Causal Inference: Matching Assignment

Magali de Bruyn

Minerva Schools at KGI

Matching Assignment

(1)

Part A

The Match function needs to run first (before the MatchBalance function).[1]

The Y assigned to Y needs to be defined (like X and Tr) or the argument omitted.

Part B

The caliper needs to be specified in Match(): the GenMatch() and Match() function should have the same parameters.

The Y assigned to Y needs to be specified/defined in order for summary(mout) below to display the causal effect. Y should be "a vector containing the outcome of interest" i.e. the data's Ys or outcome values.

Part C

The MatchBalance function should *not* include re78 (the outcome, independent variable, or Y) in its formula. By adding it in, the function tries to determine if balance is achieved on this variable, treating it as an observed covariate. However, this is the outcome, so we should *not* be trying to balance the value between the treated units and control units; this would be effectively the same as trying to get rid of a treatment effect.

[1] The GenMatch function found the optimal weight to give each covariate in 'X' to achieve balance on the covariates; we need to now estimate our causal effect of interest using those weights through the Match function. *Only* then should we determine if the balance has actually been obtained on the variables of interest using the MatchBalance() function. This is achieved by passing “mout” for the match.out value (instead of “genout”). match.out, as described per the documentation of the MatchBalance function, should be “the output object from the Match function” so it can provide balance statistics for both before and after matching.

(2)

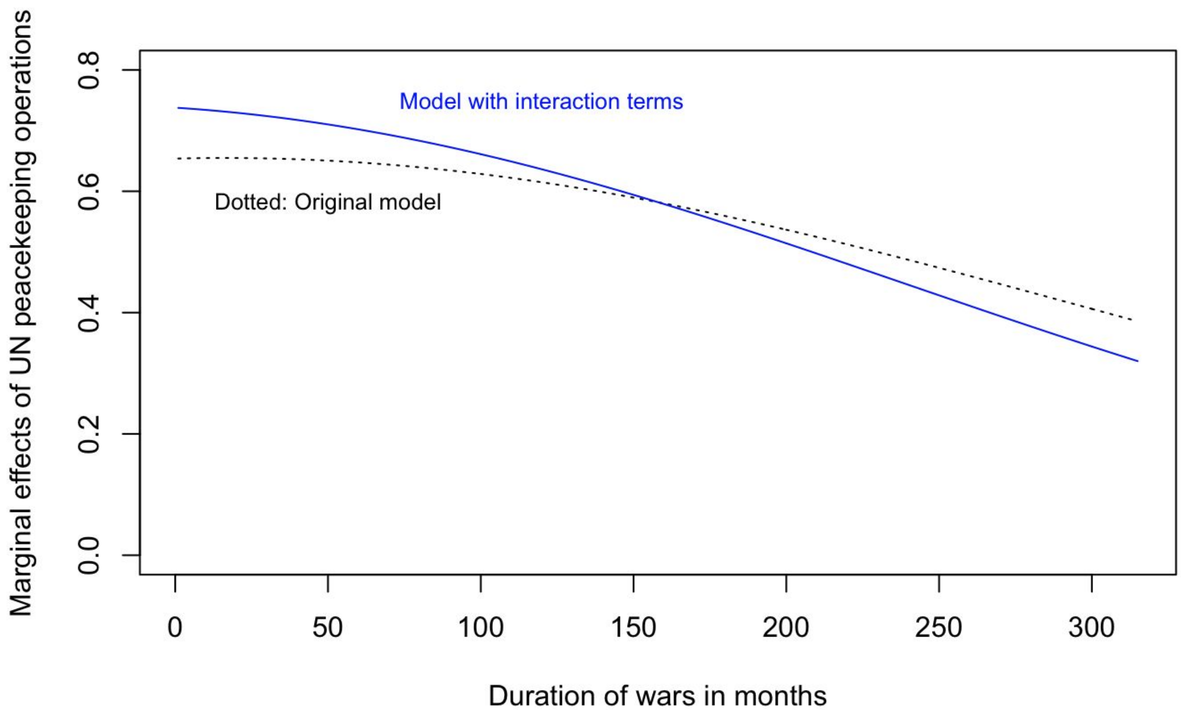

Figure 1. A replication of King & Zeng's Figure 8 but with the two other interaction terms (EXP*UNTYPE4 and WARDUR*LOGCOST). Varying interaction terms indeed provides a different model but described effects of UN intervention more similar to Doyle & Sambanis' (2000) than King & Zeng's.

(a)

The y-axis measures the marginal effects of UN peacekeeping operations; in other words, it represents the causal effect of a unit change in UN peacebuilding operations on the war duration in months (when all the other covariates are kept fixed at their means)

The x-axis measures the duration of wars in months.

(b)

Figure 8 illustrates "the extreme model dependence" that arises "from using data outside of common support that requires extrapolation," as is the case for Doyle and Sambanis (2000) analysis (King & Zeng, 2007).

(c)

My figure supports the claim that the model influences the observed relationship between the marginal effects of UN peacekeeping operations and war duration. However, it fails to support (as strongly as King & Zeng's (2007) figure) that the model shapes the conclusions drawn. Adding two interaction terms leads to a model that, similar to the original one, describes a gradual decrease in marginal effects of intervention as war duration increases; it promotes—even more strongly than the original model—the claim that the UN should focus on wars that just began (as this is where the UN has the largest possible effect).

(3)

The first line outputs a vector of a 122 (the length of foo/foo's column “uncint”) “0”s (a numeric). The second line finds the indexes of all the values in foo's column “uncint” that are not equal to “0”/None and assigns a 1 to the same indexes in Tr.

In other words, it classifies the different levels of UN intervention (i.e. “Enforcement,” “None,” “Observer,” “PKO” as either treatment (1) or control (0). All descriptors other than “None”—“Enforcement,” “None,” and “Observers”—are assigned a value of 1; and “None” is considered no treatment and assigned a value of 0.

Note: the code should use “None” instead of “0” (and add a parentheses) in order for it to work as intended when finding the indexes in foo's uncint column. Hence, Tr[which(foo$uncint != 0)] <- 1 should be Tr[which(foo$uncint != “None”)] <- 1.

(4)

(a)

How does any involvement (observation, enforcement, and traditional peacebuilding operations) from the UN impact lenient peacebuilding success, or specifically the absence of violence, 2 and 5 years after the war (Doyle & Sambanis, 2000)?

(b)

SUTVA might be violated here because the outcome of the units might be dependent on the assignment of the treatment; in other words, the peacebuilding of countries that are neighbors or have ties to the countries subject to the treatment (or not) might be affected by this assignment (e.g. more stability in the region). The restrict argument in Match()/GenMatch() allows us to define matches and specifically exclude observation pairs from being matched (through the use of negative numbers in the third column) or force them to be matched (through the use of 0 in the third column); this can be implemented to exclude countries neighboring a treated country from being matched, for example.

Alternatively, we could take into account as a covariate to match on whether the countries/observations have neighboring countries that have been treated too (and potentially the number of such neighboring countries).

(c) (i)

	Treatment effect estimate	Treatment effect standard error	p-value (from Match Balance)
Logistic regression			
Lenient peacebuilding success 2 years after	0.1265978	0.6079902	< 2.22e-16
Lenient peacebuilding success 5 years after	0.1535389	0.6289750	< 2.22e-16
Propensity score matching			
Lenient peacebuilding success 2 years after	0.02777778	0.6079902	< 2.22e-16
Lenient peacebuilding success 5 years after	0.06060606	0.6289750	< 2.22e-16
Genetic matching			
Lenient peacebuilding success 2 years after	NA	NA	0.073698
Lenient peacebuilding success 5 years after	NA	NA	0.0054657
Genetic matching with propensity scores			
Lenient peacebuilding success 2 years after	NA	NA	0.0032034
Lenient peacebuilding success 5 years after	NA	NA	0.0018742

Table 1. A summary of the estimated impact (treatment effect) of *any* involvement (observation, enforcement, and peacebuilding operations) from the UN on lenient peacebuilding success 2 and 5 years after the war using logistic regression, p-score matching, and genetic matching without and with propensity scores. "NA" indicates that the leximin p-value was below 0.10 for matching results (see Appendix B for results). The regression model include the variables: treatment (defined previously), wartype, logcost, wardur, factnum, factnum2, trnsfcap, develop, exp, decade, treaty, costcap. The functional forms of the propensity score model are: "Tr + wartype + logcost + wardur + factnum + factnum2 + trnsfcap + develop + exp + decade + treaty + costcap." The variables ultimately genetically matched on are wartype, logcost, wardur, factnum, factnum2, trnsfcap, develop, exp, decade, and treaty. The MatchBalance variables used are: "wartype + logcost + wardur + factnum + factnum2 + trnsfcap + develop + exp + decade + treaty + costcap + wartype*wartype + logcost*logcost + wardur*wardur + factnum2 + trnsfcap*trnsfcap + develop*develop + exp*exp + decade*decade + treaty*treaty + costcap*costcap + wartype*wartype*wartype + wartype*logcost + wartype*wardur + wartype*factnum + wartype*factnum2 + wartype*trnsfcap + wartype*develop + wartype*exp + wartype*decade + wartype*treaty + wartype*costcap + logcost*wartype + logcost*logcost*logcost + logcost*wardur + logcost*factnum + logcost*factnum2 + logcost*trnsfcap + logcost*develop + logcost*exp + logcost*decade + logcost*treaty + logcost*costcap + wardur*wartype + wardur*logcost + wardur*wardur*wardur + wardur*factnum + wardur*factnum2 + wardur*trnsfcap + wardur*develop + wardur*exp + wardur*decade + wardur*treaty + wardur*costcap + costcap*wartype + costcap*logcost*logcost + costcap*wardur + costcap*factnum + costcap*factnum2 + costcap*trnsfcap + costcap*develop + costcap*exp + costcap*decade + costcap*treaty + costcap*costcap*costcap" excluding for 2 years: "logcost*factnum2 + develop*costcap + factnum*costcap + logcost*treaty;" and excluding for 5 years: "wartype*exp + treaty*costcap."

(ii)

To choose covariates for the regression model, I read through Doyle & Sambanis' (2000) paper and Appendix, investigated what the variables the dataset represents mean, and understood which ones had a significant effect on outcome. This allowed me to exclude the variables "MILOUT" (military victory) and "EH" (ethnic heterogeneity); and drove me to include COSTCAP (dead/displaced per capita), in addition to those required by the assignment instructions at a minimum.

Since matching on these variables and their first-order interactions resulted in a leximin p-value from MatchBalance below 0.10, I then checked out where the imbalance came from through the KS test p-value (it was from the logcost*factnum2, develop*costcap, factnum*costcap, logcost*treaty interaction terms for 2 years and from the wartype*exp and treaty*costcap interaction terms for 5 years); and attempted caliper and exact matching for them. I checked that the resulting sample was still reflective of my overall population by looking at the size of the sample or total number of observations discarded.

Unfortunately, these attempts failed to result in a p-value above or equal to 0.10 when running MatchBalance with these variables. I attempted discarding the above low-balance interaction terms, resulting in p-values that were still below the threshold.

However, discarding achieving balance on low-balance covariates (i.e. omitting them) in order to obtain higher p-values is not appropriate because they (or their non-interactions) influence the outcome significantly (or we were requested to use them).* I therefore chose to report a model which includes the variables (and their interactions) I estimate to be important (or that were required) but whose p-values are below 0.10 over a model that fits this criteria (Appendix B). To improve the balance on the selected covariates, a larger sample size or more complete data might be necessary.

* Note: We can use backward selection to discard covariates (in the hopes of improving balance on the important ones) and check which variables have the weakest association with outcome—or the largest p-value in the regression model (above 0.10) and pruned them (Stuart, 2010; Velentgas et al., 2013); this included develop and factnum2, two covariates we were required to include in our model and therefore the method was not applicable.

(5)

TO: Ana María Menéndez (UN Secretary-General's Senior Advisor on Policy)

FROM: Magali de Bruyn

DATE: 23 November 2019

RE: Dependence of causal inference on methodological assumptions: a case study of matching and modelling applied to UN peacekeeping operations

EXECUTIVE SUMMARY

Building on analysis by Doyle & Sambanis (2000) and meta-analysis by King & Zeng (2007), evaluation of different statistical models *confirm* the model dependence of a UN peacebuilding dataset and limited possibility for achieving good balance on most covariates. I illustrate (graphically) the model dependence (or the degree to which the chosen variables influence the determined effects and correlation) by including two interaction terms in our model measuring the effect of marginal UN peacekeeping operations on war duration. Attempts to balance using propensity score matching, genetic matching, and genetic matching with propensity scores, further demonstrate the variability of the inferred effect of UN peacebuilding efforts. Evidence in the data is currently not substantial enough to support a conclusion on the effect of multidimensional UN peacekeeping operations on war duration or of *any* involvement (observation, enforcement, and peacebuilding operations) on lenient peacebuilding success 2 and 5 years from now. In order to draw valid inferences, better balance—or matches for countries—must first be achieved (Appendix B). Collection of more recent data would be helpful.

CONCLUSION

The preceding exploration provides evidence for the number of methodological choices left to the researcher and the subjectivity of consequent conclusions when the observational dataset is limited. In general, researchers should seek to minimize bias among comparisons (the treatment and control group). Currently, (genetic) matching attempts to promote this by balancing the data on covariates as measured through the highest minimum p-value among them, but this only serves as a proxy for bias. Consequently, even if almost perfect balance is achieved through (genetic) matching, an unbiased dataset is not guaranteed; RCT remain one of the only ways to guarantee this necessary assumption for causal inference. Caution with regards to reliance on data analyses is advised and peer-validation of assumptions and choices when modelling recommended.

References

Doyle, M., & Sambanis, N. (2000). International Peacebuilding: A Theoretical and Quantitative Analysis. *American Political Science Review*, 94(4), 779–801. Retrieved from http://web.worldbank.org/archive/website01241/WEB/IMAGES/INTERNAT.PDF

King, G. & Zeng, L. (2007). When Can History Be Our Guide? The Pitfalls of Counterfactual Inference. *International Studies Quarterly*, 51, 183–210. Retrieved from https://gking.harvard.edu/files/counterf.pdf

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Stat Sci*, 25(1), 1-21. doi:10.1214/09-STS313

Velentgas, P., Dreyer, N., Nourjah, P., Smith, S., & Torchia, M. (2013). Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. *Agency for Healthcare Research and Quality*, 12(13). Retrieved from www.effectivehealthcare.ahrq.gov/ Methods-OCER.cfm

APPENDIX A

Code

Please find the code for:

question 1 here.

question 2 here.

question 3 here.

question 4 here.

APPENDIX B

Results of Question 4: Matching

Final results used to complete the assignment have been copy and pasted here.