Causal Inference in Case-Control and Case-Population Studies: Vignette

Overview

This vignette describes how to use package "ciccr" that is based on the paper entitled "Causal Inference under Outcome-Based Sampling with Monotonicity Assumptions" (Jun and Lee, 2023).

Causal Inference on Relative and Attributable Risk

Case-Control Sampling

We first load the ciccr and MASS packages.

```
library(ciccr)
library(MASS)
```

To illustrate the usefulness of the package, we use the dataset ACS_CC that is included in the package. This dataset is an extract from American Community Survey (ACS) 2018, restricted to white males residing in California with at least a bachelor's degree. The ACS is an ongoing annual survey by the US Census Bureau that provides key information about the US population. We use the following variables:

```
  y = ACS_CC$topincome
  t = ACS_CC$baplus
  x = ACS_CC$age
```

- The binary outcome 'Top Income' (Y) is defined to be one if a respondent's annual total pre-tax wage and salary income is top-coded. In our sample extract, the top-coded income bracket has median income $565,000 and the next highest income that is not top-coded is $327,000.
- The binary treatment (T) is defined to be one if a respondent has a master's degree, a professional degree, or a doctoral degree.
- The covariate (X) is age in years and is restricted to be between 25 and 70.

The original ACS sample is not a case-control sample but we construct one by the following procedure.

1. The case sample ($Y = 1$) is composed of 921 individuals whose income is top-coded.
2. The control sample ($Y = 0$) of equal size is randomly drawn without replacement from the pool of individuals whose income is not top-coded.

We now construct cubic b-spline terms with three inner knots using the age variable.

```
  x = splines::bs(x, df = 6)
```

Causal Inference on Relative Risk Using Case-Control Samples

Define $\beta(y) = E[\log \mathrm{OR}(X)|Y = y]$ for $y = 0, 1$, where $\mathrm{OR}(x)$ is the odds ratio conditional on $X = x$:

$$\mathrm{OR}(x) = \frac{P(T = 1|Y = 1, X = x)}{P(T = 0|Y = 1, X = x)} \frac{P(T = 0|Y = 0, X = x)}{P(T = 1|Y = 0, X = x)}.$$

Using the retrospective sieve logistic regression model, we estimate $\beta(1)$ by

```r
  results_case = avg_RR_logit(y, t, x, 'case')
  results_case$est
#>         y
#> 0.7286012
  results_case$se
#>         y
#> 0.1013445
```

Here, option `'case'` refers to conditioning on $Y = 1$.

Similarly, we estimate $\beta(0)$ by

```r
  results_control = avg_RR_logit(y, t, x, 'control')
  results_control$est
#>         y
#> 0.5469094
  results_control$se
#>         y
#> 0.1518441
```

Here, option `'control'` refers to conditioning on $Y = 0$.

We carry out causal inference on relative risk by

```r
results = cicc_RR(y, t, x, 'cc', 0.95)
```

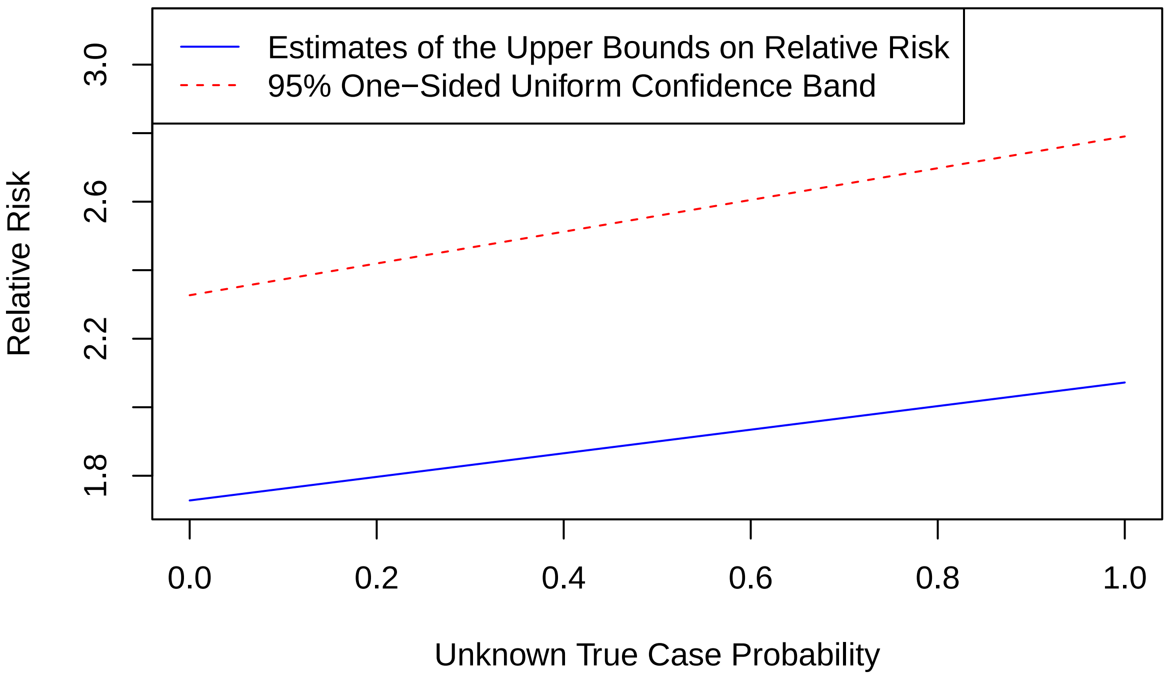
Here, 'cc' refers to case-control sampling and 0.95 refers to the level of the uniform confidence band; 0.95 is the default choice.

The S3 object `results` contains estimates `est`, standard errors `se`, and one-sided confidence bands `ci` at `p = 0` and `p = 1`.

```r
 # point estimates
 results$est
#>         y         y
#> 0.5469094 0.7286012
 # standard errors
 results$se
#>         y         y
#> 0.1518441 0.1013445
 # confidence intervals
 results$ci
#>         y         y
#> 0.8445183 1.0262101
```

It is handy to examine the results by plotting a graph.

```r
cicc_plot(results)
```



To interpret the results, we assume both marginal treatment response (MTR) and marginal treatment selection (MTS). In this setting, MTR means that everyone will earn no less by obtaining a degree higher than bachelor's degree; MTS indicates that those who selected into higher education have higher potential to earn top incomes. Based on the MTR and MTS assumptions, we can conclude that the treatment effect lies in between 1 and the upper end point of the one-sided confidence interval with high probability. Thus, the estimates in the graph above suggest that the effect of obtaining a degree higher than bachelor's degree is anywhere between 1 and the upper end points of the uniform confidence bands. This roughly implies that the chance of earning top incomes may increase up to by a factor as large as the upper end points of the uniform confidence band, but allowing for possibility of no positive effect at all. The results are shown over the range of the unknown true case probability. See Jun and Lee, 2023 for more detailed explanations regarding how to interpret the estimation results.

Comparison with Logistic Regression

We can compare these results with estimates obtained from logistic regression.

```
logit = stats::glm(y~t+x, family=stats::binomial("logit"))
est_logit = stats::coef(logit)
ci_logit = stats::confint(logit, level = 0.9)
#> Waiting for profiling to be done...
# point estimate
exp(est_logit)
#> (Intercept)           t          x1          x2          x3          x4          x5          x6
#>  0.05461156  2.06117153  4.42179639 12.99601849 19.03962976 26.83565737  6.42381406 26.14359394
# confidence interval
exp(ci_logit)
#>                     5 %       95 %
#> (Intercept)  0.01960819  0.1304108
#> t            1.75166056  2.4271287
#> x1           1.05679997 21.6604223
#> x2           5.50583091 33.8909622
#> x3           6.79458010 61.3258710
#> x4          10.22943808 78.7353953
```

```
#> x5          2.00536450 22.8509008
#> x6          8.66983039 87.6311482
```

Here, the relevant coefficient is 2.06 (`t`) and its two-sided 90% confidence interval is [1.75, 2.43]. If strong ignorability were plausible and causal relative risk were homogeneous, then the treatment effect would be about 2 and its two-sided confidence interval would be between [1.75, 2.43]. However, it is unlikely that the higher BA treatment satisfies the strong ignorability condition.

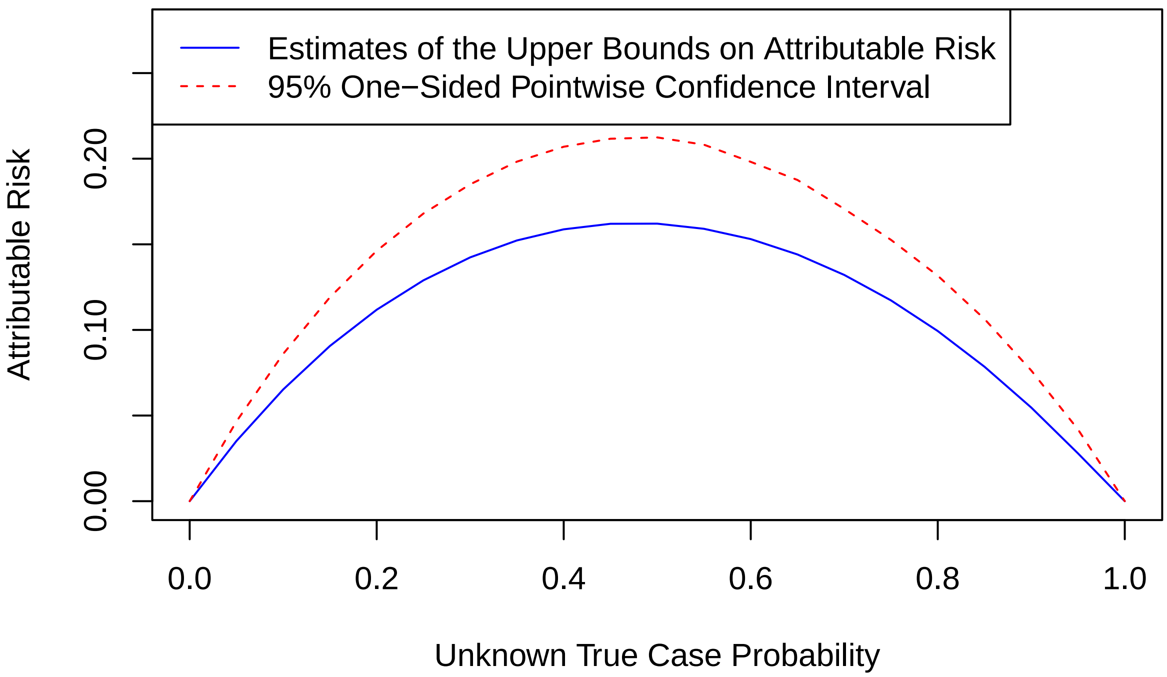
Causal Inference on Attributable Risk Using Case-Control Samples

We now consider attributable risk, that is the difference in probabilities. We carry out causal inference on attributable risk by

```
results_AR = cicc_AR(y, t, x, sampling = 'cc', no_boot = 100)
```

The results can be plotted as before.

```
cicc_plot(results_AR, parameter = 'AR')
```



The upper bounds are approximately inverted-U shaped. When $p = 0$ or $p = 1$, there can be no causal effect; the upper bound is maximized around $p = 0.5$.

Causal Inference on Relative Risk Using Case-Population Samples

We now consider an example of case-population samples. For this purpose, we use the dataset ACS_CP that is included in the package. This dataset is again an extract from American Community Survey (ACS) 2018. The original ACS sample is not a case-population sample but we construct one by the following procedure.

1. The case sample ($Y = 1$) is composed of 921 individuals whose income is top-coded.
2. The control sample ($Y = 0$) of equal size is randomly drawn with replacement from all observations and its top-coded status is coded missing.

We use the following variables:

```r
y = ACS_CP$topincome
t = ACS_CP$baplus
x = ACS_CP$age
```

We print `y` to see how the outcome variable is coded.

```r
print(head(y))
#> [1] NA  1  1 NA NA NA
```

We now code missing Y by 0 in the population sample.

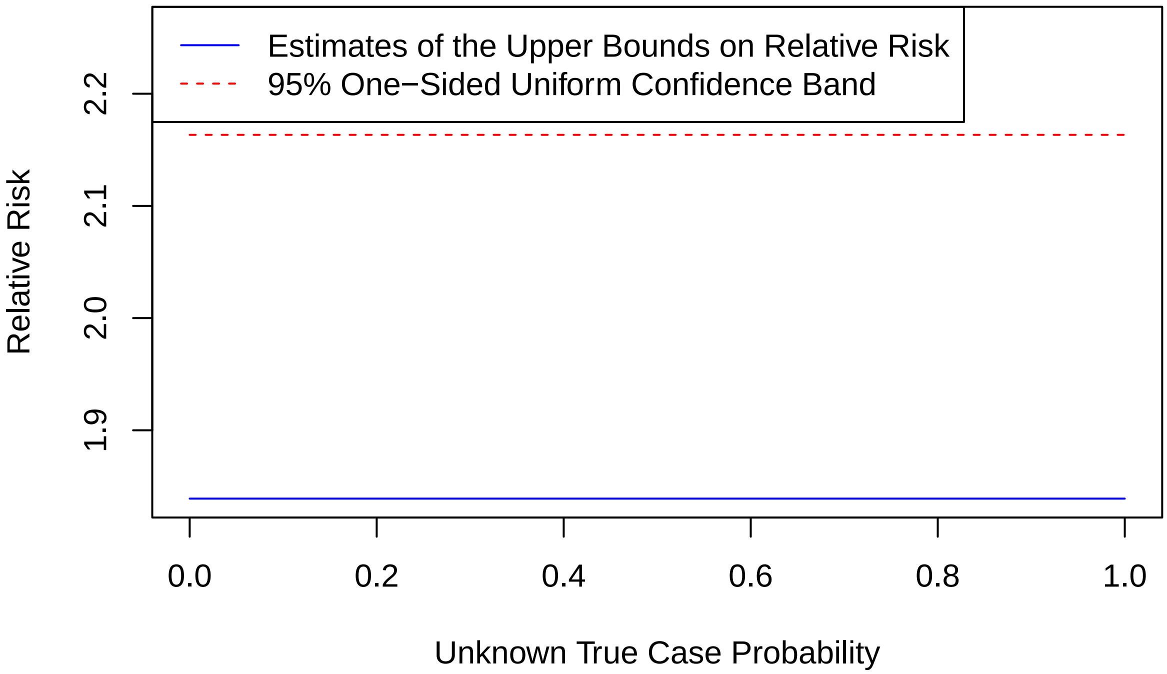
```r
y = as.integer(is.na(y)==FALSE)
```

We estimate $\beta(0)$ by

```r
results_control = avg_RR_logit(y, t, x, 'control')
results_control$est
#>         y
#> 0.6092686
results_control$se
#>        y
#> 0.098725
```

and carry out causal inference on relative risk by

```r
results = cicc_RR(y, t, x, 'cp', 0.95)
cicc_plot(results)
```



Note that the estimates and upper bounds are constant across the unknown true case probability. This is because they do not depend on the value of the case probability in the case-population sample.

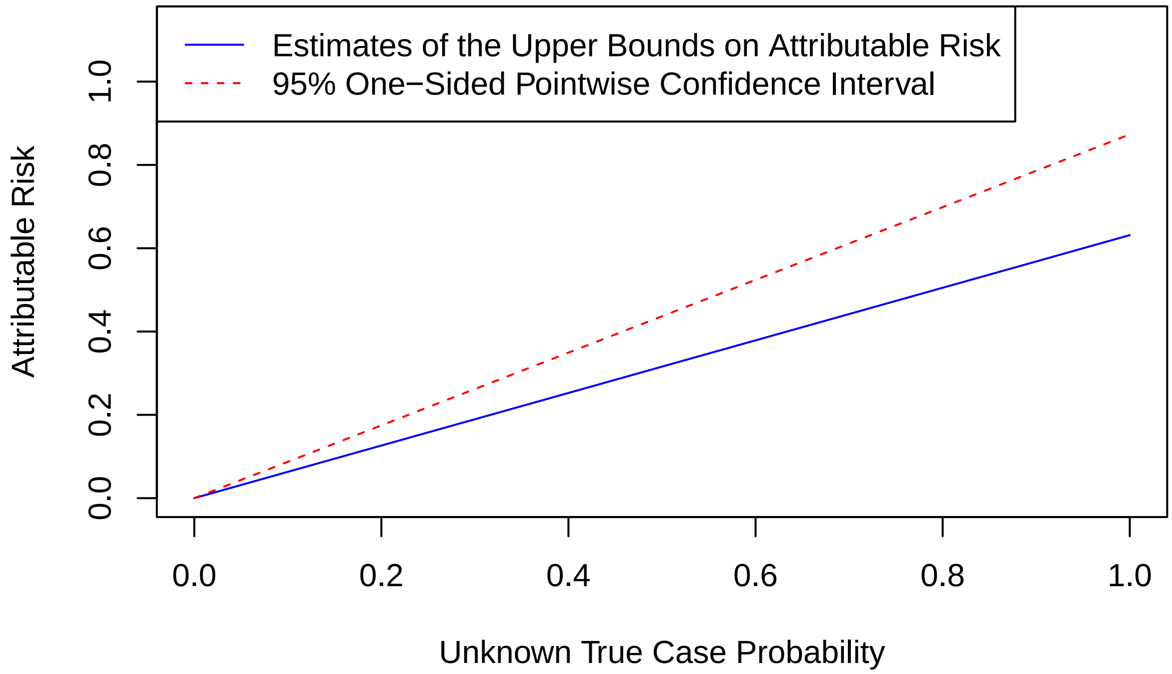
Causal Inference on Attributable Risk Using Case-Population Samples

We now consider causal inference on attributable risk by

```r
results_AR = cicc_AR(y, t, x, sampling = 'cp', no_boot = 100)
```

The results can be plotted as before.

```r
cicc_plot(results_AR, parameter = 'AR')
```



For case-population sampling, the upper bound on attributable risk is a linear function of the unknown true case probability; as a result, it increases linearly, as shown in the figure.

References

Sung Jae Jun and Sokbae Lee. (2023). Causal Inference under Outcome-Based Sampling with Monotonicity Assumptions. https://arxiv.org/abs/2004.08318.

Manski, C.F. (1997). Monotone Treatment Response. Econometrica, 65(6), 1311-1334.

Manski, C.F. and Pepper, J.V. (2000). Monotone Instrumental Variables: With an Application to the Returns to Schooling. Econometrica, 68(4), 997-1010.