)}{\bar{d}_j E\left(\bar{d}_j\right)}\right) \geq \left(z_2 - z_1\right)' g\right] \\
\leq \; & \Pr\left[\sum_{j=1}^{|S|}\frac{1}{\bar{d}_j}\left(z_{1j} - z_{2j}\right)\left(\bar{w}_j - \omega_j\right) \geq \left(z_2 - z_1\right)' g, \; \bar{d}_1 > m, ...\bar{d}_{|S|} > m\right] \\
& +1 - \Pr\left[\bar{d}_1 > m, ...\bar{d}_{|S|} > m\right] \\
\leq \; & \Pr\left[\left(\cup_{j=1}^{|S|}\left\{\frac{1}{\bar{d}_j}\left(z_{1j} - z_{2j}\right)\left(\bar{w}_j - \omega_j\right) \geq \frac{\left(z_2 - z_1\right)' g}{|S|}\right\}\right) \cap \left(\bar{d}_1 > m, ...\bar{d}_{|S|} > m\right)\right] \\
& +1 - \Pr\left[\bar{d}_1 > m, ...\bar{d}_{|S|} > m\right]
\end{aligned}
$$

$$
\begin{aligned}
= \; & \Pr\left[\cup_{j=1}^{|S|}\left(\left\{\frac{1}{\bar{d}_j}\left(z_{1j} - z_{2j}\right)\left(\bar{w}_j - \omega_j\right) \geq \frac{\left(z_2 - z_1\right)' g}{|S|}\right\} \cap \left(\bar{d}_1 > m, ...\bar{d}_{|S|} > m\right)\right)\right] \\
& +1 - \Pr\left[\bar{d}_1 > m, ...\bar{d}_{|S|} > m\right] \\
\leq \; & \sum_{j=1}^{|S|}\Pr\left[\left\{\frac{1}{\bar{d}_j}\left(z_{1j} - z_{2j}\right)\left(\bar{w}_j - \omega_j\right) \geq \frac{\left(z_2 - z_1\right)' g}{|S|}\right\} \cap \left(\bar{d}_1 > m, ...\bar{d}_{|S|} > m\right)\right] \\
& +1 - \Pr\left[\bar{d}_1 > m, ...\bar{d}_{|S|} > m\right] \\
\leq \; & \sum_{j=1}^{|S|}\Pr\left[\left(z_{1j} - z_{2j}\right)\left(\bar{w}_j - \omega_j\right) \geq \frac{m\left(z_2 - z_1\right)' g}{|S|}\right] + 1 - \Pr\left[\bar{d}_1 > m, ...\bar{d}_{|S|} > m\right] \\
= \; & \sum_{j=1}^{|S|}\Pr\left[\frac{1}{n}\sum_{i=1}^{n}\left(z_{1j} - z_{2j}\right)\left(w_{ij} - \omega_j\right) \geq \frac{m\left(z_2 - z_1\right)' g}{|S|}\right] + 1 - \Pr\left[\bar{d}_1 > m, ...\bar{d}_{|S|} > m\right]
\end{aligned}
$$

We now invoke the principle of large deviations. By condition (v) and Cramer's theorem (see, e.g. Hollander (2000), page 5) the first probability

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}\left(z_{1j}-z_{2j}\right)\left(w_{ij}-\omega_j\right) \geq \frac{m\left(z_2-z_1\right)' g}{|S|}\right]$$

is $O\left(e^{-\rho_1 n}\right)$ for some $\rho_1 > 0$ since $\left(z_2-z_1\right)' g > 0$. The random variables D_{ij} are binary and so will satisfy Cramer's condition trivially. Therefore,

$$\begin{aligned} & 1-\Pr\left[\bar{d}_1 > m, \ldots \bar{d}_{|S|} > m\right] \\ = & \Pr\left[\cup_{j=1}^{|S|}\left(\bar{d}_j < m\right)\right] \leq \sum_{j=1}^{|S|}\Pr\left(\bar{d}_j < m\right) = \sum_{j=1}^{|S|}\Pr\left(\bar{d}_j-\delta_j < m-\delta_j\right) \end{aligned}$$

is $O\left(e^{-\rho_2 n}\right)$ for some $\rho_2 > 0$ by Cramer's theorem since $m-\delta_j < 0$ for all j by condition (iii). Thus, we have shown that if $z_1' g < z_2' g$, then $\Pr\left(z_1'\hat{g} > z_2'\hat{g}\right) = O\left(e^{-\rho n}\right)$ for some $\rho > 0$.

Now, if $p_0 = z_1$, i.e. $z_1' g > z_j' g$ for all $j = 2, \ldots |S|$, then $\Pr\left(z_1'\hat{g} > z_j'\hat{g}\right) = 1 - O\left(e^{-\rho n}\right)$ for any $j = 2, \ldots |S|$. This implies that

$$\begin{aligned} \Pr\left(\hat{p} = z_1\right) &= \Pr\left(z_1'\hat{g} > z_2'\hat{g}, z_1'\hat{g} > z_3'\hat{g}, \ldots z_1'\hat{g} > z_{|S|}'\hat{g}\right) \\ &= 1-\Pr\left(\cup_{j=1}^{|S|}\left\{z_1'\hat{g} < z_j'\hat{g}\right\}\right) \\ &\geq 1-\sum_{j=1}^{|S|}\Pr\left(z_1'\hat{g} < z_j'\hat{g}\right) \\ &= 1-O\left(|S|\, e^{-\rho n}\right) = 1-O\left(e^{-\rho n}\right). \end{aligned}$$

Analogously, if $p_0 = z_j$, then $\Pr\left(\hat{p} = z_j\right) = 1 - O\left(e^{-\rho n}\right)$ for all $j = 1, 2, \ldots |S|$. ■

From the analysis above, it is clear that if true problem has a unique solution z_j, then for any finite n, $\hat{p}$ corresponding to the sample problem will have a discrete distribution with $\Pr\left(\hat{p} = z_j\right) = 1 - O\left(e^{-\rho n}\right)$ and $\Pr\left(\hat{p} = z_k\right) = O\left(e^{-\rho n}\right)$ for all $k \neq j$. Consequently, $e^{-n\rho/2}\left(\hat{p}-p_0\right)$ will have a finite asymptotic variance where the constant ρ will be related to the rate function in Cramer's theorem.

4.3 Optimal value

Consistency of the optimal value is corollary 1. To see the rate of convergence of the sample optimal value, notice that

$$\sqrt{n}\left(\hat{g}'\hat{p}-g'p_0\right) = \sqrt{n}\left(\hat{g}-g\right)'\hat{p} + \sqrt{n}\left(\hat{p}-p_0\right)' g. \tag{6}$$

Since the second term is $o_p(1)$ under the assumptions of the previous subsection, the asymptotic distribution of the sample optimal value will be normal with asymptotic variance given by $p_0' Avar(\hat{g}) p_0$. The asymptotic variance of $\hat{g}_j$ (note that assumption (iv) in proposition 2 implies that $\hat{g}_j$ has finite variance for all j) by standard delta method arguments (and condition (iii) of proposition 2) equals the asymptotic variance of

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_{ij}score_i-\mu_j}{\delta_j}-\frac{\mu_j}{\delta_j^2}(D_{ij}-\delta_j)\right]$$

which can be consistently estimated by replacing μ_j and δ_j by $\bar{y}_j$, $\bar{d}_j$ respectively while p_0 is consistently estimated by the solution vector $\hat{p}$. From now on, we will use the notation Σ to denote $Avar(\hat{g})$. These observations are collected into a proposition for easy reference later.

Proposition 6 *Under the assumptions of proposition 5*

$$\sqrt{n}\left(\hat{g}'\hat{p}-g'p_0\right)\xrightarrow{d}N\left(0,p_0'\Sigma p_0\right)$$

where Σ is diagonal. A consistent estimate of the (j,k)th element of Σ is given by $\frac{1}{n}\sum_{i=1}^{n}s_{ij}s_{ik}$ where

$$s_{ij}=\frac{D_{ij}score_i}{\frac{1}{n}\sum_{i=1}^{n}D_{ij}}-\frac{\frac{1}{n}\sum_{i=1}^{n}D_{ij}score_i}{\left(\frac{1}{n}\sum_{i=1}^{n}D_{ij}\right)^2}\times D_{ij}$$

The fact that the optimal solution converges much faster than the optimal value is interesting and can be "intuitively" explained as follows. Consider an oversimplified version of the above problem.

$$\begin{aligned}
&\max \hat{g}_1p_1+\hat{g}_2p_2\\
&s.t.\\
p_1+p_2 &= 1\\
p_1 &\geq 0, p_2\geq 0.
\end{aligned}$$

The extreme points are $(1,0)$ and $(0,1)$. Whenever $\hat{g}_1/\hat{g}_2>1$, the solution is at $(1,0)$ (and whenever $\hat{g}_1/\hat{g}_2<1$, the solution is at $(0,1)$). Observe that at this point, the optimal value is $\hat{g}_1$. Even if $\hat{g}_1,\hat{g}_2$ vary a lot from one sample to another, as long as $\hat{g}_1/\hat{g}_2>1$, the solution continues to be $(1,0)$ but the maximum value $\hat{g}_1$ varies a lot. In other words, the optimal solution is affected much less by the variation in $\hat{g}_1,\hat{g}_2$ relative to the optimal value.

4.4 Which covariates?

Obviously, including more covariates or refining the support of a given covariate leads to higher maxima and lower minima in both the population (i.e. with known g's) and in the sample (with estimated $\hat{g}$'s). To see this, consider the following simple example. First consider the case with no covariate and let mean outcome equal $\hat{\mu}$. Now consider a binary covariate called X which takes two values H and L. Let the proportion of those be π_H and π_L respectively and let the random assignment produce f_1, f_2 and f_3 fractions of room types HH, HL and LL respectively with $\hat{g}_1, \hat{g}_2, \hat{g}_3$ be the respective categorical means of the outcome. Then by definition, we have

$$\begin{aligned}
\hat{\mu} &= f_1\hat{g}_1 + f_2\hat{g}_2 + f_3\hat{g}_3 \\
2f_1 + f_2 &= 2\pi_H \\
f_2 + 2f_3 &= 2\pi_L.
\end{aligned}$$

Now, with these categories, the optimality problem solves

$$\begin{aligned}
& \max_p \left[p_1\hat{g}_1 + p_2\hat{g}_2 + p_3\hat{g}_3\right] \\
& s.t. \\
2p_1 + p_2 &= 2\pi_H \\
p_2 + 2p_3 &= 2\pi_L.
\end{aligned}$$

Clearly, (f_1, f_2, f_3) is feasible for the above problem and so the maximum for the above problem must be at least $\hat{\mu}$.

So when one knows average outcome for every potential combination of all covariates, using all covariates will yield the largest maximum mean outcome- which we will call $g'p_0$- obtained as the optimal value for the problem displayed above with the $\hat{g}$ replaced by g. When no covariates are used, the sample value is $\hat{\mu}$ whose mean (and plim) for large n is $\pi_H^2 g_1 + 2\pi_h\pi_l g_2 + \pi_L^2 g_3$. In terms of being closer to $g'p_0$ in plim, we have shown above that $p\lim \hat{g}'\hat{p} = g'p_0$. So for large n, it is advisable to use all covariates. Also, for large n, $E\left(g'p_0 - \hat{g}'\hat{p}\right)^2 \to 0$ if the outcomes have finite variance.[5] So the same conclusion holds if one wants to minimize a MSE type functional.

For small n, these approximations are poor and one needs to calculate $E\left(g'p_0 - \hat{\mu}\right)^2$ and $E\left(g'p_0 - \hat{g}'\hat{p}\right)^2$ or simply $g'p_0 - E\left(\hat{\mu}\right)$ and $g'p_0 - E\left(\hat{g}'\hat{p}\right)$, where $E\left(.\right)$ denotes the exact finite

[5] since, $E\left(\hat{g}'\hat{p} - g'p\right)^2 \leq E\left\|\hat{g} - g\right\|^2 + mE\left\|\hat{p} - p\right\|^2 + 2m\left(E\left\|\hat{g} - g\right\|^2\right)\left(E\left\|\hat{p} - p\right\|^2\right)^{1/2}$ and $E\left\|\hat{g} - g\right\|^2 \to 0$ if outcomes have finite variance.

sample expectation, to see when it is better not to condition on covariates. This is the approach taken in Manski (2004) in the treatment choice context[6]. Intuitively, finer categorization implies smaller precision in the estimation of $\hat{g}$'s making it more likely that $E\left(\hat{\mu}-g'p_0\right)^2 \leq E\left(\hat{g}'\hat{p}-g'p_0\right)^2$.

Remark 1 *Virtually the same argument can be used to justify that solution to the LP problem provides asymptotically a higher (not lower) expected mean than any other allocation mechanism- like positive assortative (PA) and negative assortative (NA) matching or random assignment, considered in Graham et al (2005). This is because all these other allocations have to be feasible and the LP one maximizes the mean among all feasible allocations.*

Remark 2 *One can compare relative efficiency of a given allocation, say PA matching, relative to optimal matching. In the above example, positive assortative matching is the outcome with* $p_2 = 0, p_1 = \pi_H$ *and* $p_3 = \pi_L$. *The value of the outcome is* $\pi_H \hat{g}_1 + \pi_L \hat{g}_3 \overset{P}{\to} \pi_H g_1 + \pi_L g_3$. *Thus one can estimate the relative efficiency of PA matching by* $\left(\pi_H \hat{g}_1 + \pi_L \hat{g}_3\right)/\hat{g}'\hat{p}$, *that of NA by* $\left(\pi_H - \min\left\{\pi_H, \pi_L\right\}\right)\hat{g}_1 + \min\left\{\pi_H, \pi_L\right\}\hat{g}_2 + \left(\pi_L - \min\left\{\pi_H, \pi_L\right\}\right)\hat{g}_3/\hat{g}'\hat{p}$ *and of random allocation by* $\hat{\mu}/\hat{g}'\hat{p}$. *Using the usual delta method and the asymptotic distributions derived above, one can also form confidence intervals for these relative efficiencies.*

5 Nonuniqueness and pretest CI

It turns out that the asymptotic distribution of the estimated optimal value is quite different if the population solution is nonunique. To see this, reconsider the maximization problem described in (1) and recall the definition of $S = \left\{z_1, ...z_{|S|}\right\}$, the set of feasible extreme points defined in section 4. Let $\hat{p}$ denote the sample solution and let $\hat{v} = \hat{p}'\hat{g}$. Observe that the fundamental theorem of linear programming implies that one can alternatively define

$$\hat{v} = \max\left\{z_1'\hat{g}, ...z_{|S|}'\hat{g}\right\}.$$

The asymptotic distribution of the sample maximum value under nonunique population solution is described in the following proposition.

[6] Manski's approach actually considers the suprema of these differences over all possible values of g's. The asymptotic performance of one specific "decision rule" , the main subject of this paper, is valid for all possible values of g's except ones which lead to a nonunique solution to the population problem- discussed in the following section.

Proposition 7 *Assume that* $\sqrt{n}\left(\hat{g}-g\right) \xrightarrow{d} W \equiv N\left(0,\Sigma\right)$. *If* $\min_{p\in\mathcal{P}}\left(g'p\right) = \max_{p\in\mathcal{P}}\left(g'p\right) = v$, *then* $plim_{n\to\infty}\hat{v} = v$ *and*

$$\sqrt{n}\left(\hat{v}-v\right) \xrightarrow{d} W^{\max} = \max\left\{W_1,...W_{|S|}\right\}$$

where $W_j = z_j'W$ *for* $j = 1,...|S|$. $W^{\max}$ *is not Gaussian in general.*

Proof. Under nonuniqueness we have $v = z_1'g = ... = z_{|S|}'$ and continue to have that for all finite n,

$$\hat{v} = \sup_{p\in\mathcal{P}} p'\hat{g} = \max\left\{z_1'\hat{g},...z_{|S|}'\hat{g}\right\}.$$

Since $plim_{n\to\infty}\hat{g} = g$, $\max\left\{z_1'\hat{g},...z_{|S|}'\hat{g}\right\} \xrightarrow{P} \max\left\{z_1'g,...z_{|S|}'g\right\} = v$, by the continuous mapping theorem. So $\hat{v}$ is consistent for v. Moreover,

$$\begin{aligned}\sqrt{n}\left(\hat{v}-v\right) &= \max\left\{\sqrt{n}\left(z_1'\hat{g}-v\right),...\sqrt{n}\left(z_{|S|}'\hat{g}-v\right)\right\}\\ &= \max\left\{z_1'\sqrt{n}\left(\hat{g}-g\right),...z_{|S|}'\sqrt{n}\left(\hat{g}-g\right)\right\},\end{aligned}$$

whence the conclusion follows by the continuous mapping theorem. ■

Although $W^{\max}$ does not have a pivotal distribution, its distribution can be simulated by first drawing a w from the (estimated) asymptotic distribution of $\sqrt{n}\left(\hat{g}-g\right)$, i.e. $N\left(0,\hat{\Sigma}\right)$, calculate the extreme points of the constraints set, viz. S and then calculate $\max\left\{z_1'w,...z_{|S|}'w\right\}$. Repeating this a large number of times should simulate the distribution of $W^{\max}$. However, this process can be very time-consuming if M is moderately large, implying that $|S|$ is very large. The following trick helps us reduce computation significantly. Solve problem (1) after replacing g by a draw from the (estimated) asymptotic distribution of $\sqrt{n}\left(\hat{g}-g\right)$, i.e. $N\left(0,\hat{\Sigma}\right)$. Repeat this a large number of times. Since the maxima will continue to be one of the extreme points, we will end up with the distribution of $W^{\max}$.

However, $E\left(W^{\max}\right) = \theta_0 \neq 0$, in general and therefore a bias corrected estimate of v will be given by $\hat{v}_{BC} = \hat{v} - n^{-1/2}\tilde{\theta}_0$, where $\tilde{\theta}_0$ equals the mean of the simulated distribution of $W^{\max}$. To see why this bias arises, assume for simplicity that $|S| = 2$, $z_1 = (1,1)$, $z_2 = (-1,-1)$. Let $\sqrt{n}\left(\hat{g}-g\right) \xrightarrow{d} \left(X_1,X_2\right) \simeq N_2\left(0,I\right)$. Then $\max\left\{z_1'\sqrt{n}\left(\hat{g}-g\right), z_2'\sqrt{n}\left(\hat{g}-g\right)\right\} \xrightarrow{d} |X_1+X_2|$. But X_1+X_2 is a mean zero normal, so its absolute value has a strictly positive mean.

Thus a bias-corrected C.I. for v can be formed as follows. Choose d_L, d_H such that $1-\alpha = \Pr\left(d_L \leq W^{\max} \leq d_H\right)$. A level $(1-\alpha)$ confidence interval for v is then given by

$$CI_{non-unique} = \left[\hat{v}-\frac{d_H}{\sqrt{n}},\, \hat{v}-\frac{d_L}{\sqrt{n}}\right] = \left[\hat{v}_{BC}-\frac{d_H-\tilde{\theta}_0}{\sqrt{n}},\, \hat{v}_{BC}+\frac{\tilde{\theta}_0-d_L}{\sqrt{n}}\right].$$

Typically, $d_L < \theta_0 < d_H$ implying that $CI_{non-unique}$ will be "centred around" $\hat{v}_{BC}$ (and not $\hat{v}$).

Since the asymptotic distribution of the maximum value is completely different depending on whether the population solution is unique, one can adopt a pretest type method for calculating confidence intervals. One first tests for additivity. Upon rejection of it, one calculates the optimal solutions ($\hat{p}$) and level $(1-\alpha)$ confidence intervals as

$$CI_{unique} = \left[\hat{g}'\hat{p} - \frac{\hat{p}'\hat{\Sigma}\hat{p}}{\sqrt{n}} z_{\alpha/2}, \, \hat{g}'\hat{p} + \frac{\hat{p}'\hat{\Sigma}\hat{p}}{\sqrt{n}} z_{\alpha/2}\right]$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile point of the standard normal. Upon failure to reject the null of nonuniqueness, one calculates the confidence interval $CI_{non-unique}$. The overall confidence interval is thus

$$\hat{I} = 1\,(T_n \geq c)\, CI_{unique} + 1\,(T_n < c)\, CI_{non-unique},$$

where c is the critical value used in the T_n-based test of uniqueness, described in proposition 3. The estimator of the maximum value is simply $\hat{\theta} = \hat{g}'\hat{p}$, no matter whether the null of nonuniqueness is rejected or not. This is the approach followed in the application. The following proposition describes the probability that $\hat{I}$ covers the maximum value $v \equiv v\,(g)$, where the notation $v\,(g)$ makes it clear that the maximum value depends on g.

Proposition 8 *Under the assumptions of propositions 3, 6 and 7, for all g,*

$$\Pr\left(v\,(g) \in \hat{I}\right) \geq 1 - \alpha - \alpha',$$

where α' is the size of the test described in proposition 3.

Proof. For values of g such that $g \notin \mathcal{R}\,(A)$, i.e.$rank\,(B) = M + 1$, we have

$$\begin{aligned}
& \Pr\left(v\,(g) \in \hat{I} | g \notin \mathcal{R}\,(A)\right) \\
= \; & \Pr\left(T_n \geq c, \, \hat{g}'\hat{p} - \frac{\hat{p}'\hat{\Sigma}\hat{p}}{\sqrt{n}} z_{\alpha/2} \leq v \leq \hat{g}'\hat{p} + \frac{\hat{p}'\hat{\Sigma}\hat{p}}{\sqrt{n}} z_{\alpha/2} | g \notin \mathcal{R}\,(A)\right) \\
& + \Pr\left(T_n < c, \hat{v} - \frac{d_H}{\sqrt{n}} \leq v \leq \hat{v} - \frac{d_L}{\sqrt{n}} | g \notin \mathcal{R}\,(A)\right). \qquad (7)
\end{aligned}$$

The second term is dominated by $\Pr\left(T_n < c | g \notin \mathcal{R}\,(A)\right)$ which converges to 0 as $n \to \infty$ since the test based on T_n is consistent. This implies that the second term converges to 0. Observe that for any two events A_1, A_2, $\Pr\left(A_1 \cup A_2\right) \leq 1$ implying

$$\Pr\left(A_1 \cap A_2\right) \geq \Pr\left(A_1\right) + \Pr\left(A_2\right) - 1. \qquad (8)$$

So the first term in (7) is at least

$$\begin{aligned}
&\Pr\left(T_n \geq c | g \notin \mathcal{R}(A)\right) - 1 \\
&+ \Pr\left(\hat{g}'\hat{p} - \frac{\hat{p}'\hat{\Sigma}\hat{p}}{\sqrt{n}} z_{\alpha/2} \leq v \leq \hat{g}'\hat{p} + \frac{\hat{p}'\hat{\Sigma}\hat{p}}{\sqrt{n}} z_{\alpha/2} | g \notin \mathcal{R}(A)\right)
\end{aligned}$$

which converges to $(1-\alpha)$ as $n \to \infty$ since the test based on T_n is consistent. So

$$\Pr\left(v(g) \in \hat{I} | g \notin \mathcal{R}(A)\right) \geq 1 - \alpha \tag{9}$$

and so $\hat{I}$ is at worst too conservative.

For values of g such that $rank(B) = M$, i.e. $g \in \mathcal{R}(A)$, we have

$$\begin{aligned}
&\Pr\left(v(g) \in \hat{I} | g \in \mathcal{R}(A)\right) \\
= &\Pr\left(T_n \geq c, \hat{g}'\hat{p} - \frac{\hat{p}'\hat{\Sigma}\hat{p}}{\sqrt{n}} z_{\alpha/2} \leq v \leq \hat{g}'\hat{p} + \frac{\hat{p}'\hat{\Sigma}\hat{p}}{\sqrt{n}} z_{\alpha/2} | g \in \mathcal{R}(A)\right) \\
&+ \Pr\left(T_n < c, \hat{v} - \frac{d_H}{\sqrt{n}} \leq v \leq \hat{v} - \frac{d_L}{\sqrt{n}} | g \in \mathcal{R}(A)\right).
\end{aligned}$$

Using (8), the second probability is at least

$$\begin{aligned}
&\Pr\left(T_n < c | g \in \mathcal{R}(A)\right) + \Pr\left(\hat{v} - \frac{d_H}{\sqrt{n}} \leq v \leq \hat{v} - \frac{d_L}{\sqrt{n}} | g \in \mathcal{R}(A)\right) - 1 \\
&\underset{n\to\infty}{\to} \left(1-\alpha'\right) + (1-\alpha) - 1 \\
= \; &1 - \alpha - \alpha',
\end{aligned}$$

where α' is the size of the test of the null $g \in \mathcal{R}(A)$ using T_n. Therefore,

$$\Pr\left(v(g) \in \hat{I} | g \in \mathcal{R}(A)\right) \geq 1 - \alpha - \alpha'. \tag{10}$$

From (9) and (10), the conclusion follows. ■

The above proof shows the natural trade-off between the power of the test of uniqueness based on T_n and the uniform coverage probability of the pretest confidence interval. The larger is α', the smaller the coverage probability and higher is the power and conversely.

Clearly, one can think of alternative pretest estimators and confidence intervals. If additivity cannot be rejected, one can simply pick any $\check{p} \in \mathcal{P}$, calculate $\hat{g}'\check{p}$ and propose that as an estimate of the (common) maximum value. This makes the overall estimator another pretest type estimator, given by

$$\check{\theta} = 1(T_n \geq c)\,\hat{g}'\hat{p} + 1(T_n < c)\,\hat{g}'\check{p}.$$

The corresponding confidence interval is given by

$$1\left(T_n \geq c\right) \times CI_{unique} + 1\left(T_n < c\right) \times \left[\hat{g}'\check{p} - \frac{\check{p}'\hat{\Sigma}\check{p}}{\sqrt{n}} z_{\alpha/2},\, \hat{g}'\check{p} + \frac{\check{p}'\hat{\Sigma}\check{p}}{\sqrt{n}} z_{\alpha/2}\right].$$

The decision theoretic comparison of these alternative estimators in terms of their risk and confidence intervals in terms of their (uniform) coverage probabilities, the "optimal" choice of α, α' etc. are outside the scope of the current paper and is reserved for future research.

6 Application

I now apply these methods to calculate optimal allocation of freshmen to dorm rooms based on observed outcomes of a random assignment process at Dartmouth College. The two outcomes I will consider are (i) freshman year GPA and (ii) eventual enrolment into a fraternity or sorority. The two separate covariates that I will use to design the optimal matching rule will be (i) an academic index, called ACA in this paper henceforth, which is a weighted sum of the student's high school GPA, high school class rank and SAT score and (ii) race classified as white and others. I will use the observed marginal distribution of covariates in the sample as a benchmark marginal distribution. But I emphasize that I still adhere to the theoretical set up which assumes that this marginal distribution is known exactly to the planner and an optimal matching rule maps each fixed marginal to an optimal joint distribution. The analysis will be done separately for men and women and will be restricted to individuals who were assigned to double rooms. Table 0 contains summary statistics for variables used. For all other details about the background and the assignment process, please see Sacerdote (2001).

First consider the case where the policy covariate is ACA. This variable assumes values between 151 and 231 in the data. I impose discreteness by dividing the sample into several ranges of ACA and show results for 2, 4, and 6 categories. Finer categorization yields larger maxima and smaller minimum values at the optima but leads to lesser precision since finer categories imply less precise estimates of the conditional means $\hat{g}$ within each category. The results are shown in table 1 separately for men and women when the outcome of interest is mean freshman year GPA and in table 2A, 2B for the outcome "joining a fraternity/sorority in junior year".

The second column in each table reports the value of the test statistic T_n and the 90% and 95% critical points of the corresponding $\varkappa^2$ distribution (df=1, 6 and 15 for 2, 4 and 6 categories respectively). When I cannot reject the null of additive effects, I report the maximum value calculated from the sample, the bias corrected one next to it, a 95% confidence interval corresponding to $CI_{non-unique}$ in the text, and the difference between the sample max and the sample min values. "95%" within "" marks refers to the idea that with pretesting, the exact coverage probability will typically differ from the nominal level by at most the size of the test of uniqueness. When the outcome of interest is freshman year GPA, it is seen from table 1 that we cannot reject the null of additive effects. The value of the test statistic is always smaller than the 90% critical value. However, as the number of categories increase, the difference between the maximum and the minimum calculated on the basis of the sample increases as expected.

In table 2A, the same exercise is repeated for the outcome "joining a Greek organization". It is seen from table 2A that the test statistic for additivity is large in almost all cases, leading to rejection of additivity. I report the 95% confidence interval for the minimum value (CI_{unique} above) together with the maximum and the minimum values. Concentrating on the panel for women, it is seen that the additivity is rejected for both small (2) and large number (6) of categories but not for the intermediate one (4). This is a consequence of the fact that one loses precision as the number of categories rises leading to a decrease in the value of the test statistic but it becomes harder to "fit" a b to a larger number of $\hat{g}$'s as described in the definition of T_n.

Table 2B describes the nature of the optimizing allocations. Columns 3 and 4 of table 2B report the fractions of rooms with two highest ACA types and two lowest ACA types. For instance, in the first row (0.51, 0.49) means that the allocation which achieves the maximum probability of fraternity enrolment is where 51% of the rooms have two high types and 49% of the rooms have two low-types (implying that no room is mixed). For men, as can be seen from the last two columns of table 2A, the minimum probability of joining a fraternity is achieved when no room has two students from the bottom category and, few rooms with two students form the very top category. This happens because a very low type experiences a larger decline in its propensity to join a Greek house when it moves in with a high type relative to how much increase the high type experiences when he moves in with a low ACA type. Overall, this can be interpreted as a recommendation for "more mixed rooms" for men, if the planner wants to reduce the probability of joining Greek houses.

For women, the picture is exactly the opposite- more segregation in terms of previous academic achievement seems to produce lower enrollment into Greek houses. The most likely explanation for this might be that women utilize Greek houses in different ways than men and look to them for psychological "comfort". When forced to live with someone very different from her, she seeks a comfort-group outside her room and becomes more likely to join a sorority which contains more women "like her". For men, peers like oneself reinforce one's tendencies and this effect dominates.

Similar exercises with race are reported in tables 3 and 4. The results here are sensitive to the definition of race. I first consider allocation based on the dichotomous covariate whether the student belongs to an under-represented minority (Black, Hispanic, Native Indian and Asian Indian) or not. These results are reported in table 3 when the outcome is joining a Greek house. There seems to be significant nonadditivity and optimal solutions are very similar to the ones obtained with ACA as the covariate. The mean probability of joining a sorority is minimum when segregation is maximum and that of joining a fraternity is minimum when dorms are almost completely mixed with no two individuals from the minorities staying with one another.

Next I consider the case where race is classified as white and others. There does not seem to be any evidence of non additivity in race related peer effects with this definition, as is evident from table 4. Finer categorization of "others" into "blacks" and "others" caused problems in estimation since the number of rooms with two black men is one- which makes it impossible to estimate the requisite variance.

In either case, no significant nonadditivites were found corresponding to outcome being freshman year GPA.

7 Conclusion

... to come

8 References

References

[1] Cooley, Jane (2006): Desegregation and the Achievement Gap: Do Diverse Peers Help?, Duke University.

[2] Cragg, J. G and Donald, S. G. (1997): Inferring the Rank of a Matrix, Journal of Econometrics, vol. 76, no. 1-2, Jan.-Feb. 1997, pp. 223-50.

[3] Dehejia, Rajeev H (2005): Program Evaluation as a Decision Problem, Journal of Econometrics, vol. 125, no. 1-2, pp. 141-73.

[4] Graham, B., G. Imbens and G. Ridder (2005): Complementarity and Aggregate Implications of Assortative Matching: A Nonparametric Analysis

[5] Hirano, K. and J. Porter (2005): Asymptotics for Statistical Treatment Rules. Manuscript.

[6] Hollander, F. (2000): Large Deviations, American Mathematical Society.

[7] Hoxby, C. and G. Weingarth (2006): Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects. Manuscript.

[8] Luenberger, D. (1984): Linear and nonlinear programming. Addison Wesley.

[9] Manski, C. (2004):Statistical Treatment Rules for Heterogeneous Populations, Econometrica, vol. 72, no. 4, pp. 1221-46.

[10] Manski, C. (1995): Identification problems in the social sciences, Cambridge and London: Harvard University Press, 1995

[11] Sacerdote, B. (2001): Peer Effects with Random Assignment: Results for Dartmouth Roommates, Quarterly Journal of Economics, vol. 116, no. 2, May 2001, pp. 681-704

Table 0: "Population" Characteristics

Women, N=428

Variable	Mean	SD	Range Min Max
ACA	202.68	12.65	156, 228
White	0.69	0.46	0, 1
Fresh GPA	3.23	0.39	1.56, 4.0
Soro	0.47	0.5	0, 1

Men, N=436

Variable	Mean	SD	Range Min Max
ACA	205.55	12.99	151, 231
White	0.72	0.45	0, 1
Fresh GPA	3.15	0.45	1.15, 3.9
Frat	0.53	0.5	0, 1

Table 1

Y=Freshman GPA
X=ACA

# Categories	test-statistic	Critical values 90%, 95%	Sample Max	Bias Corrected Max	"95%" CI	Max-Min
Men (mean=3.15)						
2	0.59	2.71, 3.84	3.17	3.16	3.13, 3.19	0.023
4	7.4	10.64, 12.59	3.205	3.17	3.14, 3.20	0.11
6	20.6	22.31, 25.00	3.293	3.25	3.21, 3.27	0.233
Women (mean=3.235)						
2	0.01	2.71, 3.84	3.238	3.23	3.21, 3.25	0.004
4	3.98	10.64, 12.59	3.275	3.25	3.22, 3.28	0.15
6	20.23	22.31, 25.00	3.347	3.3	3.27, 3.33	0.189

Table 2A

Y=Prob of joing frat
X=ACA

# Categories	test-statistic	Critical values 90%, 95%	Sample Min	"95%" CI	Max-Min
Men (mean=0.53)					
2	3.05*	2.71, 3.84	0.487	0.417, 0.557	0.084
4	9.82*	9.64, 12.59	0.463	0.357, 0.571	0.111
6	26.9**	22.31, 25.00	0.411	0.297, 0.524	0.276
Women (mean=0.47)					
2	3.65*	2.71, 3.84	0.423	0.355, 0.492	0.09
4	6.7	10.64,12.59	0.399^	0.335, 0.399	0.106
6	35.9**	22.31, 25	0.322	0.238, 0.407	0.372

^ BC min=0.3674

Table 2B

Y=Prob of joing frat
X=ACA

# Categories	test-statistic	HH and LL Max	HH and LL Min
Men (mean=0.53)			
2	3.05*	0.51, 0.49	0.037, 0.00
4	9.82*	0.01, 0.24	0.00, 0.00
6	26.9**	0.35, 0.002	0.00, 0.00
Women (mean=0.47)			
2	3.65*	0.037, 0	0.52, 0.48
4	6.7		
6	35.9**	0, 0.016	0.011, 0.22

Table 3A: Race=Non-minorities (L), minorities (H)

	test-statistic	Critical values 90%, 95%
Y=Freshman GPA		
Men (mean=3.15)	0.253	2.71, 3.84
Women (mean=3.24)	0.104	2.71, 3.84

Table 3B: Race=Non-minorities (L), minorities(H)

	test-statistic	Critical values 90%, 95%	Sample Min	"95%" CI	Max-Min
Y=Prob of joining frat					
Men (mean=0.53)	3.3*	2.71, 3.84	0.3	0.23, 0.36	0.077
women (mean=0.47)	4.66**	2.71, 3.84	0.09	0.05, 0.13	0.04

Table 3C: Race=Non-minorities (L) , minorities (H)

	test-statistic	HH and LL Max	HH and LL Min
Y=Prob of joining frat			
Men (mean=0.53)	3.3*	0.114, 0.885	0, 0.77
women (mean=0.47)	4.66**	0, 0.78	0.11, 0.89

<u>Table 4: Race=White,Others</u>

	test-statistic	Critical values 90%, 95%
Y=Freshman GPA		
Men (mean=3.15)	0.28	2.71, 3.84
Women (mean=3.235)	0.33	2.71, 3.84
Y=Prob of joining frat		
Men (mean=0.53)	0.26	2.71, 3.84
women (mean=0.47)	1.27	2.71, 3.84