Application of `lpreba` package

To showcase the practical use of the `lpreba` package, we consider the nonparametric estimation of treatment effects in the regression discontinuity design (RDD). RDD is a popular quasi-experimental research design to identify a causal effect as a jump of an outcome variable at a known threshold of a so-called assignment variable. This threshold constitutes a boundary. When measuring the size of the discontinuity, i.e. the causal effect in a valid RDD, boundary effects may severely bias the estimate and inference. RDD was first applied by Thistlethwaite and Campbell (1960). Hahn et al. (2001) were the first to suggest local linear regression to estimate the conditional expectation function at the threshold to avoid boundary bias. Also see Porter (2003).

In a widely-cited study, Lee (2008) applied RDD to analyze the causal effect of party incumbency on reelection probabilities in U.S. House elections. Figure 1 shows the Democrats' probability of winning a seat in the U.S. House of Representatives in the next election against the Democratic vote share minus the vote share of the strongest opponent (almost always the Republicans). Each data point is an average of the indicator variable for winning election $t+1$ for each interval, which is 0.005 wide. To the left of the zero-threshold the Democratic nominee lost election t; to the right, the Democratic party won. We can see a smooth relationship between the two variables (higher reelection probability for a higher vote share), except for the threshold at zero which decides about victory or defeat. Lee (2008) interprets the large discontinuity as the true electoral incumbency advantage as he argues that due to the inherent uncertainty in the exact vote share, winning a close election is "as good as random". Then, districts where the Democrats just barely lost constitute a valid counterfactual for districts where the Democrats just barely won and became incumbent (receiving treatment).

The task is now to estimate the size of the discontinuity. Figure 2 displays the fit of the Nadaraya-Watson, the local linear and the boundary-adjusted Nadaraya-Watson estimator for each side of the threshold. To select the bandwidths (we allow the degree of smoothing to differ for each side) we use leave-one-out cross-validation (LOOCV). The solution has to be obtained numerically, which can cause problems. The CV error function can have multiple local minima (Figures 3a, 4b), and the solution can be unbounded (Figure 3b). From the Figures 3 and 4 we see that the cross-validated bandwidths are larger for the left side and for the local linear estimator.

The estimates to the left side of the threshold look very similar. However, since local linear regression coincides with the global OLS fit (infinite bandwidth) the estimate is more stable yet captures the same structure. In contrast, on the right side the estimates differ considerably near the boundary. It is apparent that the Nadaraya-Watson estimate suffers from boundary bias. The explicit boundary adjustment results in a much smaller boundary estimate. The curve resembles the parametric logit fit of Lee (2008, Fig. 5a). Local linear regression also leads to a smaller discontinuity estimate, but the correction is less strong. Besides, the estimate is less wavy. The figure illustrates the need for boundary correction. Just looking at the right side to the threshold, Nadaraya-Watson suggests a ten percentage points greater incumbency

advantage compared to adjusted Nadaraya-Watson; and a more than three percentage points greater advantage compared to the local linear estimator. Based on Figure 2 and its simplicity we pick the local linear estimator for further illustration.

We can construct asymptotic confidence intervals to assess precision. Figure 5 displays the local linear estimate along with 95% approximate confidence bands. These confidence intervals account for the variance of the estimator but not its bias, i.e. they are not bias-corrected. We can see that there is more imprecision in the estimator near the threshold.

References

Hahn, J., P. Todd, and W. Van der Klaauw (2001). "Identification and estimation of treatment effects with a regression-discontinuity design". *Econometrica* 69 (1), pp. 201–209.

Lee, D. (2008). "Randomized experiments from non-random selection in U.S. House elections". *Journal of Econometrics* 142 (2), pp. 675–697.

Porter, J. (2003). "Estimation in the regression discontinuity model". `www.ssc.wisc.edu/~jrporter/reg_discont_2003.pdf`. Unpublished.

Thistlethwaite, D. and D. Campbell (1960). "Regression-discontinuity analysis: An alternative to the ex post facto experiment". *Journal of Educational Psychology* 51 (6), pp. 309–317.
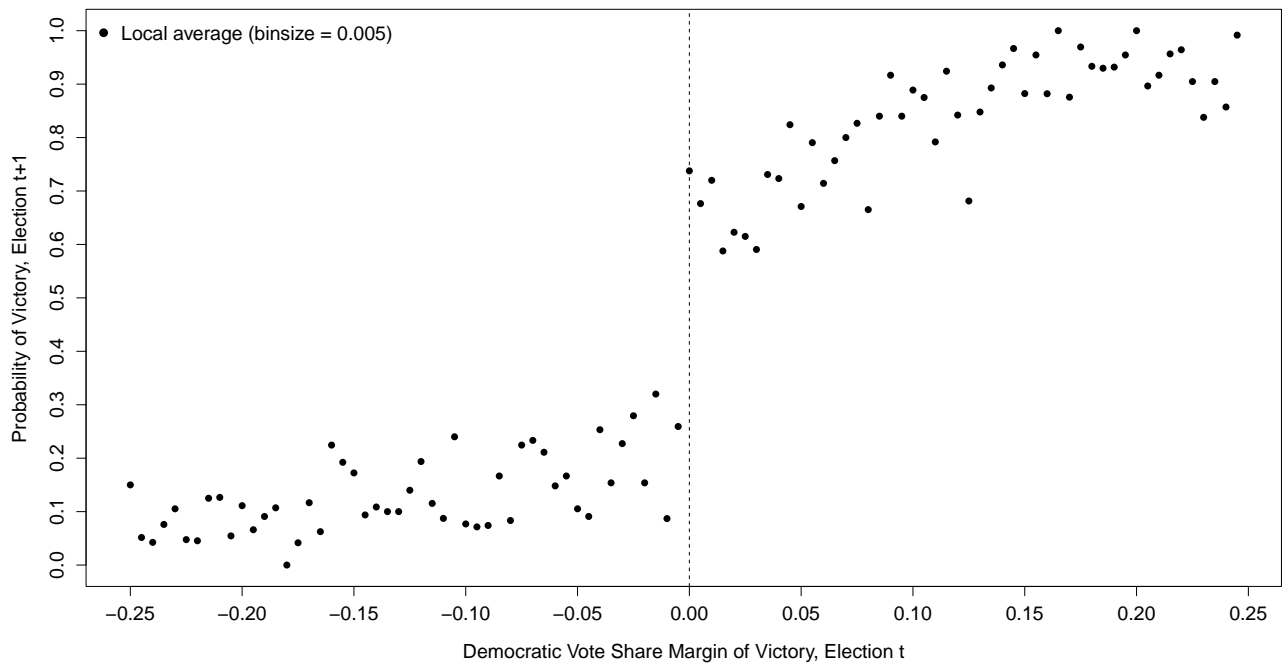
Figure 1: Democrats' probability of victory in election $t+1$, by margin of victory in election t. Local averages are given by the filled circles.
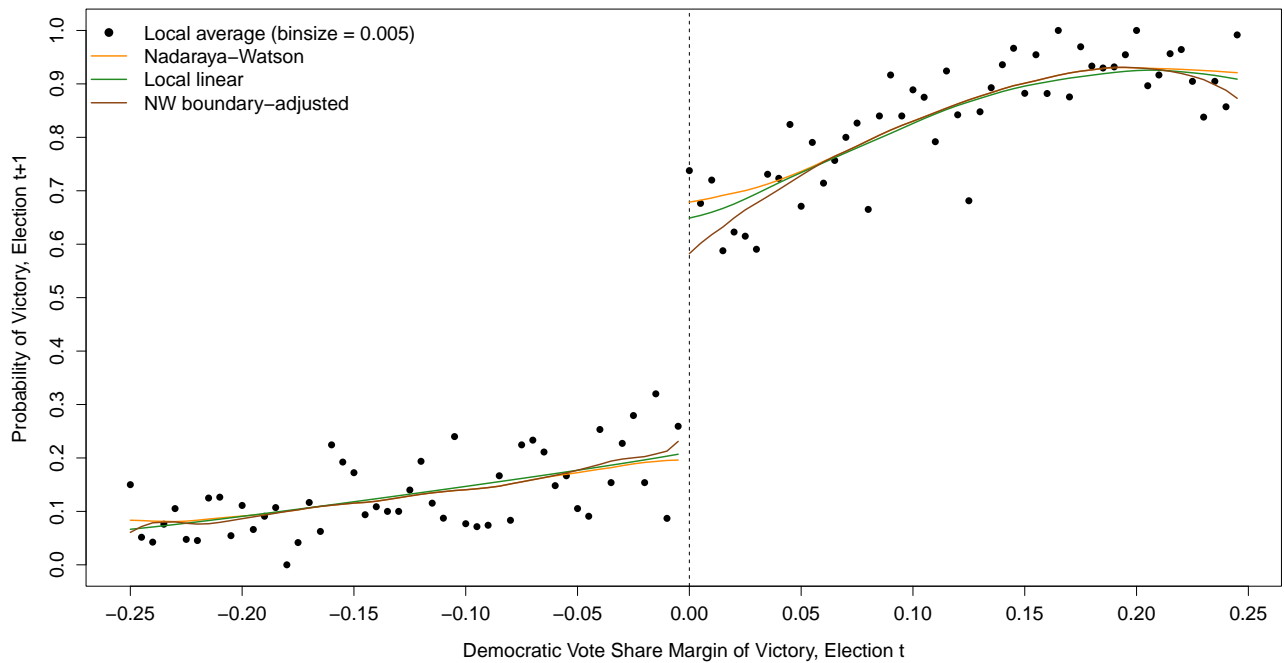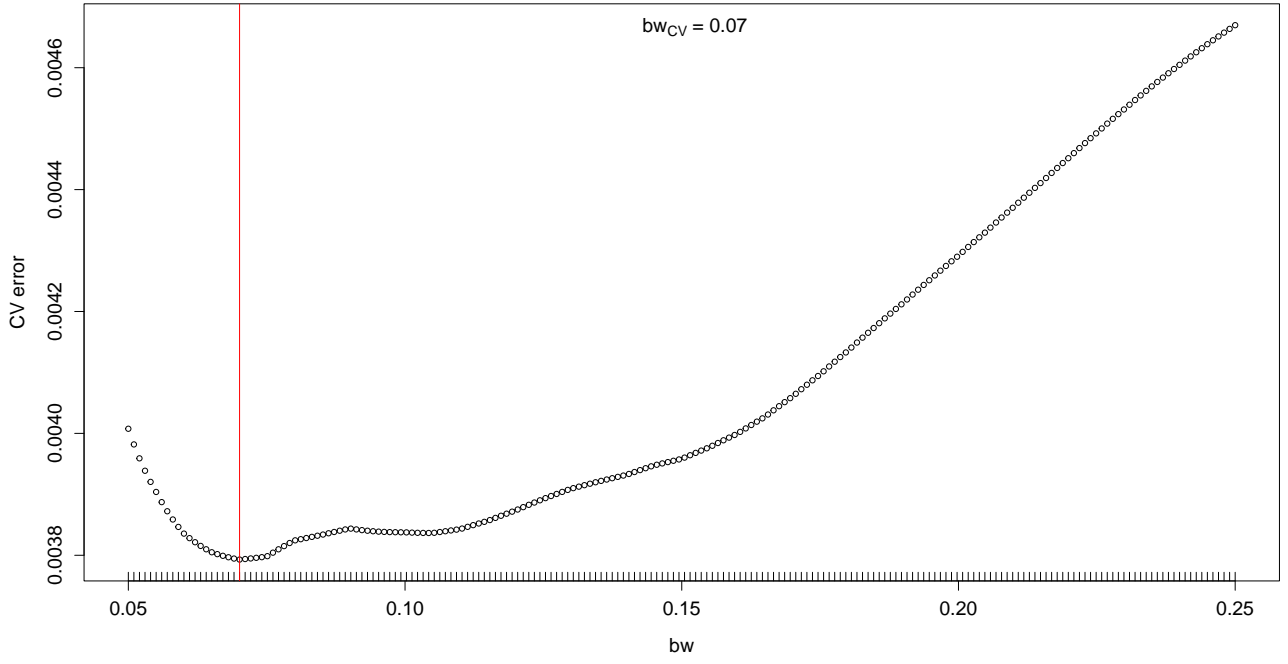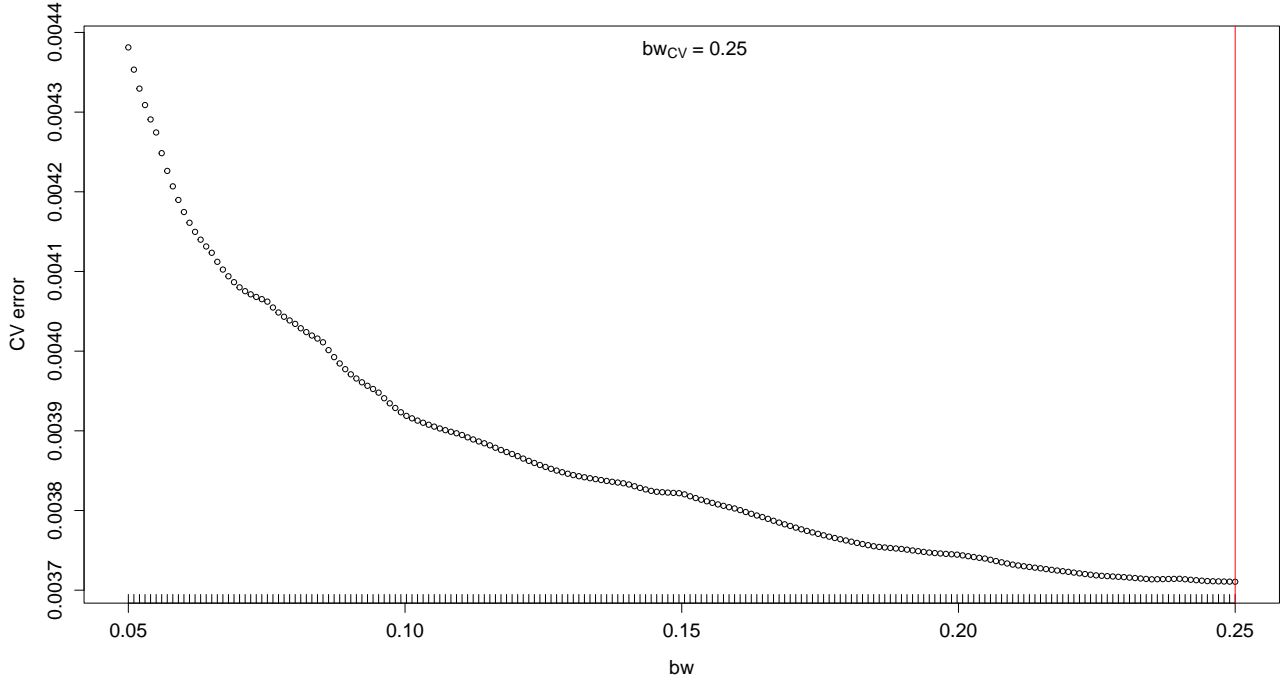

Figure 2: Democrats' probability of victory in election $t+1$, by margin of victory in election t. Estimates of Nadaraya-Watson, local linear regression and adjusted Nadaraya-Watson with cross-validated bandwidths.
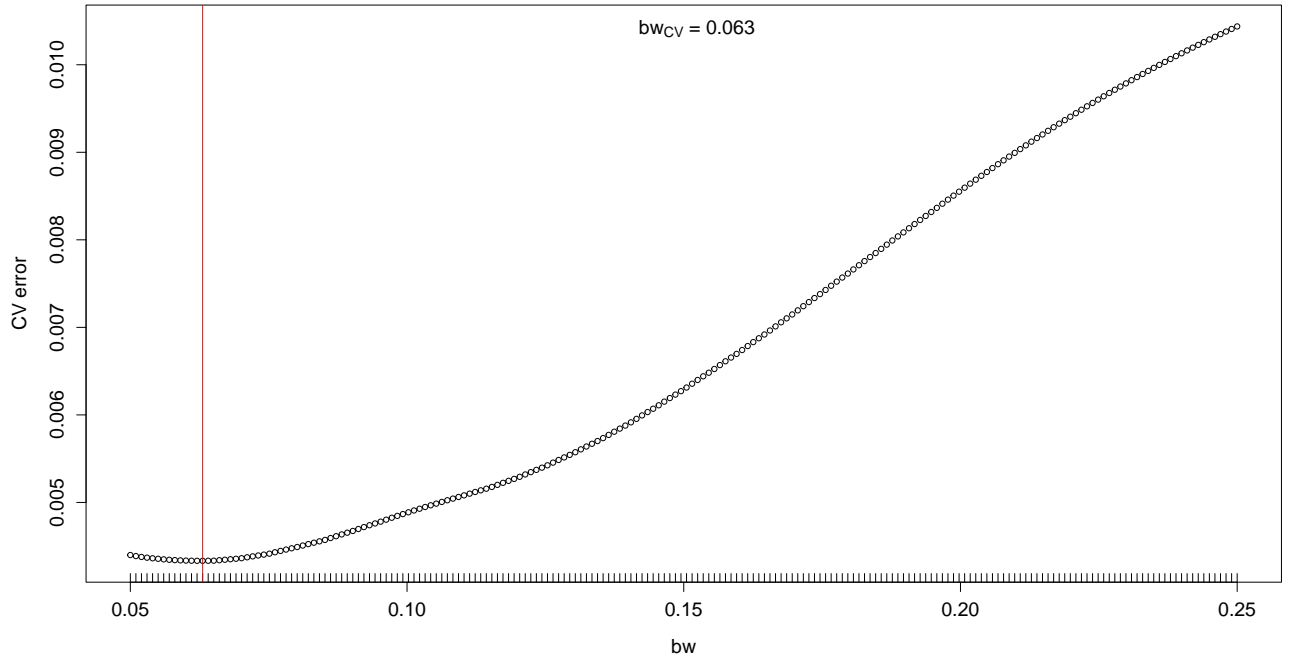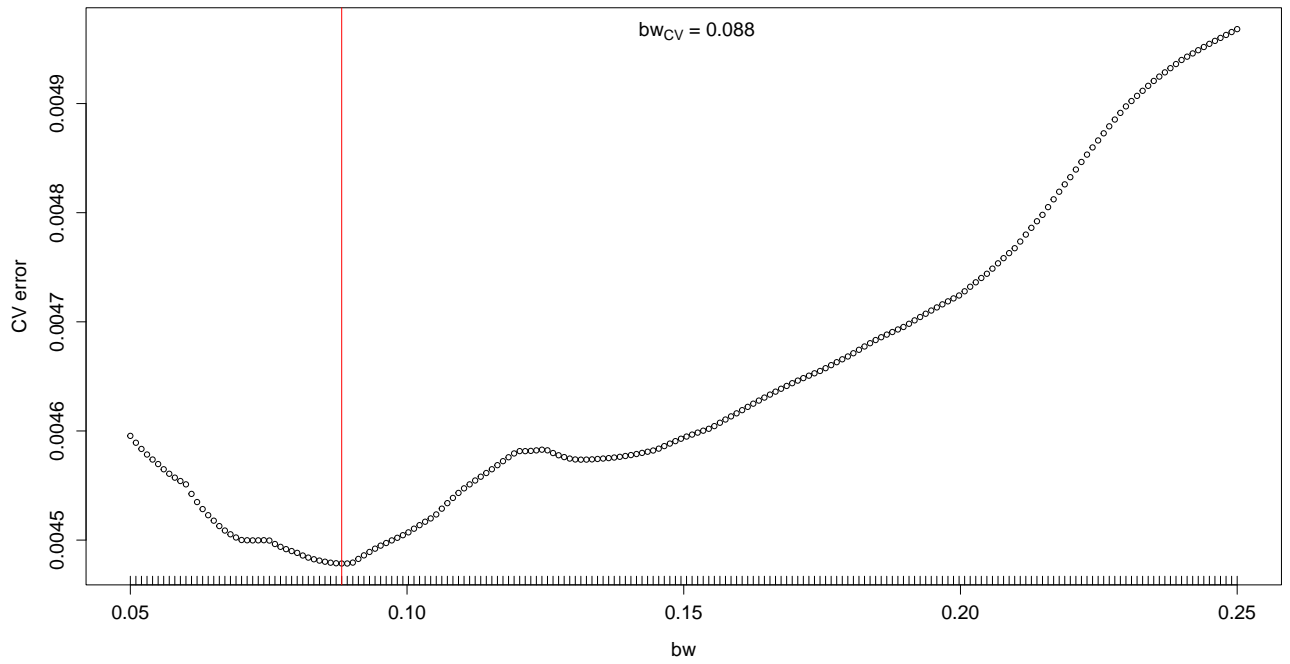
(a) Nadaraya-Watson estimator

(b) Local linear estimator

Figure 3: CV error over a grid of 200 bandwidths for the left side to the threshold in Figure 2. The optimal bandwidth is indicated by the red line.

(a) Nadaraya-Watson estimator

(b) Local linear estimator

Figure 4: CV error over a grid of 200 bandwidths for the right side to the threshold in Figure 2. The optimal bandwidth is indicated by the red line.
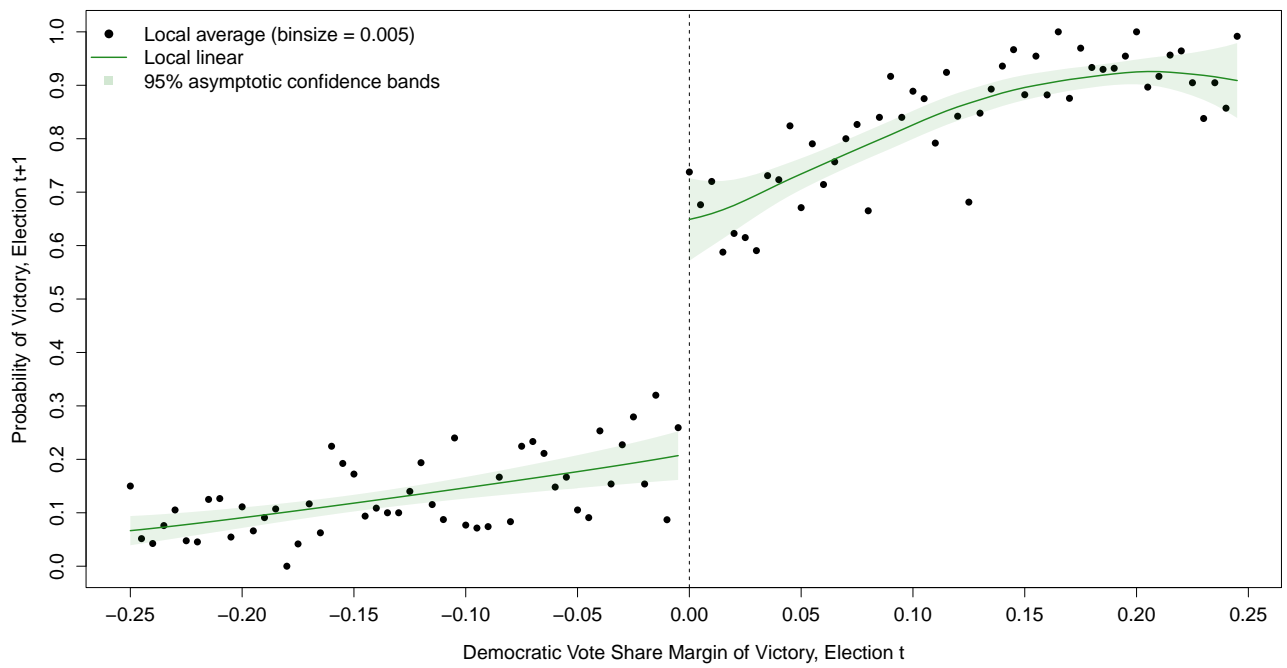
Figure 5: Democrats' probability of victory in election $t+1$, by margin of victory in election t. Local linear estimate along with 95% approximate confidence bands.