samplesizelogisticcasecontrol Package

September 24, 2021

```
> library(samplesizelogisticcasecontrol)
```

Random data generation functions

Let X_1 and X_2 be two variables with a bivariate normal ditribution with mean (0, 0) and covariance [1, 0.5; 0.5, 2]. X_2 corresponds to the exposure of interest. Let $X_3 = X_1X_2$ and define functions for generating random data from the distribution of (X_1, X_2) and (X_1, X_2, X_3).

```
> mymvn <- function(n) {
+   mu    <- c(0, 0)
+   sigma <- matrix(c(1, 0.5, 0.5, 2), byrow=TRUE, nrow=2, ncol=2)
+   dat   <- rmvnorm(n, mean=mu, sigma=sigma)
+   dat
+ }
> myF <- function(n) {
+   dat <- mymvn(n)
+   dat <- cbind(dat, dat[, 1]*dat[, 2])
+   dat
+ }
```

Generate some data

```
> data <- myF(200)
> colnames(data) <- paste("X", 1:3, sep="")
> data[1:5, ]
```

```
            X1          X2          X3
[1,] -0.1141327  1.85152786 -0.21131996
[2,]  1.2357848  1.29814042  1.60422227
[3,] -1.0944167  0.06226012 -0.06813852
[4,] -0.9736883 -2.26825489  2.20857336
[5,]  0.5367609 -1.64848892 -0.88484438
```

Examples of univariate calculations

We have the logistic model $logit = \mu + \beta X$ and are testing $\beta = 0$. Suppose the disease prevalence is 0.01, the log-odds ratio for the exposure X is 0.26 and that the exposure follows a Bernoulli(p) distribution with p = 0.15.

```
> prev  <- 0.01
> logOR <- 0.26
> p <- 0.15
```

Compute the sample sizes

```
> sampleSize_binary(prev, logOR, probXeq1=p)

$ss.wald.1
[1] 4472

$ss.wald.2
[1] 4498

$ss.score.1
[1] 4467

$ss.score.2
[1] 4441
```

The same result can be obtained assuming X is ordinal and passing in the 2 probabilities $P(X = 0)$ and $P(X = 1)$.

```
> sampleSize_ordinal(prev, logOR, probX=c(1-p, p))

$ss.wald.1
[1] 4472

$ss.wald.2
[1] 4498

$ss.score.1
[1] 4467

$ss.score.2
[1] 4440
```

Let X be ordinal with 3 levels. The vector being passed into the probX argument below is $(P(X = 0), P(X = 1), P(X = 2))$.

```
> sampleSize_ordinal(prev, logOR, probX=c(0.4, 0.35, 0.25))

$ss.wald.1
[1] 975

$ss.wald.2
[1] 985

$ss.score.1
[1] 973

$ss.score.2
[1] 963
```

Now let the exposure X be $N(0, 1)$.

```
> sampleSize_continuous(prev, logOR)

$ss.wald.1
[1] 625

$ss.wald.2
[1] 644

$ss.score.1
[1] 621

$ss.score.2
[1] 602
```

For the univariate case with continuous exposure, we can specify the probability density function of X in different ways. Consider X to have a chi-squared distribution with 1 degree of freedom. Note that the domain of a chi-squared pdf is from 0 to infinity, and that the var(X) = 2.

```
> sampleSize_continuous(prev, logOR, distF="dchisq(x, 1)",
+                       distF.support=c(0,Inf), distF.var=2)

$ss.wald.1
[1] 170

$ss.wald.2
[1] 242

$ss.score.1
[1] 168

$ss.score.2
[1] 113

> f <- function(x) {dchisq(x, 1)}
> sampleSize_continuous(prev, logOR, distF=f, distF.support=c(0, Inf),
+                        distF.var=2)

$ss.wald.1
[1] 170

$ss.wald.2
[1] 242

$ss.score.1
[1] 168

$ss.score.2
[1] 113
```

If we do not set *distF.var*, then the variance of X will be approximated by numerical integration and could yield slightly different results.

```
> sampleSize_continuous(prev, logOR, distF="dchisq(x, 1)", distF.support=c(0,Inf))

$ss.wald.1
[1] 170

$ss.wald.2
[1] 242

$ss.score.1
[1] 168

$ss.score.2
[1] 113
```

Let X have the distribution defined by column X_1 in data.

```
> sampleSize_data(prev, logOR, data[, "X1", drop=FALSE])

$ss.wald.1
[1] 614

$ss.wald.2
[1] 632

$ss.score.1
[1] 611

$ss.score.2
[1] 594
```

Examples with confounders

We have the logit model $logit = \mu + \beta_1 X_1 + \beta_2 X_2$ and are interested in testing $\beta_2 = 0$. Here we must have log-odds ratios for X_1 and X_2, and we will use the distribution function mymvn defined above to generate 200 random samples. Note that logOR[1] corresponds to X_1 and logOR[2] corresponds to X_2.

```
> logOR <- c(0.1, 0.13)
> sampleSize_data(prev, logOR, mymvn(200))

$ss.wald.1
[1] 1350

$ss.wald.2
[1] 1370

$ss.score.1
[1] 1347
```

```
$ss.score.2
[1] 1328
```

Now we would like to perform a test of interaction, $\beta_3 = 0$, where $logit = \mu + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ and $X_3 = X_1 X_2$. The vector of log-odds ratios must be of length 3 and in the same order as (X_1, X_2, X_3).

```
> logOR <- c(0.1, 0.15, 0.11)
> sampleSize_data(prev, logOR, myF(1000))

$ss.wald.1
[1] 1405

$ss.wald.2
[1] 1447

$ss.score.1
[1] 1391

$ss.score.2
[1] 1350
```

Pilot data from a file

Suppose we want to compute sample sizes for a case-control study where we have pilot data from a previous study. The pilot data is stored in the file:

```
> file <- system.file("sampleData", "data.txt", package="samplesizelogisticcasecontrol")
> file

[1] "/tmp/RtmpEcqBJ2/Rinst526d2f7f8840/samplesizelogisticcasecontrol/sampleData/data.txt"
```

Here the exposure variable is "Treatment", and "Gender_Male" is a dummy variable for the confounder gender. We will use the data from only the controls and define a new variable of interest which is the interaction of gender and treatment. In our model, both gender and treatment will be confounders. First, read in the data.

```
> data <- read.table(file, header=1, sep="\t")
```

Create the interaction variable

```
> data[, "Interaction"] <- data[, "Gender_Male"]*data[, "Treatment"]
> data[1:5, ]

  Casecontrol Gender_Male Treatment Interaction
1           0           0         0           0
2           1           1         0           0
3           0           0         0           0
4           1           1         1           1
5           1           1         0           0
```

Now subset the data to use only the controls

```
> temp  <- data[, "Casecontrol"] %in% 0
> data2 <- data[temp, ]
```

The data that gets passed in should only contain the columns that will be used in the analysis with the variable of interest being the last column.

```
> vars  <- c("Gender_Male", "Treatment", "Interaction")
> data2 <- data2[, vars]
```

Define the log-odds ratios for gender, treatment, and the interaction of gender and treatment. The order of these log-odds ratios must match the order of the columns in the data.

```
> logOR <- c(0.1, 0.13, 0.27)
```

Compute the sample sizes

```
> sampleSize_data(prev, logOR, data2)

$ss.wald.1
[1] 9402

$ss.wald.2
[1] 9404

$ss.score.1
[1] 9403

$ss.score.2
[1] 9400
```

Note that the same results can be obtained by not reading in the data and creating a new interaction variable, but by setting the input argument of data to be of type *file.list*.

```
> data.list <- list(file=file, header=1, sep="\t",
+             covars=c("Gender_Male", "Treatment"),
+             exposure=c("Gender_Male", "Treatment"))
> data.list$subsetData <- list(list(var="Casecontrol", operator="%in%", value=0))
> sampleSize_data(prev, logOR, data.list)

$ss.wald.1
[1] 9402

$ss.wald.2
[1] 9404

$ss.score.1
[1] 9403

$ss.score.2
[1] 9400
```

Power calculation using pilot data

Using the pilot data, estimate the log-odds ratios from a logistic regression:

```
> fit <- glm(Casecontrol ~ Gender_Male + Treatment + Interaction, data=data, family=binomi
> summary(fit)

Call:
glm(formula = Casecontrol ~ Gender_Male + Treatment + Interaction,
    family = binomial(), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8870  -0.8398  -0.8108   1.4989   1.5996

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.94369    0.13890  -6.794 1.09e-11 ***
Gender_Male  0.08297    0.19627   0.423    0.672
Treatment    0.21373    0.19447   1.099    0.272
Interaction -0.30629    0.27771  -1.103    0.270
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1213.1  on 999  degrees of freedom
Residual deviance: 1211.5  on 996  degrees of freedom
AIC: 1219.5

Number of Fisher Scoring iterations: 4
```

Extract the estimates needed

```
> coef     <- fit$coefficients
> logOR    <- coef[-1]
> logOR

Gender_Male   Treatment Interaction
 0.08296883  0.21372855 -0.30628697
```

Estimate the power assuming a study size of 15000 subjects with 10 percent of them cases.

```
> power_data(prev, logOR, data[, vars], sampleSize=15000, cc.ratio=0.1)

$pow.wald.1
[1] 0.8020756

$pow.wald.2
[1] 0.8010627
```

```
$pow.score.1
[1] 0.8035882

$pow.score.2
[1] 0.8045952
```

Session Information

```
> sessionInfo()

R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: CentOS Linux 7 (Core)

Matrix products: default
BLAS/LAPACK: /usr/local/intel/compilers_and_libraries_2020.2.254/linux/mkl/lib/intel64_lin

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] samplesizelogisticcasecontrol_2.0.0 mvtnorm_1.1-2

loaded via a namespace (and not attached):
[1] compiler_4.1.0 tools_4.1.0
```