Response to Peer Review: RJournal 2021-52

Aikens et al.

May 27, 2021

To the Editor

This peer review directed our attention to a number of practical improvements which we believe have strengthened both the paper and the `stratamatch` package. The most substantial updates are listed below:

- Per reviewer 1, we have added a commentary on effect estimation with citations to the relevant literature (page 7, "A brief comment on estimation").

- In response to suggestions by both reviewer 1 and reviewer 2, the latest release of `stratamatch` now supports Mahalanobis distance matching and full matching. This update is reflected in the text.

- We have released a new vignette for the `stratamatch` package, "Advanced Functionality," which elaborates on some of the more nuanced possible usages of the package. This includes a tutorial on new matching functionality as well as flexible implementation of other matching schemes, including the propensity score caliper design suggested by reviewer 1. We also include code tutorials and suggestions on alternative modeling schemes for the prognostic score, in response to suggestions by reviewer 2.

Several other minor comments on text, citations, and package documentation have also been addressed. Below, please find our responses to individual points raised by the reviewers.

Suggestions by Reviewer 1

Summary. *The authors describe the use of the R package stratamatch, which simplifies the use of prognostic score estimation and stratification for matching large datasets in observational studies... All of my edits are minor. I have some concerns about the options available in the package, but these are also minor. Overall, I am very impressed with this paper and package and recommend its publication with a few slight reversions.*

This review was very helpful at directing our attention to several practical improvements. Thank you for your thorough inspection of the paper and the accompanying package.

Minor comments from Reviewer 1

Minor Comment: *I request that the authors include a brief section on effect estimation after stratification and matching. It's not immediately clear from the paper how effect estimation would proceed, i.e., would the standard errors have to account for the random partitioning of the pilot data, estimation of prognostic or propensity scores, stratification, or matching, all of which have the potential to add uncertainty to the effect estimate. This could be as simple as a single paragraph explaining how to do this. I am aware that the literature may not be fully settled on this, but there is definitely enough to go off for some solid recommendations (see the MatchIt vignette on effect estimation for example).*

This is a good suggestion. We have added a brief section on effect estimation to the illustrations section (page 7). Since – as this reviewer suggests – there is some disagreement in the literature on the nuances of this point, we avoid a single prescriptive recommendation and include a set of citations so that interested readers can develop their own opinion. We also reference the new MatchIt vignette on effect estimation, since we think this is a great resource.

Minor Comment: *Similarly, I think it would be a good idea and easy to add full matching and Mahalanobis distance matching as options other than k:1 optimal matching, since these two methods also rely on functions in optmatch and may perform better than optimal pair matching on the propensity score (e.g., in Aikens, Greaves, & Baiocchi, 2020).*

We agree! To make `strata_match` more flexible, we've updated it to allow full matching in addition to fixed-ratio matching, and we've added the option to use Mahalanobis distance rather than a propensity score to create the distance matrix for matching. The revised text reflects these changes, and an updated tutorial is available in the new `stratamatch` vignette on "Advanced Matching"

Minor Comment: *I do wish more methods were natively available so that the user would not have to manually program estimation of these additional values, which would defeat some of the purpose of the package's convenience. The MatchIt package provides a nice example of how straightforward it would be to add these additional methods (e.g., random forests, BART) for propensity score and prognostic score estimation.*

Thank you for the excellent suggestion. The MatchIt package provides a great example

of deeper integration with several modeling options. In our case, we tried to balance against the maintenance challenges of reliably supporting multiple package dependencies. As a design choice, we tried to maximize flexibility by allowing the user to plugin any estimation model of their choice. Nonetheless, to help users with varied backgrounds with their own implementation of an alternative modeling scheme, a large portion of the new `stratamatch` vignette is dedicated to illustrating some more advanced options for fitting the prognostic score. Therein, we also include brief intuition and suggestions of other resources for modeling in R (e.g. the `caret` package). We especially wanted to emphasize the manual use of lasso regression in the manuscript and vignette, since we think this is a particularly practical option for users who are interested in incorporating some aspect of feature selection to their modeling process, but who don't want to stray too far from traditional logistic regression approaches.

Minor Comment: *In the paper, I request that some of the citations be updated. For example, King & Nielsen has since been published in Political Analysis and the citation should reflect that. The citation for Rubin (2008) is also incorrect. "MatchIt" should be capitalized correctly since R packages are case-sensitive.*

Thank you. All of the requested corrections have been made.

Suggestions by Reviewer 2

Summary. *This paper introduces the "stratamatch" package which implements the methods proposed in Aikens et al. (2020). "stratamatch" allows researchers to easily perform the following steps: (1) divide the control group into a pilot set and analysis set, (2) fit a prognostic score model on the pilot set, (3) stratify the analysis set based on predicted prognostic scores, (4) calculate propensity scores within strata, and (5) pairmatch within strata based on propensity scores... We found use of the package, alone or in combination with existing related packages, to be straightforward and seamless. We like the emphasis on diagnostics. The article has helpful illustrative examples (one a toy example and the other real) and we found it easy to read. Still, we have a suggestion or two for the article, and we found rough spots in the software here and there.*

This review helped to catch several errors and places where the package or manuscript could be strengthened. Thank you.

Main comments from Reviewer 2

Main Comment: *One part of the article's motivation for the software is that optimal matching can scale poorly to large data sets; strata matching addresses that issue. This argument would be strengthened by presenting run times for matching on large data sets with and without strata matching.*

We agree! This is an important motivating point. We have summarized empirical runtimes of stratified and unstratified matching on differently-sized simulated data sets in figure 5. To draw the readers' attention to this result, we've added references to this analysis in the introduction and conclusion.

Main Comment: *strata_match() only seems to allow fixed-ratio matching on the propensity score. strata_match() does not allow passing in optional arguments. More complex pair matching (perhaps including calipers on the propensity score or matching on another covariate in addition to the propensity score) is not compatible with strata_match(). Perhaps you'd deem calipers on the propensity score to be unnecessary, as propensity score matching already occurs within strata. Still, what advice would you offer to users who want to combine benefits of your auto-stratification with matching devices not reflected within strata_match() options?*

This is a great comment. To make `strata_match` more flexible, we've updated it to allow full matching in addition to fixed-ratio matching, and we've added the option to use Mahalanobis distance rather than a propensity score to create the distance matrix for matching. We have not yet implemented propensity score caliper options for `stratamatch`. The reason for this stems from what this reviewer suggests: there is a partial redundancy between stratification and the addition of a caliper that makes the combination somewhat less attractive for matching. As a result, we've opted not to include calipers as an option thus far to keep the arguments to `strata_match` as straightforward as possible.

However, we think that the caliper matching design makes a great pedagogical example of a matching scheme that a user might reasonably want to implement by hand with `optmatch` after stratifying a dataset with `stratamatch`. For this reason, we've included a code tutorial on

this design in the new `stratamatch` vignette on "Advanced Functionality." This speaks to the broader point that this reviewer is making: Even as we increase the flexibility of `stratamatch`, there will inevitably be users who want something very specific out of their matching that we haven't implemented yet. For these researchers, passing the results of `auto_stratify` to another package such as `optmatch` or `MatchIt` may give them the flexibility they desire in the matching stage while giving them acess to the stratification support and diagnostics supplied by `stratamatch`. We hope that the caliper matching example in the new vignette will be a good introduction to help users get started implementing creative matching schemes with a prognostic score stratification.

Minor comments from Reviewer 2

Minor Comment: *As an "Potential Issues" entry in an issues_table, "small treatment:control ratio" seems less question-begging than "Not enough treatment samples". Similarly, it seems better simply to flag that a stratum has a small number of treatments and/or controls than to declare that it as "Too few samples" – strictly speaking, as long ask there's one of each there's enough to make a comparison.*

This is a good catch. We have changed the implementation to replace "Not enough treatment samples" with "small treatment:control ratio". To the second point, our focus for the "Too few samples" flag stems from a concern that a stratum too small will result in poor match quality. Even if there is a favorable ratio of treated to control individuals (say, 1:1), it may be difficult to find high-quality matches within a stratum that contains very few samples overall. We agree that this issue does not guarantee that the stratification produced is unusable; it is merely a red flag that issues may arise with matching downstream. We've updated the documentation for `manual_stratify` and `auto_stratify` to clarify this label, and we've adjusted the manuscript text to clarify this point where the label is first introduced.

Minor Comment: *Also, the Potential Issues flags for the two situations above ought to mark the distinction between them more sharply: "Too few samples" and "Not enough treatment samples" could get mixed together in the user's memory, particularly when they don't appear side-by-side in a single report.*

This is a good point! The error message for "Not enough [treatment/control] samples" has been updated to reference the *ratio* rather than the *number* of samples (see above), which should help to make these issue labels more distinct.

Minor Comment: *Not certain, but the following passage of the stratamatch() documentation seems potentially misleading:*
```
(optional) formula for propensity score.  If left blank, all columns of the
dataset will be used as covariates (except outcome, treatment and stratum)
```
Is "the dataset" the entirety of the data= argument that was fed to *_stratify(), or only the columns of that data frame that were specifically referenced in the *_stratify() call? We glean it's the latter. It'll help your users to disambiguate within the documentation.

It's actually the former. We have updated the documentation to make this more clear.

Some calls to `auto_stratify` don't specify covariates (e.g. those in which the prognostic scores are provided a priori), so the whole analysis set is used by default when the propensity model formula is not specified in `strata_match`.

Minor Comment: *We liked the "Assigment-Control" plot. Is this the same as what Aikens, Greaves, and Baiocchi (2020) call the "Fisher-Mill plot"? If so, the connection seems appropriate to mention in the documentation for the plot method.*

Yes, they are the same. Preprint versions of Aikens, Greaves, and Baiocchi used the phrase "Fisher-Mill plot" for these visualizations. However, in the final published manuscript in *Statistics in Medicine* these plots are instead called "Assignment-Control plots". The documentation for `plot.auto_strata` now references the updated text in the "See Also" section, with a brief note about the name discrepancy.

Minor Comment: *Speaking of which, it looks like you have documentation for plot.strata, but nothing for plot.auto_strata. Seeing that the example object had type auto_strata, I typed ?plot.auto_strata and didn't get anything. Can you arrange to have my help query pull up your plot.strata help page, perhaps by declaring aliases in an appropriate place?*

This is a great catch. We've updated the package so that the commands `?plot.auto_strata` and `?plot.manual_strata` each pull up their own documentation page.

Minor Comment: *We once received the error "Error in if (cj == upper) next : missing value where TRUE/FALSE needed". As far as I could tell, the error arose because the size value we passed into "auto_stratify" was too small, possibly creating strata without treatment observations. A more informative error message would have helped me along here.*

Thanks for finding this. We've replicated your bug and updated the error handling to print a more informative message.