New IV Setup

Rachael Caelie (Rocky) Aikens

3/24/2021

Setup

While the prior formulation of the IV simulations is valid, it is somewhat simplified. Often, the IV is correllated with other measured covariates, requiring adjustment. Here, we tweak the simulation set-up in a minor way to add this layer of complexity.

We take the same set-up as before:

$$\begin{aligned}
X_i &\sim Normal(0, I_{10})\\
T_i &\sim Bernoulli\left(\frac{1}{1+exp(\phi(X_i))}\right)\\
Y_i(0) &= \Psi(X_i) + \epsilon_i\\
\epsilon_i &\sim N(0,1)
\end{aligned}$$

where $\phi(X)$ and $\Psi(X)$ represent the true propensity and prognostic score functions, given by

$$\begin{aligned}
\phi(X_i, Z_i) &= c_1 X_{i1} + c_2 Z_i - c_0,\\
\Psi(X_i) &= \rho X_{i1} + \sqrt{(1-\rho^2)} X_{i2}.
\end{aligned}$$

Where c_2, c_1, c_0, and ρ are constants. As before, Z is an instrumental variable in that it is unconfounded and has no effect on the outcome except via the treatment. However, we will additionally suppose that the correllation between Z and X_1 is given by some constant, ρ_z.

An easy way to fabricate this is to set:

$$Z_i = \rho_Z X_{i1} + \sqrt{1-\rho_z^2}\zeta_i$$

Where the ζ_i are iid standard normal and completely independent of all other covariates. In this setup, ζ might be thought of as the "pure" IV, while Z represents an IV that - while still a valid instrumental variable - is only independent of the outcome given the treatment assignment *and* the observed covariates. Importantly, while Z is measured, the "pure" IV ζ is unmeasured.

A Two-Stage Least Squares Approach

Using the notation in the set-up, a two-stage least squares approach to estimating the treatment effect might be:

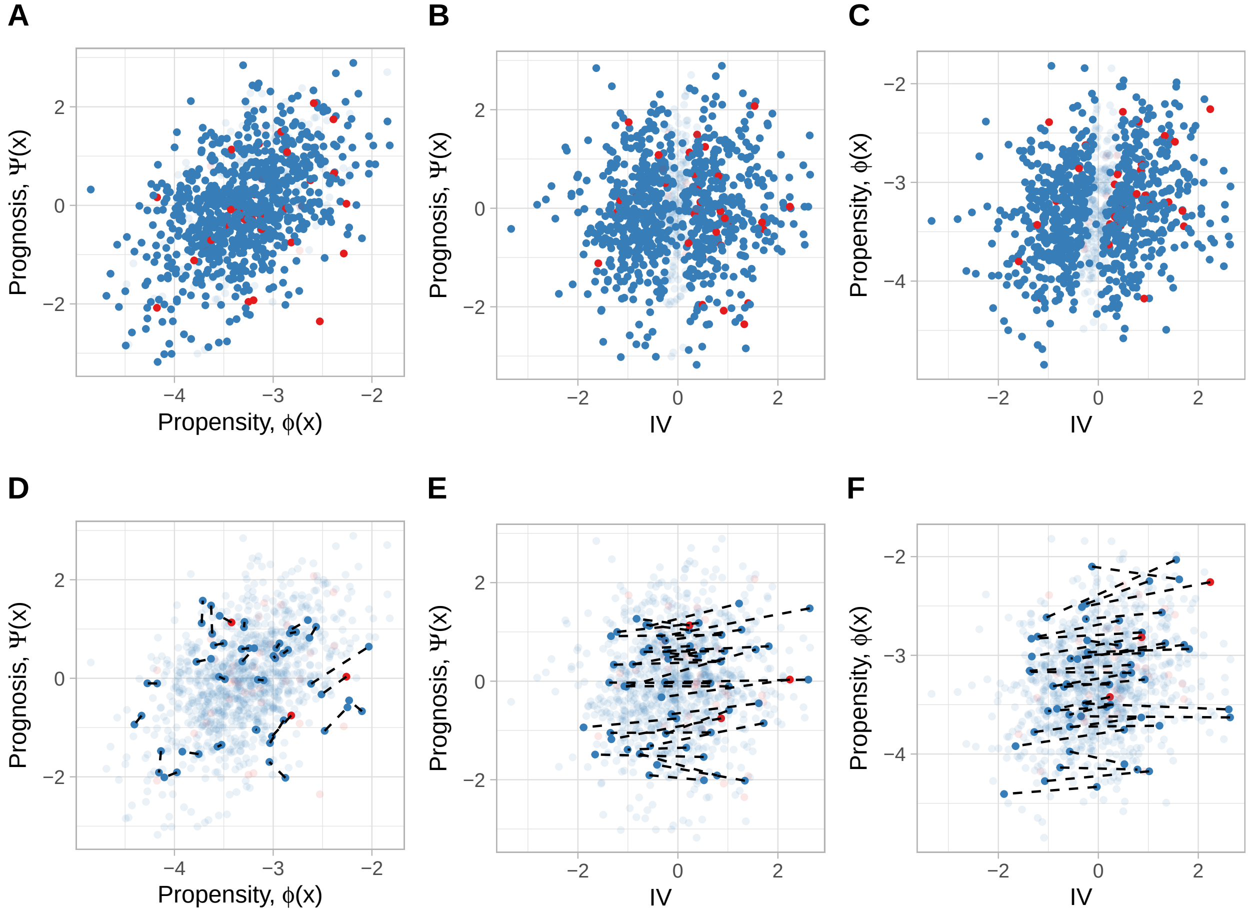
1. Regress $T_i \sim \beta_z Z + \beta^t X$ to obtain $\hat{T}_i$. Notice that an ideal regression would have $\hat{T}_i \approx \frac{1}{1-e^{-\phi(X_i,Z_i)}}$. This $\hat{T}_i$ might be thought of as a "cleaned" treatment assignment based on all the measured covariates. We suppose that deviations between T_i and $\hat{T}_i$ may capture unobserved confounding.
2. Regress $Y_i \sim \tau\hat{T}_i + \gamma^t X$ to obtain an estimate of treatment effect.

A related approach is two-stage residual inclusion. While two-stage least squares uses $\hat{T}_i$ from the first-stage regression to represent the treatment assignments with unobserved confounding "regressed out," two-stage residual inclusion uses the residuals, $T_i - \hat{T}_i$ explicitly as an expression of the unobserved confounding affecting subject i, and uses these residuals as a regressor in the second stage.

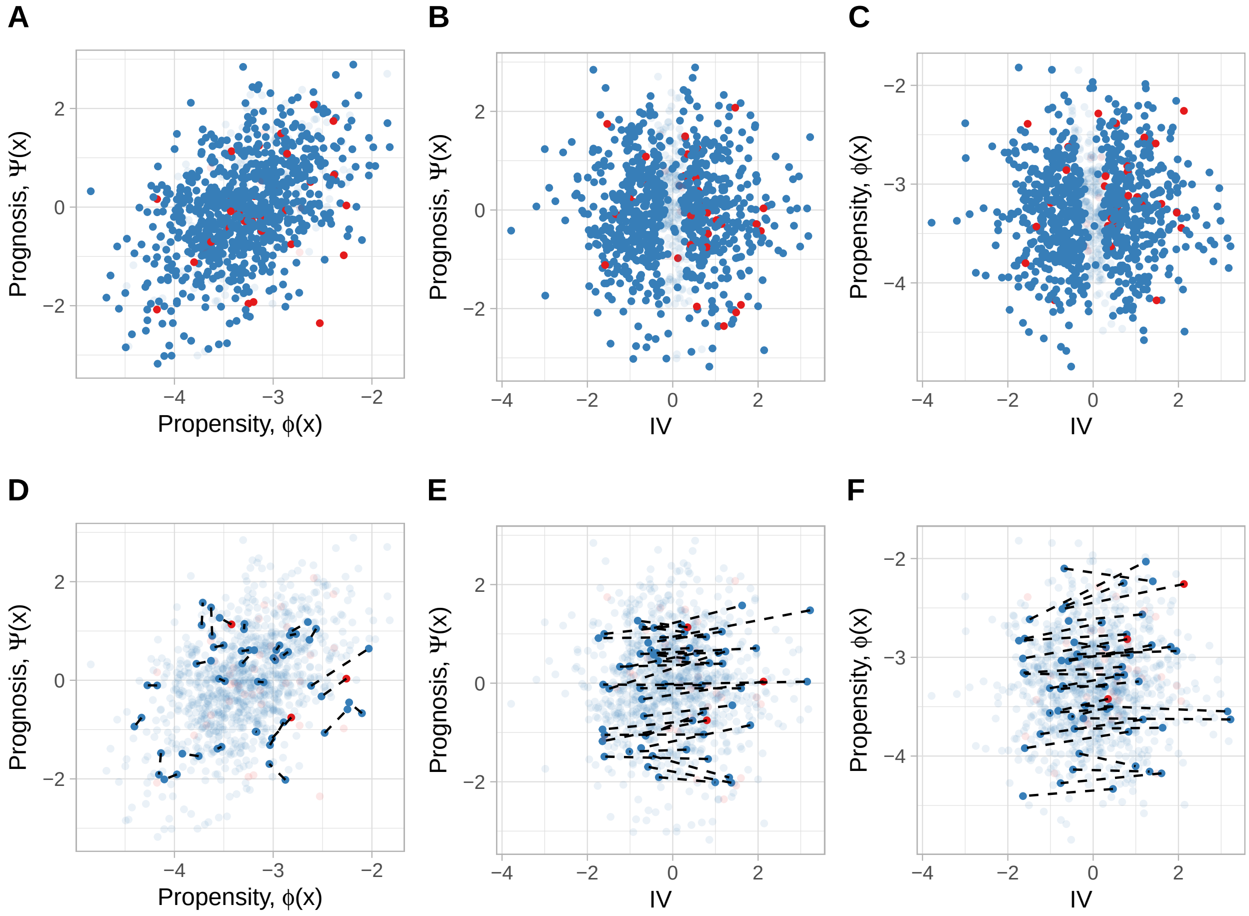
Nearfar Matching

Nearfar matching handles situations like this in a different fashion. Inessense, nearfar matching pairs individuals which are far apart in terms of the measured IV (Z), but near in terms of other important covariates. When we select individuals who have very different levels of Z but very similar levels of X_1, we are essentially trying to select matches which isolate variation in the "pure" IV, ζ.

There are two ways one might imagine visualizing the dataset after nearfar matching. One possibility is to consider the IV axis as Z, the measured IV, since this is the variable that is directly considered in the nearfar matching:



However, the variable of greater import is ζ, the "pure" IV. We would like matched individuals to be distant in terms of ζ in order to reduce bias, but close in terms of propensity and prognosis. Reassuringly, when we visualize the nearfar match in terms of the true IV, ζ, we find that matches are distant in terms of this variable, even though it is not directly observed!


A
B
C
Prognosis, Ψ(x)
Propensity, ϕ(x)
Prognosis, Ψ(x)
IV
Propensity, ϕ(x)
IV
D
E
F
Prognosis, Ψ(x)
Propensity, ϕ(x)
Prognosis, Ψ(x)
IV
Propensity, ϕ(x)
IV