ATbounds-vignette

```
library(ATbounds)
```

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

Case 1: The Reference Propensity Score is Correctly Specified

We start with the case that the reference propensity score is correctly specified.

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

  nsw_control <- read.table("http://users.nber.org/~rdehejia/data/nswre74_control.txt")
  colnames(nsw_control) <- c("treat","age","edu","black","hispanic",
                           "married","nodegree","RE74","RE75","RE78")
```

The outcome variable is `RE78` (earnings in 1978). The binary treatment indicator is `treat` (1 if treated, 0 if not treated). 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.

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)
  X <- cbind(age,edu,black,hispanic,married,nodegree,RE74/1000,RE75/1000)
```

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] 0.1106029
```

We now introduce our bounds.

```
  bns_nsw <- atebounds(Y, D, X, rps, q = 3, permute_max = 100)
```

In implementing `atebounds`, the polynomial order `q` is set at $q = 3$. The default choice 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 `permute_max` number of times and take the average of the estimates of the bounds. The default choice of `permute_max = 0`, under which there is no permutation.

There are two more options which are not used above:

- `discrete`: `TRUE` if `X` inclues 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] 0.7567568
#> 
#> $y1_ub
#> [1] 0.7567568
#> 
#> $y0_lb
#> [1] 0.6349873
#> 
#> $y0_ub
#> [1] 0.6530452
#> 
#> $ate_lb
#> [1] 0.1037116
#> 
#> $ate_ub
#> [1] 0.1217695
#> 
#> $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

Notice that `bns_nsw$y1_lb` is equal to `bns_nsw$y1_ub` above. This is because it turns out that the estimated upper bound was smaller than the estimated lower bound, which indicates that $\mathbb{E}[Y(1)]$ may be point-identified, that is, the assumed reference propensity score is correctly specified. If this happens, we assign the following for both the upper and lower bounds:

$$\widehat{\mathbb{E}[Y(1)]} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i}{p_*(X_i)},$$

where $p_*(\cdot)$ is the reference propensity score. On the other hand, `bns_nsw$y0_lb` is indeed smaller than `bns_nsw$y0_ub`. Finally, we obtain the lower bound for ATE by `ate_lb = y1_lb - y0_ub` and the upper bound by `ate_ub = y1_ub - y0_lb`.

We find that the estimated bounds for ATE are tight and includes `ate_rps`, as expected since the assumed reference propensity score is correctly specified.

Case 2: The Reference Propensity Score is misspecified

We now move to the case that the reference propensity score is misspecified.

```
  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(nsw_treated,psid2_control)
  attach(psid)
#> The following objects are masked from nsw:
#> 
#>     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 = 3, permute_max = 100)
  print(bns_psid)
#> $y1_lb
#> [1] 0.2749302
#> 
#> $y1_ub
#> [1] 0.8085899
#> 
#> $y0_lb
#> [1] 0.4816034
#> 
#> $y0_ub
#> [1] 0.9670768
#> 
#> $ate_lb
#> [1] -0.6921466
#> 
#> $ate_ub
#> [1] 0.3269865
#> 
#> $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, permute_max = 100)
  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 wider than those with $q = 1$. This is mainly 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 = 3, permute_max = 100)
  print(bns_psid_att)
#> $lb
#> [1] -0.6286155
#>
#> $ub
#> [1] 0.532657
#>
#> $att_rps
#> [1] 0.2031435
#>
#> attr(,"class")
#> [1] "ATbounds"
```

Again, we find that the bounds are wide; they seem similar to those for the ATE.