

Matching for Covariate Balance in a Regression Discontinuity Design

Ben Gruenbaum
Harvard

Overview

- RD requires covariate balance across treatment groups
- Matching methods are explicitly designed for this task
- Matching on only one covariate, can be used on close U.S. House elections to correct imbalance on many covariates and improve RD estimates
- Estimating models across a range of matching and estimation bandwidths gives new perspective on data
- Results indicate that imbalance in original data had little effect on incumbency advantage estimates.

Background and Motivation

- RD design for elections requires that outcomes of close elections are randomly distributed.
- By implication covariates must be balanced across treatment groups.
- Elections in U.S. House post-WWII elections are significantly imbalanced and seem (uniquely perhaps) unsuitable for RD analysis.

The Matching Model

- Assume some incumbents can manipulate results if they are “close enough”:
 $E[Y_i|Inc_i=1, v_i < v_0] \neq 0$.
- Then V_i is a function of, latent “true” vote Z_i and ability to manipulate that total, $U_i \in \{0,1\}$:

$$v_i(z_i) = z_i + u_i \Delta, \forall v_i \in [v_0 - \Delta, v_0 + \Delta], \\ D_i(v_i) = 1(v_i > v_0)$$

- Matching on Inc_i implies

$$E[D_i|Inc_i=1] = E[D_i] = E[D_i|Inc_i=0]$$

If $Z_i \perp Inc_i$ we have

$$E[D_i|Inc_i=1, Z_i] = E[D_i|Inc_i=0, Z_i]$$

- Since $E[U_i|Inc_i=0]=0$ implies $E[D_i|z_i > v_0]=1$ and $E[D_i|z_i < v_0]=0$:

$$E[Y_i|z_i > v_0] = E[Y_i|D_i=1], \\ E[Y_i|z_i < v_0] = E[Y_i|D_i=0] \forall z_i \in [v_0 - \Delta, v_0 + \Delta]$$

- Thus causal effects from matched samples can be estimated similarly as in an “ideal” RD.

Interpreting the Matching Estimand

Main assumptions for matching model

- First, manipulation of treatment is a function of observable data features.
- Second, z_i is random in some bandwidth.
- Note that both must also hold in “normal” RD too.

Interpretation and features of model

- An exclusion restriction on Z_i is not required for valid causal effects if we assume $E[D_i|U_i=1]=1$ and $E[U_i|Inc_i=0]=0$. This is a version of a no defiers assumption.
- Estimand from matching-RD is akin to “LATE for compliers”.
- This is the best traditional RD can do also: If $E[U_i] > 0$ then matching recovers an unbiased RD estimate for compliers.

Estimation and Inference

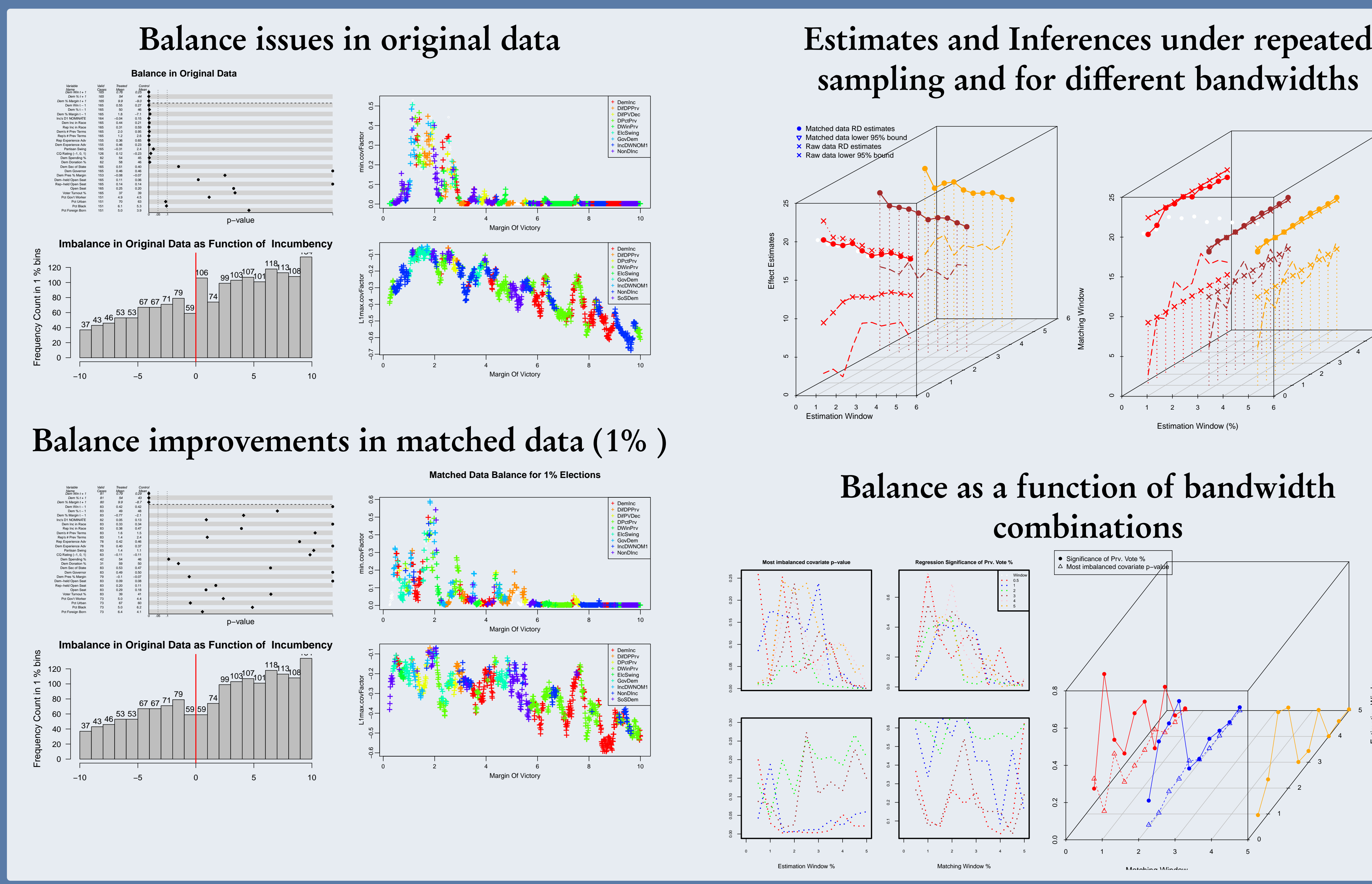
Estimates may be sensitive to three model choices

- Bandwidth for local linear regression: h_r
- Bandwidth within which matching occurs: h_m
- Sampling variance from matching

Procedure for modeling bandwidth sensitivity:

- Specify (h_r, h_m) pair
- Match observations in $[v_0 - h_m, v_0 + h_m]$ to achieve exact balance on Inc_i
- Note proportion of total matches in this bandwidth that are discarded, p_m
- Randomly sample p_m proportion of data **not** in $[v_0 - h_m, v_0 + h_m]$
- Estimate effects using bandwidth h_r (for local linear regression)
- Repeat 100 times per (h_r, h_m) pair.

Matching creates balance, improves inferences for close U.S. House Elections



Balance Measures, Results and Discussion

Assessing Balance

- Balance plots:** Are covariates within a given margin significantly different on either side of that margin?
- Histograms:** Are incumbents more likely to be found just above the cutpoint? (Do incumbents win more close elections than the lose?)
- Balance Trends:** As observations approach cutpoint (from right to left) do covariate observations diverge or converge?
- Conditional Independence Tests:** In regressions of Y_i on Z_i is Z_i significant a significant predictor after matching?
- Balance Frontier:** How do bandwidth specifications effect the balance matching achieves across treatment groups?

Estimate sensitivity to bandwidth specifications:

- How do estimated treatment effects vary as a function of h_r and h_m ?

Conclusion

Graphs illustrate three main findings

- Matching within small windows (0.5 – 5% pictured) significantly effects balance observed near the cutpoint.
- Balance is not a uniform function of either h_r or h_m . Under some circumstances, wider bandwidths appear to actually reduce bias contrary to expectations.
- Overall, there seems to be little reason to worry that estimates from original data are biased.