Comments on paper by Ho, Imai, King, and Stuart (version of September 7, 2004)

Ben B Hansen

29 September 2004

pp. 10–11: Researchers willing to assume that a particular parametric model (up to some unknown parameters) generated their data should specify and directly estimate this model. Preprocessing data with matching procedures is suboptimal in this situation.

This a good thing to say at the outset – there are people out there who think they know just what the functional form looks like, and it is good to separate them off from the much broader class of people who do not, and who may have use for matching.

p. 16: We can also select, duplicate, or selectively drop observations from an existing sample without bias, as long as...

as an assertion about bias, I would be willing to accept this (although a general proof of it, particularly one that applies to nonlinear models as well as linear ones, is not at all obvious to me). But it carries a suggestion of something more than a narrow claim about bias: that you won't cause any trouble by selecting, duplicating or selectively drop in observations, something to that effect. That, I think, is questionable. First, unless the treatment effect is constant across the entire range your considering, removing or duplicating observations before the analysis is likely to shift the target, that is to say the causal parameter to be estimated. With the matching strategy you describe later in the paper, one will on occasion be forced to reject treatment units as well as controls – thus, one does not alleviate this tension by restricting attention to estimation of effect of treatment on the treated parameters. Second, when one duplicates observations, one creates a horrible mess as far as estimating variances or conducting statistical tests is concerned. My impression is that this mess is often ignored; I have certainly seen cases in which it mucked up the standard errors but the authors showed no awareness of a problem. I am aware that there has been some work in the direction of making sense of that muck; if your assessment is that this work is far enough along as to be useful to the practical person, then perhaps you should provide a reference to it. Otherwise, I think you should recommend that *duplication* only be done with great caution.

p20: The propensity score tautology in our view is the main justification for using this technology in practice: The estimated propensity score is a balancing score when we have a consistent estimate of the true propensity score. We know we have a consistent estimate of the propensity score when matching on the propensity score balances the raw covariates.

I would disagree with this description of the rationale for the use of estimated propensity scores, on two counts. First, I am aware of no general theorem that says that a consistent estimate of the true propensity score is a balancing score, in virtue of its consistency. The early papers of Rosenbaum and Rubin discuss true propensity scores and estimated propensity scores of a particular, narrow class but do not discuss consistency of estimated scores of more general types; and in the econometrics literature, one finds consistency results for propensity scores estimated in particular ways — but these particular ways often differ from the most common methods of estimating propensity scores, such as those implemented in MatchIt. Consistency has to do with what happens as the sample size grows to infinity, whereas whether or not a particular score is a balancing score has to do with a particular finite sample and a particular finite n. Second, it is prima facie circuitous to raise concerns about consistency,

an asymptotic property of estimators, when asking whether a particular score is a balancing score — a finite sample property not of estimators but of stratifications. One might say that a score is a balancing score in the weak sense if matchings or stratifications on it balance covariates, and that it is a balancing score in the strong sense if matchings or stratifications on it balance both covariates and potential outcomes. Whether a score is a balancing score in the weak sense can be checked directly, whereas whether it is a balancing score in the strong sense cannot. Propensity score theory establishes that weak balancing entails strong balancing under certain stylized conditions, some referring to the method by which the score was generated and others referring to the richness of the set of covariates available. These conditions are rarely known to hold in practice, and in observational studies they can never be known to hold (because one never knows for sure that one's set of covariates is rich enough), so there is no usable theorem to the effect that weak balancing entails strong balancing. However, the conclusions of theorems that place stylized conditions are often true somewhat more broadly than under the stylized conditions; and at this point, accrued experience with propensity scores would suggest that weak balancing is about as good a guide to the presence of strong balancing as one can hope to find.

p19-22 (steps 2 and 3): Why should one match as a first resort, rather than subclassify as a first resort? Despite a personal affinity for matching, I believe that subclassification at quantiles of a propensity score is often as effective a method of incorporating that score, and is certainly much easier in most cases. If you disagree with this, it would be interesting to hear the reasons for the disagreement. And for your intended audience, presumably people who have little exposure either to matching or stratification, it might be helpful to have some justification for pursuing matching as the default.

pp 34-35: Third, the matching literature offers a large number of possible and seemingly ad hoc procedures. From one perspective, we might be concerned about the sensitivity of our results to changes in this process, just as we have been concerned with the sensitivity of causal effect estimates to parametric modeling assumptions. In our view, this is not a major issue since the right procedure is the one that maximizes balance (with an as large as possible n), no matter how many procedures we try.

Small point: Need there be one "right" procedure? I would say no. In the large, I am basically in agreement that one should be satisfied once one has achieved balance using as much of the sample as possible. I do think it is somewhat preferable to pose the question of selecting the best match as one of negotiating a trade-off between variance and bias, as this formulation generalizes better to the cases of matching with a variable number of controls and full matching. For details, see Ming and Rosenbaum's papers on matching with a variable number of controls or my paper in the current issue of the Journal of the American Statistical Association on full matching.

pp23-24: This section should at least mention the possibilities for parametric models explicitly taking matched sets into account, e.g. fixed and random effects. In the case of subclassification, such models are an appealing alternative to "running the analysis separately". Here are the references I would offer to someone seeking my statistical advice, as a function of their data type and of their preferred mode of inference.

Preferred mode of inference	Type of outcome	
	Categorical	Continuous
Randomization	Agresti (2002), *Categorical Data Analysis*; Rosenbaum (JASA, 2002)	Rosenbaum (2002), *Observational Studies*; Rosenbaum (Statistical Science, 2002)
Conditional [a]	Agresti (2002); Cox & Snell (1989), *Analysis of binary data*	ordinary OLS[b] is fine; see also Rubin (JASA, 1979)
Bayes, esp. hierarchical linear models [c]	Agresti (2002)	Raudenbush & Bryk (2002), *Hierarchical linear models*

[a] Uses a fixed effect for each matched sets.

[b] i.e., OLS with a fixed effect for each matched set plus treatment effect(s), potentially interacting with matched set indicators

[c] Uses a random effect for each matched set.