Attributable Risk

Louis Schenck
Elizabeth J. Atkinson
Cynthia S. Crowson
Terry M. Therneau

Mayo Clinic, Rochester, MN

November 18, 2014

1 Introductions and Definitions

The attribrisk package has been created to make attributable risk (AR) inference easier to calculate. This vignette gives only a small amount of theory, users who wish more are directed to Khan [3].

The term *attributable risk* is used to refer to the amount of disease in a population that can be attributed to a particular exposure. The terms *population attributable risk* and *etiologic fraction* have also been used to refer to the same concept. Consider the data from Whisnant et al [6] shown in table 1. Assume for the moment that this data were a simple sample of size 2644 from the population at large (which it is not). Then the rate of ischemic stroke among those without high blood pressure would be estimated as 559/943=0.41 while that in those with high blood pressure it is 0.55. The attributatable risk is the fraction of the strokes in the second group that is presumed to be directly due to their high blood pressure, namely $(0.55 - 0.41)/0.55 = 0.25$.

In reality the data set above is not a population sample, but a collection of all ischemic stroke patients in Olmsted County over a defined period along with a set of age matched controls, and the proper calculation of AR becomes more complex in that case.

Define the following probabilities. They may be interpeted as prevalence rates, incidence rates or proportions depending on the circumstances.

Normal   High Blood

```
Controls & 559  & 763 \\
Stroke   & 384  & 938 \\ \hline
Total    & 943  & 1701
```

Table 1: Table of data from Whisnant et al [6].

$\mathrm{P}(D)$ - probability of disease

$\mathrm{P}(F)$ - probability of risk factor

$\mathrm{P}(D|F)$ - conditional probability of disease among those with the factor

$\mathrm{P}(D|\overline{F})$ - conditional probability of disease among those without the factor

$\mathrm{P}(F|D)$ - conditional probability of the factor being present among those who are diseased

$\mathrm{P}(F|\overline{D})$ - conditional probability of the factor being present among those not diseased

$\mathrm{P}(D|F)/\mathrm{P}(D|\overline{F})$ - relative risk (RR)

The population attributable risk is defined as I^*/I where I is the total number of diseased individuals in the population and I^* is defined as the number attributed to a factor. Assuming a homogeneous population, $I = N\mathrm{P}(D)$ = expected number of diseased individuals in the population, $I^* = N\mathrm{P}(D) - N\mathrm{P}(D|\overline{F})$ = excess number of diseased individuals. Then

$$\mathrm{AR} = I^*/I = \frac{\mathrm{P}(D) - \mathrm{P}(D|\overline{F})}{\mathrm{P}(D)} \tag{1}$$

This formula for AR requires estimates of the rate (prevalence or incidence) of disease in the total population and the rate of disease in those with the factor. Using Bayes' theorem the equation can be expressed as

$$\mathrm{AR} = \frac{\mathrm{P}(F)(\mathrm{RR}-1)}{\mathrm{P}(F)(\mathrm{RR}-1)+1} \tag{2}$$

The formula for AR can also be represented by

$$\mathrm{AR} = \frac{\mathrm{P}(F|D)(\mathrm{RR}-1)}{(RR)} \tag{3}$$

This formula is useful in case-control studies where $\mathrm{P}(F|D)$ in the prevalence of the factors among the cases, and RR can be estimated using an odds ratio obtained from a logistic regression with case/control status as the outcome. Logisic regression, either stratified (for matched case-control studies) or unstratified, is the basis for the attribrisk function.

2 Examples

2.1 Unmatched case-control design with dichotomous exposure and no confounders

The data from Whisnant, et al 1996 concerning cerebral infarction (CI) and high blood pressure (HBP) is treated as an unmatched design [6]. Data can be found in the data frame chapter.dat in the attribrisk package.

```
> require(attribrisk)
> data(chapter.dat)
>
> #Show first and last row.
> chapter.dat[c(1,2644),]

       hbp   dm   ihd cases match.id
1    HBP-0 DM-0 IHD-0     0        1
2644 HBP-1 DM-1 IHD-1     1     1322

> #Summarize the relationship between hbp and case/control status.
> count <- table(chapter.dat$hbp, chapter.dat$cases)
> count

          0   1
  HBP-0 559 384
  HBP-1 763 938
```

From this table the odds ratio and prevalence are

$$\widehat{OR} = \frac{(938 * 559)}{(763 * 384)} = 1.79$$

$$\hat{\mathrm{P}}(F|D) = \frac{938}{938 + 384} = 0.71$$

Using equation 3 we get

$$\widehat{AR} = \frac{0.71 * (1.79 - 1)}{1.79} = 0.313$$

The R call for this is below.

```
> example1 <- attribrisk(cases ~ expos(hbp), data=chapter.dat)
>
> example1

Call:
attribrisk(formula = cases ~ expos(hbp), data = chapter.dat)



 Degrees of freedom: 2643
                  coefficient std. err lower 0.95 upper 0.95
attributable risk      0.3131   0.0347     0.3811     0.2451
```

Attributable risk is one of the few cases where a user may want to calculate the estimate from summary data found in a paper, e.g., data such as in table 1. The function allows this via the use of case weights, as shown below. The example starts by recreating the data from the table.

```
> tdata <- data.frame(case=c(0, 1, 0, 1),
+                     hbp =c(0, 0, 1, 1),
+                     count = c(559, 384, 763, 938))
> example1b <- attribrisk(case ~ expos(hbp), data=tdata, weight=count)
> example1b

Call:
attribrisk(formula = case ~ expos(hbp), data = tdata, weights = count)


 Degrees of freedom: 3
                  coefficient std. err lower 0.95 upper 0.95
attributable risk      0.3131   0.0369     0.3854     0.2407
```

The AR estimates in the above two fits are identical, but the standard errors are not. This is because by default the `attribrisk` function uses a grouped jackknife estimate of variance based on $k = 20$ groups, which leaves out one group at a time. Subjects are divided into the k groups randomly and hence different runs will have slightly different jackknife errors. The full jackknife can be obtained by setting $k \geq n$ where n is the data set size, but with a consequent increase in the compute time. Bootstrap standard errors and confidence intervals are also available.

```
> example1boot <- attribrisk(cases ~ expos(hbp), data=chapter.dat,
+    varmethod = "boot")
> example1boot

Call:
attribrisk(formula = cases ~ expos(hbp), data = chapter.dat,
    varmethod = "boot")


 Degrees of freedom: 2643
                  coefficient std. err lower 0.95 upper 0.95
attributable risk      0.3131  0.03968     0.2293      0.387
```

2.2 Matched case-control design with dichotomous exposure and no confounders

This example uses attribrisk to estimate AR and its standard error in a matched case-control design. The data used in this example is the same as the previous example. The pairs were matched on age and sex. The variable `match.id` is used as a matching index. For example, row 1 and 1323 were a matched pair since both have `match.id` = 1.

```
> chapter.dat[chapter.dat$match.id==1,]

       hbp   dm   ihd cases match.id
1    HBP-0 DM-0 IHD-0     0        1
1323 HBP-0 DM-0 IHD-0     1        1

> example2 <- attribrisk(cases ~ strata(match.id) + expos(hbp),
+    data=chapter.dat)
>
> example2

Call:
attribrisk(formula = cases ~ strata(match.id) + expos(hbp), data = chapter.dat)



n =  2644
                  coefficient std. err lower 0.95 upper 0.95
attributable risk      0.3311  0.03791     0.4054     0.2568
```

The special function `strata` is used in the formula to tell the attribrisk function which observations go together. There can be at most one matching index. The AR estimated in the unmatched design was 0.3311 which is similar to the unmatched value of 0.3131 from the unmatched analysis.

2.3 More complex target values

The next two examples look at the computation when one or more adjusting factors are continuous, using a data set on cerebral infarctions (stroke). The following covariates are considered for each patient: age at study entry (age), diastolic blood pressure (diastolic) and smoking level (smoke). The age of the individual needs to be accounted for since older individuals are more likely to suffer a stroke. Age is a confounder, however, rather than an exposure variable since everyone ages; we cannot “adjust” it to a different value via some change in lifestyle. The public health impact of changing smoking or diatolic blood pressure levels are the questions of interest.

Consider ”smoke” and ”diastolic” factors as exposure variables. Smoking status has 5 levels — Current, Former, Never, Unknown, and Uncertain — and diastolic blood pressure is a continuous measure. The default action of the attribrisk function is to compute the effect of making all exposure variables equal to 0, which is not sensible for blood pressure; a target value needs to be chosen by the user. The target values can be chosen to represent a hypothetical intervention or the desired level of the exposure, such as absence of the risk factor.

Consider target values where all patients *Never* smoked and all subjects’ diastolic blood pressure values were lowered by 10%.

```
> # Build Targe
> data(stroke.dat)
```

```
>
> stroke.target <- data.frame(smoke = "Never",
+                             diastolic = .9*stroke.dat$diastolic)
>
> set.seed(21790)
> example4a <- attribrisk(
+    cases ~ age + expos(smoke) + expos(diastolic),
+    data=stroke.dat, varmethod="boot", baseline = stroke.target)
>
> example4a

Call:
attribrisk(formula = cases ~ age + expos(smoke) + expos(diastolic),
    data = stroke.dat, varmethod = "boot", baseline = stroke.target)



 Degrees of freedom: 2022
                  coefficient std. err lower 0.95 upper 0.95
attributable risk      0.4282  0.03779     0.3537     0.5081
```

The estimated AR is 0.428 with the standard error of about 0.038. The estimated 95% bootstrap confidence interval suggests that if the entire population had never started smoking and had 10% lower blood pressure than they do currently, there would be between 35 and 51 fewer strokes.

Next, consider the target values where all current smokers quit (i.e., change to former smokers) and the percent reduction in diastolic blood pressure depends on the diastolic blood pressure value, as in described in this table.

Current Diastolic BP	Percent Reduction
>150	25%
121 to 150	15%
101 to 120	10%
86 to 100	5%
$\leq$ 85	no reduction

```
> # Build baseline
> target <- cut(stroke.dat$diastolic, c(0, 85, 100, 120, 150, 500))
> reduce <- c(0, .05, .1, .15, .25)[as.numeric(target)]
> newbp <-with(stroke.dat, diastolic *(1-reduce))
> newsm <- with(stroke.dat, ifelse(smoke=="Current", "Former", smoke))
>
> stroke.target2 <- data.frame(diastolic = newbp,
+                               smoke = newsm)
> example4b <- attribrisk(
+    cases ~ age + expos(smoke) + expos(diastolic),
+    data=stroke.dat, baseline = stroke.target2)
>
> print(example4b, digits=3)

Call:
attribrisk(formula = cases ~ age + expos(smoke) + expos(diastolic),
    data = stroke.dat, baseline = stroke.target2)


 Degrees of freedom: 2022
                  coefficient std. err lower 0.95 upper 0.95
attributable risk       0.117   0.0218       0.16     0.0743
```

So if all current smokers quit and people with high diastolic BP reduced their BP, the AR would be about 11.7% The different target values yielded quite different potential reductions for the amount of stroke, where the second is more realistic and the first perhaps only a pipe dream.

3 Conclusion

The paper describes how to use the attribrisk function in R for unmatched and matched case-control designs. The attribrisk function enables modeling with any number of confounders/adjustors and exposures, which can be discrete or continuous, and allows for flexibility in defining target values.

A Comparing Results to Benichou (1991)

Data from table I of Benichou and Gail [1], which computes attributable risks for esophogeal cancer due to smoking and alchohol, is be found in the data frame *benichou* in the attribrisk package. The following table compares all the AR results reported by Benichou in tables III, IV, and V; and the results calculated using the attribrisk package. The standard errors shown for the attribrisk function are from the default 20-fold grouped jackknife. When the full n-fold jackknife is used they all agree with Benichou and Gail, but the resulting vignette takes longer to run than CRAN guidelines allow.

		Benichou & Gail		attribrisk	
Model	Formula	$\widehat{AR}$	$\widehat{SD}$	$\widehat{AR}$	$\widehat{SD}$
1	expos(alcohol80)	0.40	0.042	0.39	0.036
2	age * smoke + expos(alcohol80)	0.38	0.044	0.38	0.043
3	age * (smoke + expos(alcohol80))	0.38	0.044	0.38	0.036
4	smoke * (age + expos(alcohol80))	0.38	0.044	0.38	0.045
5	age * smoke * expos(alcohol80)	0.38	0.044	0.38	0.041
6	expos(alcohol40)	0.71	0.051	0.71	0.048
7	age * smoke + expos(alcohol40)	0.72	0.050	0.72	0.065
8	age * (smoke + expos(alcohol40))	0.72	0.050	0.72	0.059
9	smoke * (age + expos(alcohol40))	0.70	0.054	0.70	0.055
10	age * smoke * expos(alcohol40)	0.70	0.056	0.70	0.056
11	expos(alcohol)	0.71	0.051	0.71	0.052
12	age * smoke + expos(alcohol)	0.72	0.050	0.72	0.060
13	age * (smoke + expos(alcohol))	0.73	0.050	0.73	0.059
14	smoke * (age + expos(alcohol))	0.70	0.054	0.70	0.050
15	age * smoke * expos(alcohol)	0.70	NA	0.70	0.058
16	expos(fsmoke.alc)	0.86	0.046	0.86	0.048
17	age + expos(fsmoke.alc)	0.87	0.045	0.87	0.040
18	age * expos(fsmoke.alc)	0.87	0.044	0.87	0.047

In models 1–5, alcohol consumption was considered a binary exposure factor (0–79; 80+ g/day). Age, smoking, and their interactions were used as main effects in models 2–5. The attributable risk shows the estimated effect if all those consuming more than 80 g/day were to convert to the lowest category of 0–39.

In models 6–10, alcohol consumption was split at 40g/day. The estmated effect of having all these subjects decrease to 0–39 is much larger, largely due to the greater number of subjects affected by the intervention. Age, smoking, and their interactions were used as main effects in models 7–10.

In models 11–15, alcohol consumption was considered a polychotomus exposure factor (0–39; 40–79, 80–119, 120+ g/day). The target alcohol value is still the first category of 0–39, as in the prior set of models and the same subjects are affected. The impact changes slightly since the underlying model has a different risk estimate for each of the 40–79, 80–119 and 120+ groups. In model 15, twenty-four parameters are used to model the interactions of alcohol consumption with smoking and age. Model 15 was a saturated model and the delta method cannot provide an estimate of the standard error. Note:

Even though model 15 was saturated, the jackknife standard error estimated using the attribrisk package is reasonable.

In models 16–18 the exposure is the single binary variable for either smoking or drinking above the baseline level.

References

[1] Benichou J; *Methods of Adjustment for Estimating the Attributable Risk in Case-Control Studies: A Review.* Statistics in Medicine (1991) 10:1753-1773.

[2] Benichou J, Gail M; *Variance Calculations and Confidence Intervals for Estimates of the Attributable Risk Based on Logistic Models.* Biometrics (1990) 46:991-1003.

[3] Kahn M, O'Fallon W, Sicks J; *Technical Report Series No. 54, Generalized Population Attributable Risk Estimation.* Department of Health Science Research, Mayo Clinic, Rochester, Minnesota (2000).

[4] Kleinbaum D, Kupper L and Morgenstern H; *Epidemiologic Research: Principles and Quantitative Methods Lifetime Learning Publications.* Belmont, California (1982).

[5] Levin M; *The occurrence of lung cancer in man.* Acta Un Intern Cancer (1953) 19:531-541.

[6] Whisnant J, Wiebers D, O'Fallon W, Sicks J, Frye R; *A population-based model of risk factors for ischemic stroke: Rochester, Minnesota.* Neurology (1996) 47:1420-1428.