ATbounds: An R Vignette

Sokbae Lee and Martin Weidner

Abstract

ATbounds is an R package that provides estimation and inference methods for bounding average treatment effects (on the treated) that are valid under an unconfoundedness assumption. The bounds are designed to be robust in challenging situations, for example, when the the conditioning variables take on a large number of different values in the observed sample, or when the overlap condition is violated. This robustness is achieved by only using limited "pooling" of information across observations.

Introduction

ATbounds is an R package that provides estimation and inference methods for bounding average treatment effects (on the treated) that are valid under an unconfoundedness assumption. The bounds are designed to be robust in challenging situations, for example, when the the conditioning variables take on a large number of different values in the observed sample, or when the overlap condition is violated. This robustness is achieved by only using limited "pooling" of information across observations. Namely, the bounds are constructed as sample averages over functions of the observed outcomes such that the contribution of each outcome only depends on the treatment status of a limited number of observations. No information pooling across observations leads to so-called "Manski bounds" (Manski 1989, 1990), while unlimited information pooling leads to standard inverse propensity score weighting. The ATbounds package provides inference methods for exploring the intermediate range between these two extremes.

The methodology used in the ATbounds package is described in detail in Lee and Weidner (2021), "Bounding Treatment Effects by Pooling Limited Information across Observations," "Bounding Treatment Effects by Pooling Limited Information across Observations," available at https://arxiv.org/abs/2111.05243.

We begin by calling the ATbounds package.

```
library(ATbounds)
```

Case Study 1: Bounding the Effects of a Job Training Program

To illustrate the usefulness of the package, we first use the well-known LaLonde (1986) dataset available on Rajeev Dehejia's web page at http://users.nber.org/~rdehejia/nswdata2.html.

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
```

Alternatively, we run simple regression

```
  model <- lm(Y ~ D)
  summary(model)
#>
#> Call:
#> lm(formula = Y ~ D)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -0.7744 -0.6965  0.2256  0.3035  0.3035
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.69647    0.02152  32.362   <2e-16 ***
#> D            0.07794    0.03356   2.323   0.0205 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.4437 on 720 degrees of freedom
#> Multiple R-squared:  0.007437,   Adjusted R-squared:  0.006059
#> F-statistic: 5.395 on 1 and 720 DF,  p-value: 0.02047
  confint(model)
#>                 2.5 %    97.5 %
#> (Intercept) 0.6542184 0.7387228
#> D           0.0120623 0.1438181
```

The 95% confidence interval $[0.01, 0.14]$ is rather wide but excludes zero.

Dehejia-Wahba Sample

Dehejia and Wahba (1999) and Dehejia and Wahba (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 as a random sample from 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
```

Alternatively, we run simple regression

```
  model <- lm(Y ~ D)
  summary(model)
#>
#> Call:
#> lm(formula = Y ~ D)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -0.7568 -0.6462  0.2432  0.3538  0.3538
#>
#> Coefficients:
```

```
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.64615    0.02849  22.679   <2e-16 ***
#> D            0.11060    0.04419   2.503   0.0127 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.4594 on 443 degrees of freedom
#> Multiple R-squared:  0.01394,    Adjusted R-squared:  0.01172
#> F-statistic: 6.265 on 1 and 443 DF,  p-value: 0.01267
```

```
  confint(model)
#>                  2.5 %    97.5 %
#> (Intercept) 0.59015824 0.7021495
#> D           0.02375725 0.1974486
```

The resulting 95% confidence interval $[0.02, 0.20]$ is wide but again excludes zero.

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.

The clusters are constructed via hierarchical, agglomerative clustering with complete linkage, where the distance is measured by the Euclidean distance after studentizing each of the covariates. As mentioned above, the number m of clusters is set by

$$m = \left\lceil \frac{n}{L} \right\rceil$$

for $L = 10$.

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
```

Note that the estimate of the lower bound is larger than that of the upper bound. This crossing problem can occur in finite samples due to random sampling errors. While statistical theory guarantees that this problem cannot occur asymptotically, it is desirable to have a non-empty confidence interval in applications. We therefore use the method in Stoye (2020) to obtain a valid confidence interval that is never empty.

The 95% confidence interval $[-0.01, 0.19]$ obtained here is similar to the previous interval $[0.02, 0.20]$, which was obtained under the assumption that the Dehejia-Wahba sample is a random sample from NSW. This suggests that first, there is no evidence against violation of the random sampling assumption in the Dehejia-Wahba sample and second, our inference method does not suffer from unduly enlargement of the confidence interval to achieve robustness, although a null effect is now included in our confidence interval.

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

```
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 are

```
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 estimate of ATE under the random sampling assumption was

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

Thus, the ATE estimate based on the simple mean difference well within the 95% confidence intervals across $Q = 2, 3, 4$.

Recall that covariates X include non-discrete variables: `RE74` and `RE75`. As a default option, `atebounds` chooses the number of hierarchical clusters to be `n_hc = ceiling(length(Y)/10)`. To check sensitivity to `n_hc`, we now run the following:

```
summary(atebounds(Y, D, X, rps))
#> 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
summary(atebounds(Y, D, X, rps, n_hc = ceiling(length(Y)/5)))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, n_hc = ceiling(length(Y)/5))
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate          0.108089    0.002946
#> Confidence Interval    -0.047948    0.168398
summary(atebounds(Y, D, X, rps, n_hc = ceiling(length(Y)/20)))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, n_hc = ceiling(length(Y)/20))
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate          0.101290    0.079023
```

```
#> Confidence Interval    0.001227    0.181013
```

It can be seen that the alternative estimation result is more similar to the default one with `n_hc = ceiling(length(Y)/20)` than with `n_hc = ceiling(length(Y)/5)`. Overall, the results are qualitatively similar across the three different specifications of `n_hc`.

Finally, 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 experiment with `Q`.

```
  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
```

We also experiment with `n_hc`.

```
summary(attbounds(Y, D, X, rps))
#> 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
```

```
summary(attbounds(Y, D, X, rps, n_hc = ceiling(length(Y)/5)))
#> ATbounds: ATT
#> Call: attbounds(Y = Y, D = D, X = X, rps = rps, n_hc = ceiling(length(Y)/5))
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate          0.221885     0.13724
#> Confidence Interval     0.028392     0.32925
summary(attbounds(Y, D, X, rps, n_hc = ceiling(length(Y)/20)))
#> ATbounds: ATT
#> Call: attbounds(Y = Y, D = D, X = X, rps = rps, n_hc = ceiling(length(Y)/20))
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate          0.124074     0.10252
#> Confidence Interval    -0.009527     0.23778
```

Bound estimates cross (that is, the lower bound is larger than the upper bound) and furthermore sensitive to the choice of `Q` and `n_hc`. This is likely to be driven by the relatively small sample size. However, once we factor into sampling uncertainty and look at the confidence intervals, all the results are more or less similar.

NSW treated and PSID control

The Dehejia-Wahba sample can be regarded as a data scenario where the propensity score is known and satisfies the overlap condition. We now turn to a different data scenario where it is likely that the propensity score is unknown and may not satisfy the overlap condition.

```
  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 the sample proportion and obtain the new bound estimates:

```
  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
```

The confidence interval $[-0.35, 0.31]$ here is much larger than the confidence interval $[-0.01, 0.19]$ with the Dehejia-Wahba sample. It is worth noting that the sample proportion is unlikely to be correctly specified in the NSW-treated/PSID-control sample. Thus, it seems that our inference method produces a wider confidence interval in order to be robust against misspecification of the propensity scores.

We now consider the Manski bounds, which can be obtained by setting $Q = 1$.

```
  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 even 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 specify the reference propensity score).

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

```
  summary(attbounds(Y, D, X, rps_sp))
#> ATbounds: ATT
#> Call: attbounds(Y = Y, D = D, X = X, rps = rps_sp)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate           -0.10085     0.28040
#> Confidence Interval      -0.27838     0.67865
  detach(psid)
```

We find that the bounds on ATT are wide, as in ATE.

Case Study 2: Bounding the Effect of Right Heart Catheterization

As a second empirical example, we revisit the well-known Right Heart Catheterization Dataset. In particular, we apply our methods to Connors et al. (1996)'s study of the efficacy of right heart catheterization (RHC), which is a diagnostic procedure for directly measuring cardiac function in critically ill patients. This dataset has been subsequently used in the context of limited overlap by Crump et al. (2009), Rothe (2017), and Li, Morgan, and Zaslavsky (2018) among others. The dataset is publicly available on the Vanderbilt Biostatistics website at https://hbiostat.org/data.

In this example, the dependent variable is 1 if a patient survived after 30 days of admission, and 0 if a patient died within 30 days. The binary treatment variable is 1 if RHC was applied within 24 hours of admission, and 0 otherwise. The sample size was $n = 5735$, and 2184 patients were treated with RHC. There are a large number of covariates: Hirano and Imbens (2001) constructed 72 variables from the dataset and the same number of covariates were considered in both Crump et al. (2009) and Li, Morgan, and Zaslavsky (2018) and 50 covariates were used in Rothe (2017). In our exercise, we constructed the same 72 covariates. A cleaned version of the dataset is available in the package.

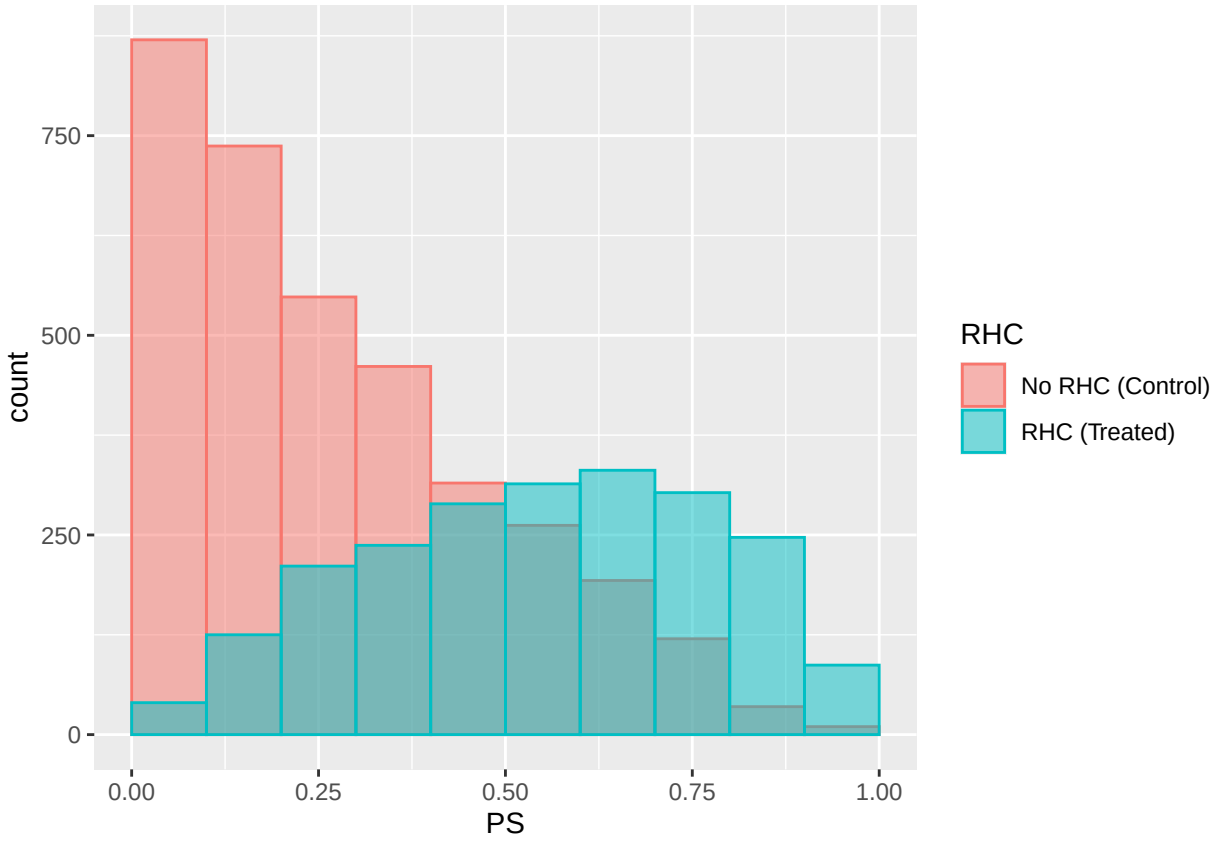
```
  Y <- RHC[,"survival"]
  D <- RHC[,"RHC"]
  X <- as.matrix(RHC[,-c(1,2)])
```

As in the aforementioned papers, we estimated the propensity scores by a logit model with all 72 covariates being added linearly.

```
  # Logit estimation of propensity score
  glm_ps <- stats::glm(D~X,family=binomial("logit"))
  ps <- glm_ps$fitted.values
  ps_treated <- ps[D==1]
  ps_control <- ps[D==0]
  # Plotting histograms of propensity scores
```

```r
df <- data.frame(cbind(D,ps))
colnames(df)<-c("RHC","PS")
df$RHC <- as.factor(df$RHC)
levels(df$RHC) <- c("No RHC (Control)", "RHC (Treated)")

ggplot2::ggplot(df, ggplot2::aes(x=PS, color=RHC, fill=RHC)) +
  ggplot2::geom_histogram(breaks=seq(0,1,0.1),alpha=0.5,position="identity")
```



The figure above shows the histograms of estimated propensity scores for treated and control groups. It is very similar to Fig.1 in Crump et al. (2009) and to Figure 2 in Rothe (2017). The support of the estimated propensity scores are almost on the unit interval for both treated and control units, although there is some visual evidence on limited overlap (that is, control units have much fewer propensity scores close to 1).

Bounding the Average Treatment Effect on the Treated

In this section, we focus on ATT. We first estimate ATT by the normalized inverse probability weighted estimator:

$$\widehat{\text{ATT}}_{\text{PS}} := \frac{\sum_{i=1}^{n} D_i Y_i}{\sum_{i=1}^{n} D_i} - \frac{\sum_{i=1}^{n} (1-D_i) W_i Y_i}{\sum_{i=1}^{n} (1-D_i) W_i},$$

where $W_i := \widehat{p}(X_i)/[1-\widehat{p}(X_i)]$ and $\widehat{p}(X_i)$ is the estimated propensity score for observation i based on the logit model described above. See, e.g., equation (3) and discussions in Busso, DiNardo, and McCrary (2014) for details of the normalized inverse probability weighted ATT estimator. The estimator $\widehat{\text{ATT}}_{\text{PS}}$ requires that the assumed propensity score model is correctly specified and the overlap condition is satisfied. The resulting estimate is $\widehat{\text{ATT}}_{\text{PS}} = -0.0639$.

```r
  # ATT normalized estimation
  y1_att <- mean(D*Y)/mean(D)
  att_wgt <- ps/(1-ps)
  y0_att_num <- mean((1-D)*att_wgt*Y)
  y0_att_den <- mean((1-D)*att_wgt)
  y0_att <- y0_att_num/y0_att_den
  att_ps <- y1_att - y0_att
  print(att_ps)
#> [1] -0.06388047
```

We now turn to our methods. As before, we take the reference propensity score to be $\widehat{p}_{\text{RPS}}(X_i) = n^{-1}\sum_{i=1}^{n} D_i$ for each observation i. That is, we assign the sample proportion of the treated to the reference propensity scores uniformly for all observations. Of course, this is likely to be misspecified; however, it has the advantage that $1/\widehat{p}_{\text{RPS}}(X_i)$ is never close to 0 or 1.

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

The resulting inverse reference-propensity-score weighted ATT estimator is

$$\widehat{\text{ATT}}_{\text{RPS}} := \frac{\sum_{i=1}^{n} D_i Y_i}{\sum_{i=1}^{n} D_i} - \frac{\sum_{i=1}^{n}(1-D_i)Y_i}{\sum_{i=1}^{n}(1-D_i)} = -0.0507.$$

```r
  att_rps <- mean(D*Y)/mean(D) - mean((1-D)*Y)/mean(1-D)
  print(att_rps)
#> [1] -0.05072115
```

When the sample proportion is used as the propensity score estimator, there is no difference between unnormalized and normalized versions of ATT estimates. In fact, it is simply the mean difference between treatment and control groups.

```r
   Xunique <- mgcv::uniquecombs(X) # A matrix of unique rows from X
   print(c("no. of unique rows:", nrow(Xunique)))
#> [1] "no. of unique rows:" "5735"
   print(c("sample size       :", nrow(X)))
#> [1] "sample size       :" "5735"
```

Note that none of the covariates in the observed sample are identical. We therefore implement hierarchical, agglomerative clustering with complete linkage.

We show empirical results using the default option.

```r
  summary(attbounds(Y, D, X, rps))
#> ATbounds: ATT
#> Call: attbounds(Y = Y, D = D, X = X, rps = rps)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate         -0.077412   -0.039280
#> Confidence Interval    -0.116903   -0.006371
```

to check sensitivity with respect to tuning parameters, we vary L and Q.

```r
  # Bounding  ATT: sensitivity analysis
  # not run to save time
  nhc_set <- c(5, 10, 20)
  results_att <- {}

  for (hc in nhc_set){
    nhc <- ceiling(length(Y)/hc)
```

```
  for (q in c(1,2,3,4)){
    res <- attbounds(Y, D, X, rps, Q = q, n_hc = nhc)
    results_att <- rbind(results_att,c(hc,q,res$est_lb,res$est_ub,res$ci_lb,res$ci_ub))
  }
}
colnames(results_att) = c("L","Q","LB","UB","CI-LB","CI-UB")
print(results_att, digits = 3)
```

The table below reports estimation results of ATT bounds for extended values of L and Q. When $Q = 1$, our estimated bounds correspond to Manski bounds, which includes zero and is wide with the interval length of almost one in all cases of L. Our bounds with $Q = 1$ are different across L because we apply hierarchical clustering before obtaining Manski bounds. With $Q = 2$, the bounds shrink so that the estimated upper bound is zero for all cases of L; with $Q = 3$, they shrink even further so that the upper end point of the 95% confidence interval excludes zero. Among three different values of L, the case of $L = 5$ gives the tightest confidence interval but in this case, the lower bound is larger than the upper bound, indicating that the estimates might be biased. In view of that, we take the bound estimates with $L = 10$ as our preferred estimates $[-0.077, -0.039]$ with the 95% confidence interval $[-0.117, 0.006]$. When $Q = 4$, the lower bound estimates exceed the upper bound estimates with $L = 5, 10$. However, the estimates with $L = 20$ give almost identical results to our preferred estimates. It seems that the pairs of $(L, Q) = (10, 3)$ or $(L, Q) = (20, 4)$ provide reasonable estimates.

Table 1: ATT Bounds: Right Heart Catheterization Study

L	Q	LB	UB	CI-LB	CI-UB
5	1	-0.638	0.282	-0.700	0.330
	2	-0.131	-0.000	-0.174	0.033
	3	-0.034	-0.048	-0.076	-0.007
	4	-0.006	-0.073	-0.079	-0.006
10	1	-0.664	0.307	-0.766	0.376
	2	-0.169	0.004	-0.216	0.039
	3	-0.077	-0.039	-0.117	-0.006
	4	-0.049	-0.057	-0.090	-0.016
20	1	-0.675	0.316	-0.843	0.430
	2	-0.178	-0.005	-0.238	0.034
	3	-0.099	-0.046	-0.149	-0.007
	4	-0.065	-0.060	-0.112	-0.017

The study of Connors et al. (1996) offered a conclusion that RHC could cause an increase in patient mortality. Based on our preferred estimates, we can exclude positive effects with confidence. This conclusion is based solely on the unconfoundedness condition, but not on the overlap condition, nor on the correct specification of the logit model. Overall, our estimates seem to be consistent with the qualitative findings in Connors et al. (1996).

Bounding the Average Treatment Effect

We now turn to bounds on ATE. Using again the sample proportion of the treated as the reference propensity score, we bound the ATE.

We start with $Q = 1$.

```
  summary(atebounds(Y, D, X, rps, Q = 1))
#> ATbounds: ATE
```

```
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 1)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate         -0.47742     0.48753
#> Confidence Interval    -0.54972     0.55664
```

We now consider $Q = 2$

```
  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         -0.17275    0.018032
#> Confidence Interval    -0.21518    0.063421
```

and $Q = 3$.

```
  summary(atebounds(Y, D, X, rps, Q = 3))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 3)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate         -0.11213   -0.058280
#> Confidence Interval    -0.14658   -0.022432
```

Finally, we take $Q = 4$.

```
  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         -0.083816   -0.122172
#> Confidence Interval    -0.143475   -0.057158
```

Overall, the results are similar to ATT.

Case Study 3: EFM

The electronic fetal monitoring (EFM) and cesarean section (CS) dataset from Neutra, Greenland, and Friedman (1980) consists of observations on 14,484 women who delivered at Beth Israel Hospital, Boston from January 1970 to December 1975. The purpose of the study is to evaluate the impact of EFM on cesarean section (CS) rates. Neutra, Greenland, and Friedman (1980) report that relevant confounding factors are: nulliparity (nullipar), arrest of labor progression (arrest), malpresentation (breech), and year of study (year). The dataset included in the R package is from the supplementary materials of Richardson, Robins, and Wang (2017), who used this dataset to illustrate their proposed methods for modeling and estimating relative risk and risk difference. In this daatset, all covariates are discrete.

```
  Y <- EFM[,"cesarean"]
  D <- EFM[,"monitor"]
  X <- as.matrix(EFM[,c("arrest", "breech", "nullipar", "year")])
  year <- EFM[,"year"]
```

```
  ate_rps <- mean(D*Y)/mean(D) - mean((1-D)*Y)/mean(1-D)
  print(ate_rps)
#> [1] 0.05005644
```

We take the reference propensity score to be the sample proportion of the treatment.

```
  rps <- rep(mean(D),length(D))
  print(rps[1])
#> [1] 0.5040044
```

Bounding the Average Treatment Effect

Using again the sample proportion of the treated as the reference propensity score, we bound the ATE.

We start with $Q = 1$.

```
  summary(atebounds(Y, D, X, rps, Q = 1, x_discrete = TRUE))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 1, x_discrete = TRUE)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate           -0.47811     0.52168
#> Confidence Interval      -0.67020     0.77400
```

We now consider $Q = 2$

```
  summary(atebounds(Y, D, X, rps, Q = 2, x_discrete = TRUE))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 2, x_discrete = TRUE)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate           -0.16105     0.15856
#> Confidence Interval      -0.26486     0.26448
```

and $Q = 3$.

```
  summary(atebounds(Y, D, X, rps, Q = 3, x_discrete = TRUE))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 3, x_discrete = TRUE)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate         -0.067808    0.091566
#> Confidence Interval    -0.107024    0.174704
```

In this example, the reference propensity score is close to 0.5, thus implying that the results will be robust even if we take a very large Q. In view of that, we take $Q = 5, 10, 20, 50, 100$.

```
  summary(atebounds(Y, D, X, rps, Q = 5, x_discrete = TRUE))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 5, x_discrete = TRUE)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate         -0.021842    0.034158
#> Confidence Interval    -0.043975    0.075407
  summary(atebounds(Y, D, X, rps, Q = 10, x_discrete = TRUE))
#> ATbounds: ATE
```

```
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 10, x_discrete = TRUE)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate       -0.0027345   0.0051908
#> Confidence Interval  -0.0208590   0.0228555
  summary(atebounds(Y, D, X, rps, Q = 20, x_discrete = TRUE))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 20, x_discrete = TRUE)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate        0.0019348   -0.000096
#> Confidence Interval  -0.0169394    0.018765
  summary(atebounds(Y, D, X, rps, Q = 50, x_discrete = TRUE))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 50, x_discrete = TRUE)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate         0.017103  -0.0028133
#> Confidence Interval   -0.016110   0.0290373
  summary(atebounds(Y, D, X, rps, Q = 100, x_discrete = TRUE))
#> ATbounds: ATE
#> Call: atebounds(Y = Y, D = D, X = X, rps = rps, Q = 100, x_discrete = TRUE)
#> Confidence Level: 0.95
#>                      Lower_Bound Upper_Bound
#> Bound Estimate         0.046554   -0.013941
#> Confidence Interval   -0.023382    0.041148
```

Overall, the empirical results suggest that there is no significant effect of EFM on cesarean section rates.

References

Busso, Matias, John DiNardo, and Justin McCrary. 2014. "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators." *Review of Economics and Statistics* 96 (5): 885–97.

Connors, Jr, Alfred F., Theodore Speroff, Neal V. Dawson, Charles Thomas, Jr Harrell Frank E., Douglas Wagner, Norman Desbiens, et al. 1996. "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically III Patients." *JAMA* 276 (11): 889–97.

Crump, Richard K, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika* 96 (1): 187–99.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–62.

———. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84 (1): 151–61.

Hirano, Keisuke, and Guido W Imbens. 2001. "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services and Outcomes Research Methodology* 2 (3-4): 259–78.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review* 76 (4): 604–20.

Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. 2018. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association* 113 (521): 390–400.

Manski, Charles F. 1989. "Anatomy of the Selection Problem." *Journal of Human Resources*, 343–60.

———. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80 (2): 319–23.

Neutra, R. R., S. Greenland, and E. A. Friedman. 1980. "Effect of Fetal Monitoring on Cesarean Section Rates." *Obstetrics and Gynecology* 55 (2): 175–80.

Richardson, T. S., J. M. Robins, and L. Wang. 2017. "On Modeling and Estimation for the Relative Risk and Risk Difference." *Journal of the American Statistical Association* 112 (519): 1121–30.

Rothe, Christoph. 2017. "Robust Confidence Intervals for Average Treatment Effects Under Limited Overlap." *Econometrica* 85 (2): 645–60.

Stoye, Jörg. 2020. "A Simple, Short, but Never-Empty Confidence Interval for Partially Identified Parameters." *arXiv:2010.10484 [Econ.EM]*. https://arxiv.org/abs/2010.10484.