ATbounds: An R vignette using data from Dehijia's web page

```
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
```

Bounds on ATE

We now introduce our bounds on the average treatment effect (ATE).

```
bns_nsw <- atebounds(Y, D, X, rps)
```

In implementing `atebounds`, there are several options:

- `Q`: bandwidth parameter that determines the maximum number of observations for pooling information (default: $Q = 3$)
- `studentize`: `TRUE` if `X` is studentized elementwise and `FALSE` if not (default: `TRUE`)
- `alpha`: $(1 - \alpha)$ nominal coverage probability for the confidence interval of ATE (default: 0.05)
- `discrete`: `TRUE` if `X` includes only discrete covariates and `FALSE` if not (default: `FALSE`)
- `n_hc`: number of hierarchical clusters to discretize non-discrete covariates; relevant only if x_discrete is FALSE. The default choice is `n_hc = ceiling(length(Y)/10)`, so that there are 10 observations in each cluster on average.

We show the summary results saved in `bns_nsw`.

```
  summary(bns_nsw)
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps)
#> Confidence Level: 0.95
#>                       Lower_Bound Upper_Bound
#> Bound Estimate          0.105289    0.077649
#> Confidence Interval    -0.010448    0.194602
```

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

```
summary(atebounds(Y, D, X, rps, Q = 2))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 2)
#> Confidence Level: 0.95
#>                       Lower_Bound Upper_Bound
#> Bound Estimate          0.083059     0.10125
#> Confidence Interval    -0.014176     0.21192
```

With $Q = 4$, the bounds for ATE is

```
summary(atebounds(Y, D, X, rps, Q = 4))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 4)
#> Confidence Level: 0.95
#>                       Lower_Bound Upper_Bound
#> Bound Estimate          0.132590    0.041889
#> Confidence Interval    -0.032998    0.202929
```

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 is in the 95% confidence interval for $Q = 2, 3, 4$.

To see what is saved in `bns_nsw`, we now print it out.

```
  print(bns_nsw)
#> $call
#> atebounds(Y = Y, D = D, X = X, rps = rps)
#>
#> $type
#> [1] "ATE"
#>
#> $cov_prob
#> [1] 0.95
#>
#> $y1_lb
#> [1] 0.7214833
#>
#> $y1_ub
#> [1] 0.7145048
#>
#> $y0_lb
#> [1] 0.6368557
#>
#> $y0_ub
#> [1] 0.6161944
#>
#> $est_lb
#> [1] 0.105289
#>
#> $est_ub
#> [1] 0.07764911
#>
#> $est_rps
#> [1] 0.1106029
#>
#> $se_lb
#> [1] 0.05010572
#>
#> $se_ub
#> [1] 0.05471636
#>
#> $ci_lb
#> [1] -0.0104481
#>
#> $ci_ub
#> [1] 0.1946019
#>
#> attr(,"class")
#> [1] "ATbounds"
```

The output list contains:

- `call`: a call in which all of the specified arguments are specified by their full names
- `type`: ATE
- `cov_prob`: confidence level (1-alpha)
- `y1_lb`: estimate of the lower bound on the average of $Y(1)$, i.e. $\mathbb{E}[Y(1)]$,
- `y1_ub`: estimate of the upper bound on the average of $Y(1)$, i.e. $\mathbb{E}[Y(1)]$,

- `y0_lb`: estimate of the lower bound on the average of $Y(0)$, i.e. $\mathbb{E}[Y(0)]$,
- `y0_ub`: estimate of the upper bound on the average of $Y(0)$, i.e. $\mathbb{E}[Y(0)]$,
- `est_lb`: estimate of the lower bound on ATE, i.e. $\mathbb{E}[Y(1) - Y(0)]$,
- `est_ub`: estimate of the upper bound on ATE, i.e. $\mathbb{E}[Y(1) - Y(0)]$,
- `est_rps`: the point estimate of ATE using the reference propensity score
- `se_lb`: standard error for the estimate on the lower bound on ATE
- `se_ub`: standard error for the estimate of the upper bound on ATE
- `ci_lb`: the lower end point of the confidence interval for ATE
- `ci_ub`: the upper end point of the confidence interval for ATE

Bounds on ATT

We now look at bounds on the average treatment effect on the treated (ATT).

```
  bns_nsw_att <- attbounds(Y, D, X, rps)
  summary(bns_nsw_att)
#> ATbounds: ATT
#> Call: attbounds(Y = Y, D = D, X = X, rps = rps)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate          0.146992    0.097293
#> Confidence Interval    -0.010469    0.256875
```

We also experiment with $Q = 2$ and $Q = 4$.

```
  summary(attbounds(Y, D, X, rps, Q = 2))
#> ATbounds: ATT
#> Call: attbounds(Y = Y, D = D, X = X, rps = rps, Q = 2)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate          0.133659     0.12899
#> Confidence Interval     0.000365     0.27285
```

```
  summary(attbounds(Y, D, X, rps, Q = 4))
#> ATbounds: ATT
#> Call: attbounds(Y = Y, D = D, X = X, rps = rps, Q = 4)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate          0.181554    0.061652
#> Confidence Interval    -0.021856    0.263347
```

Again the results are more or less similar in terms of confidence intervals. Bound estimates are sensitive to the choice of Q and it is likely to be driven by a relative small sample size.

NSW treated plus PSID control

```
  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)
```

```
  detach(nswre)
  attach(psid)
  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 <- lm$fitted.values
  bns_psid <- atebounds(Y, D, X, rps_lm)
  summary(bns_psid)
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps_lm)
#> Confidence Level: 0.95
#>                     Lower_Bound Upper_Bound
#> Bound Estimate          -1.6823     -39.791
#> Confidence Interval     -9.7544       6.051
```

We compare the results using the logit reference propensity score with those using the sample proportion as the reference propensity score.

```
  rps_sp <- rep(mean(D),length(D))
  bns_psid <- atebounds(Y, D, X, rps_sp)
  summary(bns_psid)
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps_sp)
#> Confidence Level: 0.95
#>                     Lower_Bound Upper_Bound
#> Bound Estimate          -0.17799    0.076754
#> Confidence Interval     -0.34628    0.310176
```

Both results are similar. We compare these with the Manski bounds, which can be obtained by setting $Q = 1$.

```
  summary(atebounds(Y, D, X, rps_lm, Q=1))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps_lm, Q = 1)
#> Confidence Level: 0.95
#>                     Lower_Bound Upper_Bound
#> Bound Estimate          -0.47945     0.46347
#> Confidence Interval     -0.66847     0.77042
  summary(atebounds(Y, D, X, rps_sp, Q=1))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps_sp, Q = 1)
#> Confidence Level: 0.95
#>                     Lower_Bound Upper_Bound
#> Bound Estimate          -0.47945     0.46347
#> Confidence Interval     -0.66847     0.77042
```

We can see that the Manski bounds are much larger and the same regardless of the specification of the reference propensity scores. Recall the Manski bounds do not impose the unconfoundedness assumption and do not rely on any pooling information (hence, it does not matter how to specifiy the reference propensity score).

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

```
  summary(atebounds(Y, D, X, rps_lm))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps_lm)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate           -1.6823     -39.791
#> Confidence Interval      -9.7544       6.051
  summary(atebounds(Y, D, X, rps_sp))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps_sp)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate           -0.17799    0.076754
#> Confidence Interval      -0.34628    0.310176
  detach(psid)
```

We find that the bounds on ATT are similar to those on ATE.