`RItools`: Software for balance testing using randomization inference.

Jake Bowers and Ben B. Hansen

University of Illinois @ Urbana-Champaign and University of Michigan @ Ann Arbor

`jwbowers@illinois.edu` and `bbh@umich.edu`

Abstract

Good statistical methods *are (1) easy to use and understand for methodologists and (2) allow nonspecialists to quickly and accurately appraise the quantitative evidence those specialists produce.*

Matching is simple and thus appeals to non-technical audiences. But whether or not a given match is in some senses "good" is not so clear. Clearly, we want any matching procedure to produce sets of "treated" and "control" units which are "similar" or "balanced" enough. But what is the standard to which we ought to compare a given measure of balance? How should we diagnose the effectiveness of a given matching in reducing bias?

In Hansen and Bowers (2008) we proposed "the randomized experiment" as such a standard and we developed statistical tests to allow quick and easy comparisons of matchings against this benchmark. In this poster we describe the `RItools` software that we have written to make those tests broadly available to applied researchers.

The Latently Randomized Experiment and Why It Is a Useful Standard for Balance Assessment

The Potential Outcomes Approach to Causal Inference

An intervention, Z, is posited to change an outcome Y.

Potential outcome models stipulate that to each unit i and level of intervention z corresponds a definite value of y, observed if and only if $Z_i = z$. For the basic case that $Z \in \{0, 1\}$, $Y_i(z = 1)$ tells us how unit i would act in the presence of a treatment, and $Y_i(z = 0)$, tells us how unit i would act in the absence of a treatment[1]

The treatment effect on unit i is a difference of its potential outcomes at different levels of z: for example, $\tau_i = Y_i(1) - Y_i(0)$.

A (possibly biased) estimate of the treatment effect for i can be had by comparing i's observed outcome, $Y_i^{\text{obs}} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$, with that of some other unit j receiving a different treatment condition, $z_i \neq z_j$, but deemed similar to i in other relevant respects.

A *valid* estimate of the treatment effect requires an argument in favor of the claim that the only difference between i and j which is consequential for Y is Z.(Also known as an assumption about ignorability.)

Balance characterizes a (simple) causal comparison when

$$Z \perp \mathbf{X}. \tag{1}$$

as in a **simply randomized experiment**.

A latently randomized experiment is an observational study accurately described by

$$Z \perp (Y(z) : \text{all } z) | \mathbf{X}. \tag{2}$$

It is often practically useful to find reductions $\mathbf{X}^*$ of $\mathbf{X}$, particularly matchings or stratifications based on it, such that $Z \perp (Y(z)) | \mathbf{X}^*$ should hold if (2) does.

How balanced is balanced enough?

Answer #1 As balanced as possible.

Answer #2 When after matching, each adjusted baseline difference on each variable is smaller than some fixed multiple of a pooled s.d. — like $< .2$ s.d. (Rosenbaum and Rubin, 1984).

Answer #3 At least as close as would be the case in a (stratified or matched) randomized experiment:

$$Z \perp \mathbf{X} | S. \tag{3}$$

This doesn't preclude more balance. But establishes a minimum standard.

Each attempts in a different way to answer: "Close enough not to bias estimates of treatment effects."

We prefer answer # 2.

The Test Statistics: d for one x and d^2 for all $\mathbf{X}$

Hansen and Bowers (2008) argue that the optimal test statistic for detecting imbalance on a given x is $d = \hat{\beta}_z = (Z^T Z)^{-1}(Z^T[x s_2 \ldots s_k])$ where $s_1 \ldots s_k$ are strata indicators.

Note: This regression does **not** assume that $\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$). Nor do we assume that $E(x) = \beta[Z s_2 \ldots s_k]$.

Our test compares d to a null distribution that comes only from (3).

Equation 3 also implies a joint (χ^2) distribution for all of the d across all of the relevant covariates. This test assesses balance on all of the linear combinations of all of the covariates. For most balance testing we recommend this omnibus test, d^2.

Example: The Effect of Deadlines on Drug Safety

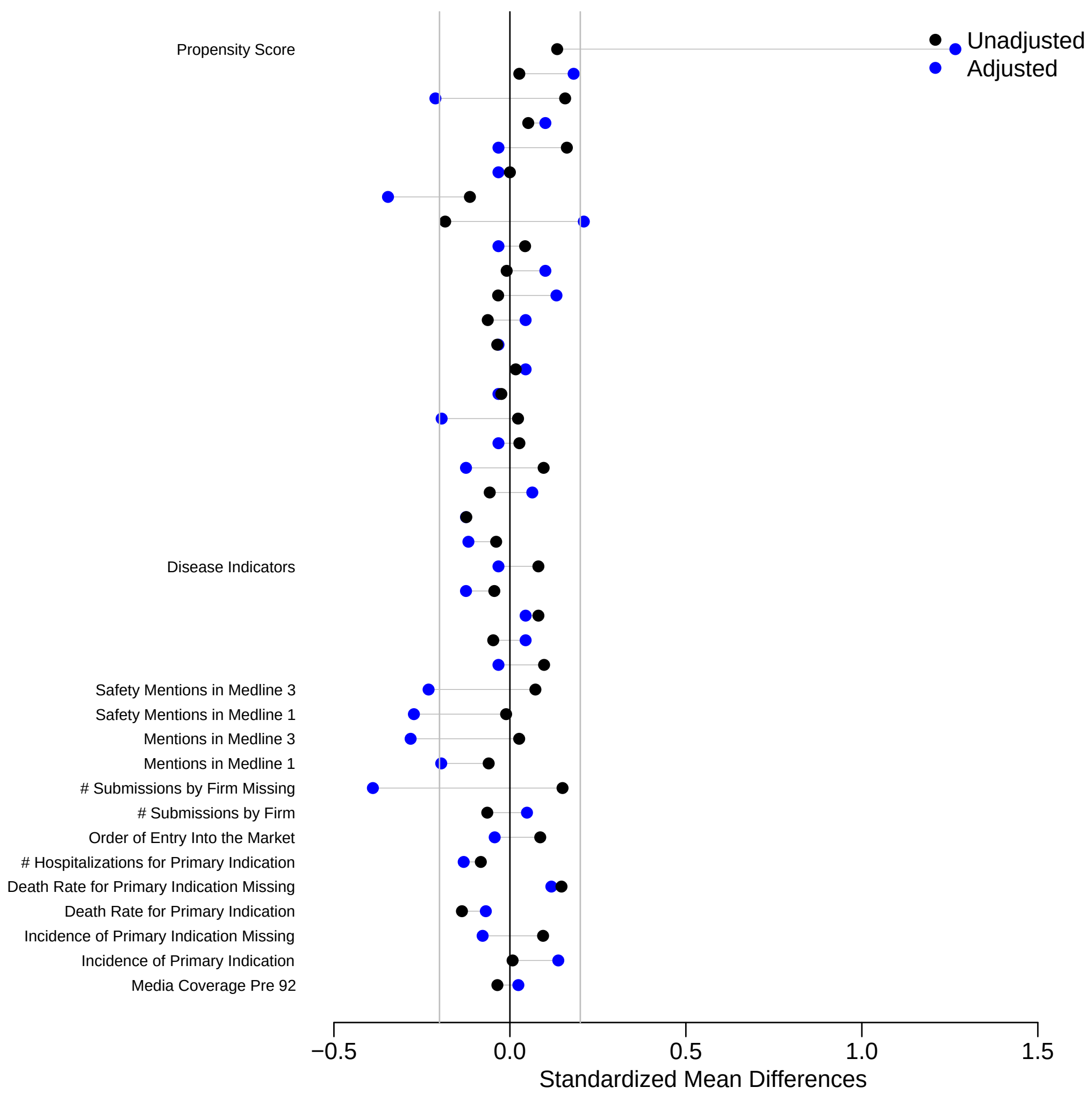
The Prescription Drug User Fee Act (PDUFA) of 1992 required the US Food and Drug Administration to act on 90% of "standard" drugs within 12 months. This was the first time that time pressure became a part of the assessment of drug safety in the USA. Our question is: **Does Haste make Waste?** Olson (2002,2004) says yes. Grabowski and Wang (2006) say no.

We compare submitted and approved during PDUFA I to drugs submitted and approved in the 4 years before PDUFA I to assess this claim.

Here we show the results of our balance assessment before and after full, optimal matching (Hansen, 2004) stratified by "priority" versus "standard" drug-type.

```
library(RItools);library(optmatch)
thefm<-fullmatch(pscoredistlist)
good<-names(thefm)[matched(thefm)]
thefmbal<-xBalance(pdufa1Z~media+
                   I(incidence/1000)+<...>,
                     strata=~thefm[good],
                     data=thedata[good,],
                     chisquare.test=TRUE)
plot(thefmbal)
<...>
```

Balance Assessment #1: A Plot



Balance Assessment #2: d^2

```
print(thefmbal)
<...>
Pre:  X-squared = 68.8, df = 38, p-value = 0.002
Post: X-squared = 21.1, df = 37, p-value = 0.984
```

Notice: Even though only 3/39 d-tests on the unadjusted data returned p-values of less than .05. One might easily see this many rejections of the null in a simple randomized experiment and thus the d tests might be interpreted to mean that no further adjustment is necessary. The d^2-test, however, recommends rejecting the null of balance and thus adjusting further.

Discussion

Other applications have included clustered/group-level treatment assignment (Hansen and Bowers, 2008) and non-random non-compliance (i.e. instrumental variables) (Bowers and Hansen, 2008).

`RItools` provides an easy and fast way for analysts with binary or continuous treatment variables to assess the strength of the arguments made in favor of unconfounded comparability. It compares the observed data against that which would be observed from a randomized experiment, and, as a true statistical test, it provides information both about the divergence of the observed from the theorized (or the standard) and also about the amount of information available and useful to assess such statements about divergence.

Our Normal approximations allow us to make these tests happen quickly, and the approximation is accurate or conservative in samples of size 20 up to 30,000+. In smaller samples, or in samples with less information, exact or simulations are easily programmable in order to check the approximation.

References

Jake Bowers and Ben Hansen (2008). *RItools: Randomization Inference Tools*, 0.1x edition.

Jake Bowers and Ben B. Hansen. Attributing effects to a cluster randomized get-out-the-vote campaign. *JASA*, 2008. to appear.

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

Ben B. Hansen. Comment: The essential role of balance tests in propensity-matched observational studies. *Statist. Med.*, 27(12), May 30 2008, 2050:2054.

Ben B. Hansen and Jake Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statist. Sci.*, 23, 2008. to appear.

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

Rubin, D. B. (1986), "Comments on "Statistics and Causal Inference"," *JASA*, 81, 961–962.

[1] Assuming "SUTVA." (Rubin, 1986)

Polmeth 2008, Political Methodology Annual Conference, University of Michigan @ Ann Arbor, July 9-12, 2008