ATbounds-vignette-NSW

```
library(ATbounds)
```

To illustrate the usefulness of the package, we first use the well-known LaLonde dataset available at Rajeev Dehejia's Data Page.

LaLonde's Experimental Sample

We fist look at LaLonde's original experimental sample.

```
  nsw_treated <- read.table("http://users.nber.org/~rdehejia/data/nsw_treated.txt")
  colnames(nsw_treated) <- c("treat","age","edu","black","hispanic",
                           "married","nodegree","RE75","RE78")

  nsw_control <- read.table("http://users.nber.org/~rdehejia/data/nsw_control.txt")
  colnames(nsw_control) <- c("treat","age","edu","black","hispanic",
                           "married","nodegree","RE75","RE78")
```

The outcome variable is `RE78` (earnings in 1978). The binary treatment indicator is `treat` (1 if treated, 0 if not treated). We now combine the treatment and control samples and define the variables.

```
  nsw <- rbind(nsw_treated,nsw_control)
  attach(nsw)
  D <- treat
  Y <- (RE78 > 0)
```

In this vignette, we define the outcome to be whether employed in 1978 (that is, earnings in 1978 are positive).

The LaLonde dataset is from the National Supported Work Demonstration (NSW), which is a randomized controlled temporary employment program. In view of that, we set the reference propensity score to be independent of covariates.

```
  rps <- rep(mean(D),length(D))
```

The average treatment effect is obtained by

```
  ate_nsw <- mean(D*Y)/mean(D)-mean((1-D)*Y)/mean(1-D)
  print(ate_nsw)
#> [1] 0.07794019
```

Dehejia-Wahba Sample

Dehejia and Wahba (1999, 2002) extract a further subset of Lalonde's NSW experimental data to obtain a subset containing information on RE74 (earnings in 1974).

```
  detach(nsw)
  nswre_treated <- read.table("http://users.nber.org/~rdehejia/data/nswre74_treated.txt")
  colnames(nswre_treated) <- c("treat","age","edu","black","hispanic",
                           "married","nodegree","RE74","RE75","RE78")

  nswre_control <- read.table("http://users.nber.org/~rdehejia/data/nswre74_control.txt")
  colnames(nswre_control) <- c("treat","age","edu","black","hispanic",
                               "married","nodegree","RE74","RE75","RE78")
  nswre <- rbind(nswre_treated,nswre_control)
  attach(nswre)
  D <- treat
  Y <- (RE78 > 0)
  X <- cbind(age,edu,black,hispanic,married,nodegree,RE74/1000,RE75/1000)
```

The covariates are as follows:

- `age`: age in years,
- `edu`: years of education,
- `black`: 1 if black, 0 otherwise,
- `hispanic`: 1 if Hispanic, 0 otherwise,
- `married`: 1 if married, 0 otherwise,
- `nodegree`: 1 if no degree, 0 otherwise,
- `RE74`: earnings in 1974,
- `RE75`: earnings in 1975.

If we assume that the Dehejia-Wahba sample still preserves initial randomization, we can set the reference propensity score to be independent of covariates. However, it may not be the case and therefore, our approach can provide a robust method to check whether the Dehejia-Wahba sample can be viewed as a randomized controlled experiment.

We first define the the reference propensity score.

```
  rps <- rep(mean(D),length(D))
```

Using this reference propensity score, the average treatment effect is obtained by

```
  ate_nswre <- mean(D*Y)/mean(D)-mean((1-D)*Y)/mean(1-D)
  print(ate_nswre)
#> [1] 0.1106029
```

We now introduce our bounds.

```
  bns_nsw <- atebounds(Y, D, X, rps, Q = 2, n_permute = 500)
```

In implementing `atebounds`, the default choice of the polynomial order `q` is $Q = 2$, which uses the nearest neighbor excluding own observations. The k-nearest neighbor estimator is used when $Q = k + 1$. Here, the nearest neighbor for observation i is always its own observation i. We use the R package FNN to carry out k-nearest neighbor search. When there are ties, the ordering of the observations may matter. To avoid this issue, we permute the dataset `n_permute` number of times and take the average of the estimates of the bounds. The default choice of `n_permute = 0`, under which there is no permutation.

There are two more options which are not used above:

- `discrete`: `TRUE` if `X` includes only discrete covariates and `FALSE` if not (default: `FALSE`)
- `studentize`: `TRUE` if `X` is studentized elementwise and `FALSE` if not (default: `TRUE`)

We print the outputs saved in `bns_nsw`.

```
  print(bns_nsw)
#> $y1_lb
#> [1] 0.7361525
#>
#> $y1_ub
#> [1] 0.7468799
#>
```

```
#> $y0_lb
#> [1] 0.6463793
#>
#> $y0_ub
#> [1] 0.6746593
#>
#> $ate_lb
#> [1] 0.06149327
#>
#> $ate_ub
#> [1] 0.1005007
#>
#> $ate_rps
#> [1] 0.1106029
#>
#> attr(,"class")
#> [1] "ATbounds"
```

The output list contains:

- `y1_lb`: the lower bound of the average of $Y(1)$, i.e. $\mathbb{E}[Y(1)]$,
- `y1_ub`: the upper bound of the average of $Y(1)$, i.e. $\mathbb{E}[Y(1)]$,
- `y0_lb`: the lower bound of the average of $Y(0)$, i.e. $\mathbb{E}[Y(0)]$,
- `y0_ub`: the upper bound of the average of $Y(0)$, i.e. $\mathbb{E}[Y(0)]$,
- `ate_lb`: the lower bound of ATE, i.e. $\mathbb{E}[Y(1) - Y(0)]$,
- `ate_ub`: the upper bound of ATE, i.e. $\mathbb{E}[Y(1) - Y(0)]$,
- `ate_rps`: the point estimate of ATE using the reference propensity score

With $Q = 2$, the bounds for ATE is

```
print(c(bns_nsw$ate_lb,bns_nsw$ate_ub))
#> [1] 0.06149327 0.10050066
```

Recall that the point ATE estimate was

```
print(ate_nswre)
#> [1] 0.1106029
```

Thus, the upper bound is slightly smaller than the point estimate using the reference propensity score. However, their difference may not be significant given the small sample size.

We now look at the case with $Q = 3$.

```
  bns_nsw <- atebounds(Y, D, X, rps, Q = 3, n_permute = 500)
  print(c(bns_nsw$ate_lb,bns_nsw$ate_ub))
#> [1] 0.1334685 0.1106208
```

With $Q = 3$, the bounds cross. This indicates that the ATE estimate may be a reasonable estimate. We need to develop an inference method to reach a more definite answer.

We now look at ATT.

```
  bns_nsw_att <- attbounds(Y, D, X, rps, Q = 2, n_permute = 500)
  print(bns_nsw_att)
#> $lb
#> [1] 0.04825963
#>
#> $ub
#> [1] 0.09150562
#>
#> $att_rps
#> [1] 0.1106029
#>
#> attr(,"class")
#> [1] "ATbounds"
```

We also set $Q = 3$.

```
  bns_nsw_att <- attbounds(Y, D, X, rps, Q = 3, n_permute = 500)
  print(bns_nsw_att)
#> $lb
#> [1] 0.09228541
#>
#> $ub
#> [1] 0.136572
#>
#> $att_rps
#> [1] 0.1106029
#>
#> attr(,"class")
#> [1] "ATbounds"
```

Again the results are a bit sensitive to the choice of q and it is likely to be deriven by a relative small sample size.

NSW treated plus PSID control

```
  attach(nswre)
#> The following objects are masked from nswre (pos = 3):
#>
#>     age, black, edu, hispanic, married, nodegree, RE74, RE75, RE78, treat
  psid2_control <- read.table("http://users.nber.org/~rdehejia/data/psid2_controls.txt")
  colnames(psid2_control) <- c("treat","age","edu","black","hispanic",
                               "married","nodegree","RE74","RE75","RE78")
  psid <- rbind(nswre_treated,psid2_control)
  attach(psid)
#> The following objects are masked from nswre (pos = 3):
#>
#>     age, black, edu, hispanic, married, nodegree, RE74, RE75, RE78, treat
#> The following objects are masked from nswre (pos = 4):
#>
#>     age, black, edu, hispanic, married, nodegree, RE74, RE75, RE78, treat
  D <- treat
  Y <- (RE78 > 0)
  X <- cbind(age,edu,black,hispanic,married,nodegree,RE74/1000,RE75/1000)
```

Here, we use one of non-experimental comparison groups constructed by LaLonde from the Population Survey of Income Dynamics, namely PSID2 controls. We now estimate the reference propensity score using a logit model and obtain the new bound estimates:

```
  lm <- stats::glm(D~X, family=binomial("logit"))
  rps <- lm$fitted.values
  bns_psid <- atebounds(Y, D, X, rps, Q = 2, n_permute = 500)
  print(bns_psid)
#> $y1_lb
#> [1] -0.5538202
#>
#> $y1_ub
#> [1] 0.8304792
#>
#> $y0_lb
#> [1] 0.4417992
#>
#> $y0_ub
#> [1] 0.8293488
#>
#> $ate_lb
#> [1] -1.383169
#>
#> $ate_ub
#> [1] 0.3886801
#>
#> $ate_rps
#> [1] 0.2715819
#>
#> attr(,"class")
#> [1] "ATbounds"
```

```
  lm <- stats::glm(D~X, family=binomial("logit"))
  rps <- lm$fitted.values
  bns_psid <- atebounds(Y, D, X, rps, Q = 3, n_permute = 500)
  print(bns_psid)
#> $y1_lb
#> [1] 0.2742846
#>
#> $y1_ub
#> [1] 0.808709
#>
#> $y0_lb
#> [1] 0.4753347
#>
#> $y0_ub
#> [1] 0.9684722
#>
#> $ate_lb
#> [1] -0.6941876
#>
#> $ate_ub
#> [1] 0.3333743
#>
#> $ate_rps
#> [1] 0.2715819
#>
#> attr(,"class")
#> [1] "ATbounds"
```

The bounds are wide and includes zero. We compare these with the Manski bounds, which can be obtained by setting $Q = 1$.

```
  bns_psid <- atebounds(Y, D, X, rps, Q = 1, n_permute = 500)
  print(bns_psid)
#> $y1_lb
#> [1] 0.3196347
#>
#> $y1_ub
#> [1] 0.8972603
#>
#> $y0_lb
#> [1] 0.3789954
#>
#> $y0_ub
#> [1] 0.8013699
#>
#> $ate_lb
#> [1] -0.4817352
#>
#> $ate_ub
#> [1] 0.5182648
#>
#> $ate_rps
#> [1] 0.2715819
#>
#> attr(,"class")
#> [1] "ATbounds"
```

We can see that the Manski bounds are qualitatively comparable to those with $Q = 2, 3$. This is interesting because the Manski bounds do not impose the unconfoundedness assumption.

Finally, we obtain the bounds for the average treatment effect on the treated (ATT).

```
  bns_psid_att <- attbounds(Y, D, X, rps, Q = 2, n_permute = 500)
  print(bns_psid_att)
#> $lb
#> [1] -0.3992898
#>
#> $ub
#> [1] 0.6032461
#>
#> $att_rps
#> [1] 0.2031435
#>
#> attr(,"class")
#> [1] "ATbounds"
```

```
  bns_psid_att <- attbounds(Y, D, X, rps, Q = 3, n_permute = 500)
  print(bns_psid_att)
#> $lb
#> [1] -0.6366576
#>
#> $ub
#> [1] 0.5265124
#>
#> $att_rps
#> [1] 0.2031435
#>
#> attr(,"class")
#> [1] "ATbounds"
  detach(psid)
```

Again, we find that the bounds are wide; they seem similar to those for the ATE. Overall, the empirical results suggest that PSID2 controls are substantially different from the treated units such that the unconfoundedness assumption does not provide any meaningful restriction once the overlap condition is dropped and the logit prepensity score can be misspecified.