Uncertainty in the Assignment-Control space

Rachael Caelie (Rocky) Aikens

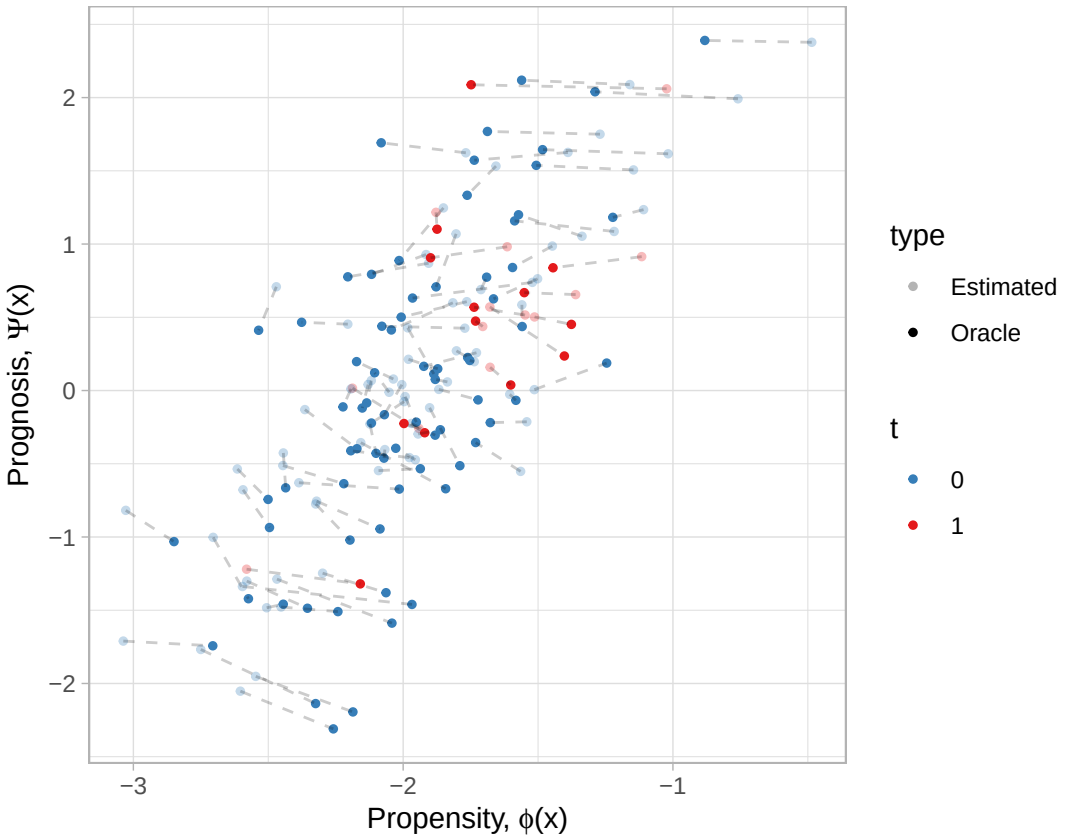
1/15/2021

Uncertainty in score models

It's important to remember that when we visualize an assignment-control plot, the coordinates of each point are estimated quantities. How do we appropriately address uncertainty in our score models when we use assignment-control plots or match in the assignment-control space?

Illustration: Comparing estimated and actual scores

As an illustration, the following plot shows the locations of a subset of indiviuals in assignment-control space. Opaque points show the true location of each observation, while translucent points show the estimated location of each observation based on fitted score models. The actual and estimated locations of each observation are connected by a dotted line.



What uncertainty?

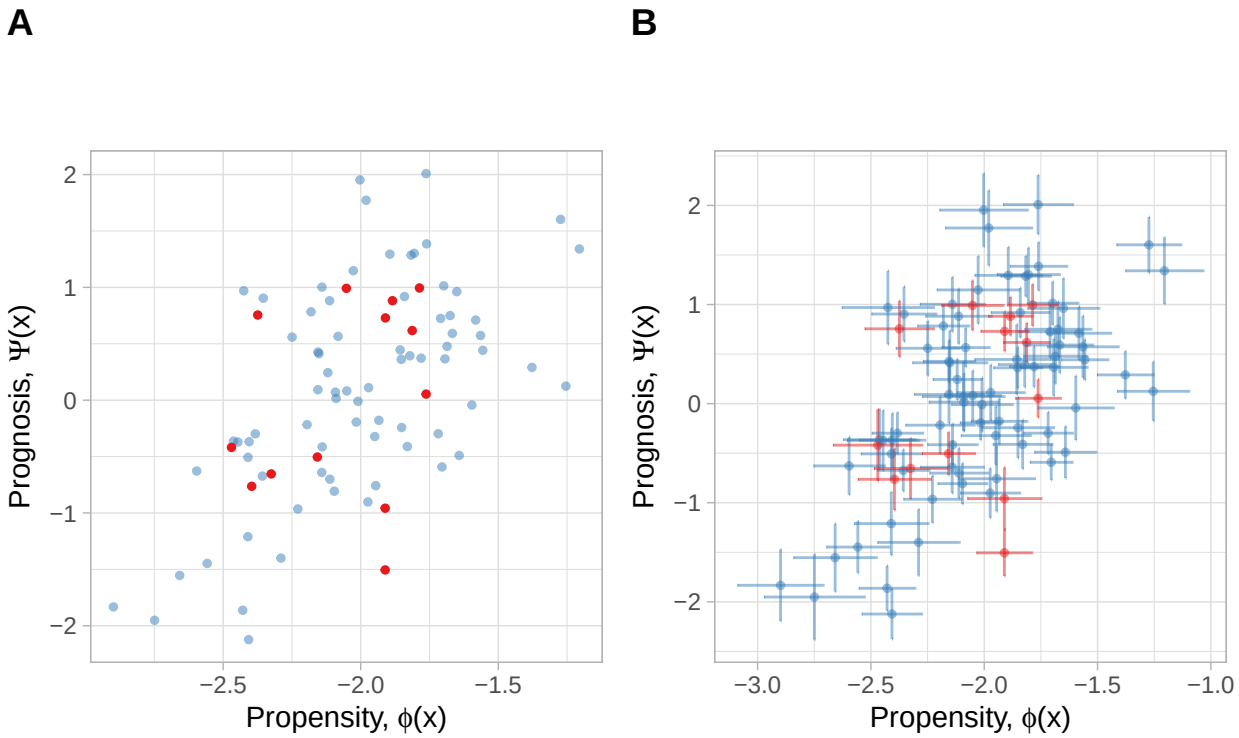
Before we get started it's important to address what we even mean by uncertainty. This is a more nuanced point than one might expect! Much of the matching literature calculates variance in the causal effect conditional on the matched pairs – the only thing that is considered "random" is the treatment assignments within matched subsets. This is to say, the traditional matching framework doesn't deal with any variability associated with sampling or match selection. So, what does uncertainty even mean before we've matched?

A common way to proceed might be to say that what we mean by "addressing uncertainty" is really accounting for sampling variability. This is a little weird because once the data set is matched, we will immediately forget everything we supposed about sampling variability. Pressing forward nonetheless, we could suppose that our dataset is a sample from some superpopulation, and we estimate imperfect score models because we are only able to look at a sample, not a superpopulation. We might then try and obtain parametric or bootstrap confidence intervals about our model parameters. It is important to note that the *confidence intervals* about the fitted values describe the estimates we care about, not the *prediction intervals*. In this case, we only care about the model surface (i.e. the expected value conditional on the observed covariates), *not* the standard errors about the model surface that a good data scientist would consider when predicting for future data points.

Even when we restrict our discussion to sampling variability, estimating sampling variability in the prognostic score is not so simple. The prognostic score is fit on the pilot set, which is a subset of the full sample. We'd need a framework for talking about variability for using this two-tiered sampling process. Unfortunately (but also interestingly?) the typical nonparamertic bootstrap approach for this problem will probably behave in unpredictable ways. Since the pilot set is itself selected using a 1:2 Mahalanobis distance matching process, the identical observations that inevitably result from bootstrapping with replacement will probably be matched to themselves, possibly resulting in strange statistical behavior. In this case, a semiparametric or parametric approach might be useful, although even semiparametric bootstrapping might fall prey to the same problem depending on the smoothing parameters used (???).

Confidence "Stars"

The plots below give a rough illustration of what uncertainty intervals might look like in an assignment control space. Panel A shows the locations of the points, and panel B shows each point with vertical and horizontal confidence bars. To avoid overplotting, only a subset of the data is shown. The confidence intervals are the usual parametric intervals from the propensity and prognostic models. This isn't perfect of course because parametric confidence intervals are not necessarily desirable and the parametric intervals and the parametric prognostic score intervals don't necessarily deal with the nuances of sampling variability described above. Also, of course, an aggregation of 95% confidence intervals doesn't necessarily retain it's statistical meaning when interpreted together.



Panel B is a little bit hard to read, even though it only shows a subset of the data. However, there are some important ideas we can take away from this illustration.

- Confidence intervals in general can be a useful reminder that the points in an assignment-control plot are not estimated with uncertainty
- The relative sizes of the horizontal and vertical confidence intervals may convey the relative amounts of uncertainty in the prognostic and propensity dimension
- The sizes of the confidence intervals compared to the spread of the points can convey how well we understand the shape of the data. Are the things we see as outliers certainly outliers? How well do we understand apparent correllations and overlap?

Other Examples