How can I perform Exact Logistic Regression in R for data analysis?

Authored by

stats writer

June 29, 2024

Exact Logistic Regression is a statistical method used for analyzing categorical data in which the outcome variable is binary. It is an extension of the traditional logistic regression model that allows for exact probability calculations, rather than relying on approximations. In order to perform Exact Logistic Regression in R, the first step is to ensure that the necessary packages are installed and loaded. Then, the data must be prepared and cleaned to ensure that it is in the correct format for analysis. Once the data is ready, the model can be fitted and evaluated using appropriate statistical techniques. The results of the analysis can then be interpreted to gain insights into the relationship between the predictor variables and the outcome variable. Overall, Exact Logistic Regression in R is a powerful tool for data analysis and can provide valuable insights into categorical data.

Exact Logistic Regression | R Data Analysis Examples

Exact logistic regression is used to model binary outcome variables in which the log odds of the outcome is modeled as a linear combination of the predictor variables. It is used when the sample size is too small for a regular logistic regression (which uses the standard maximum-likelihood-based estimator) and/or when some of the cells formed by the outcome and categorical predictor variable have no observations. The estimates given by exact logistic regression do not depend on asymptotic results.

This page uses the following packages. Make sure that you can load them before trying to run the examples on this page. If you do not have a package installed, run: `install.packages("packagename")`, or if you see the version is out of date, run: `update.packages()`.

```
require(elrm)
```

```
Version info: Code for this page was tested in R version 3.0.1 (2013-05-16)
On: 2013-08-06
With: elrm 1.2.1; coda 0.16-1; lattice 0.20-15; knitr 1.3
```

Please note: The purpose of this page is to show how to use various data analysis commands. It does not cover all aspects of the research process which researchers are expected to do. In particular, it does not cover data cleaning and checking, verification of assumptions, model diagnostics or potential follow-up analyses.

Example of exact logistic regression

Suppose that we are interested in the factors that influence whether or not a high school senior is

admitted into a very competitive
engineering school. The outcome variable is binary (0/1): admit or not admit.
The predictor variables of interest include student gender and whether or not the
student took Advanced Placement calculus in high school. Because the response variable is binary, we need
to use a model that handles 0/1 outcome variables correctly. Also, because of the number of students involved is small, we will need a procedure that can perform the estimation with
a small sample size.

Description of the data

The data for this exact logistic data analysis include the number of students admitted, the total
number of applicants broken down by gender (the variable `female`), and whether or not
they had taken AP calculus (the variable `apcalc`). Since the dataset is so small,
we will read it in directly.

```
dat <- read.table(text = "
```

female apcalc admit num
0 0 0 7
0 0 1 1
0 1 0 3
0 1 1 7
1 0 0 5
1 0 1 1
1 1 0 0
1 1 1 6",
header = TRUE)

The `num` variable indicates frequency weight. We use this to expand the dataset and
then look at some frequency tables.

expand dataset by repeating each row num times and drop the num## variable
dat <- dat

look at various tablesxtabs(~female + apcalc, data = dat)

apcalc
female 0 1

```
## 0 8 10
## 1 6 6

xtabs(~female + admit, data = dat)

## admit
## female 0 1
## 0 10 8
## 1 5 7

xtabs(~apcalc + admit, data = dat)

## admit
## apcalc 0 1
## 0 12 2
## 1 3 13

xtabs(~female + apcalc + admit, data = dat)

## , , admit = 0
##
## apcalc
## female 0 1
```

```
## 0 7 3
## 1 5 0
##
## , , admit = 1
##
## apcalc
## female 0 1
## 0 1 7
## 1 1 6
```

The tables reveal that 30 students applied for the Engineering program. Of those, 15 were admitted and 15 were denied admission. There were 18 male and 12 female applicants. Sixteen of the applicants had taken AP calculus and 14 had not. Note that all of the females who took AP calculus were admitted, versus only 70% the males.

Analysis methods you might consider

Below is a list of some analysis methods you may have encountered. Some of the methods listed are quite reasonable, while others have

either fallen out of favor or have limitations.

(Approximate) Exact logistic regression

Let's run an (approximate) exact logistic analysis using the `elrm` command
in the `elrm` package. This is based on MCMC sampling. It requires a collapsed
data set with number of trials and number of successes, so we make that first.

```
x <- xtabs(~admit + interaction(female, apcalc), data = dat)
x # view cross tabs

## interaction(female, apcalc)
## admit 0.0 1.0 0.1 1.1
## 0 7 5 3 0
## 1 1 1 7 6

cdat <- cdat <- data.frame(female = rep(0:1, 2), apcalc = rep(0:1, each = 2),
admit = x, ntrials = colSums(x))
cdat # view collapsed data set
```

```
## female apcalc admit ntrials
## 0.0 0 0 1 8
## 1.0 1 0 1 6
## 0.1 0 1 7 10
## 1.1 1 1 6 6
```

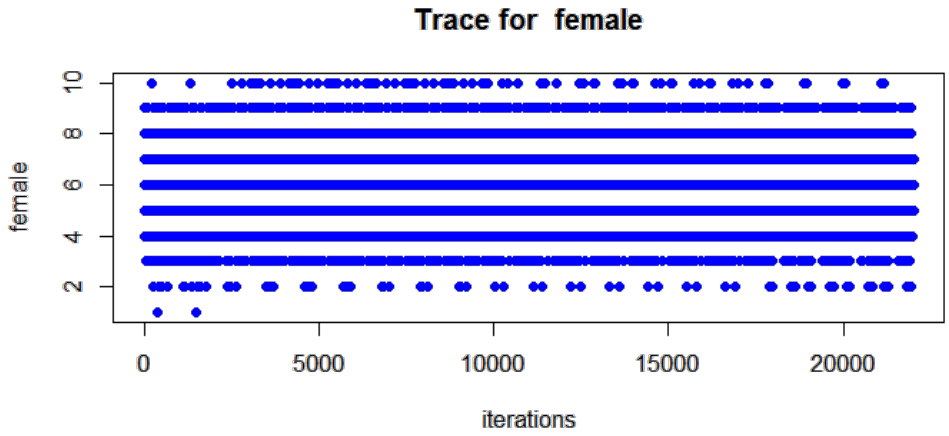
Now we can estimate the approximate logistic regression using `elrm` and
MCMC sampling. We will do 22,000 iterations with a 2,000 burnin for a
final chain of 20,000. Note that for the combined model of female and apcalc, we
use a chain of 5 million. This is because for inference, each effect needs at least
1,000, but because the conditional joint distribution is degenerate, for the female effect
the ratio of useable trials is low, meaning that to achieve over 1,000, the total iterations
must be extremely high.

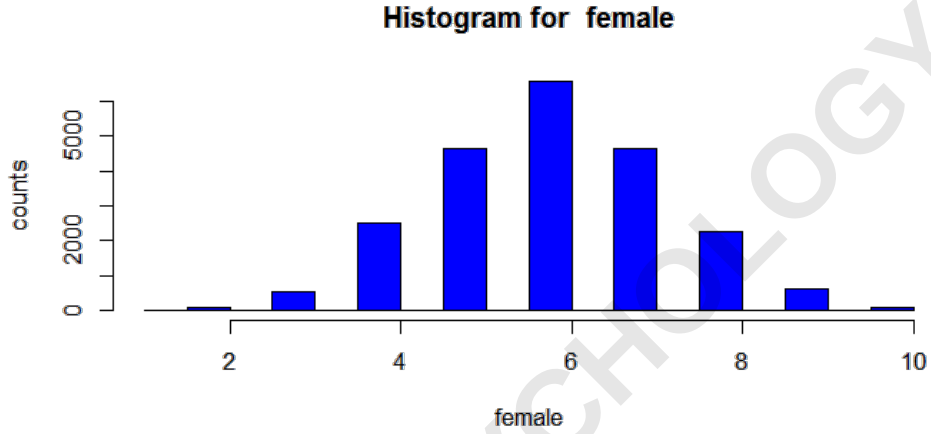
```
## model with female predictor only
m.female <- elrm(formula = admit/ntrials ~ female,
interest = ~female, iter = 22000,
dataset = cdat, burnIn = 2000)
```

```
## summary of model including estimates and CIssummary(m.female)

##
## Call:
## ]
## elrm(formula = admit/ntrials ~ female, interest = ~female, iter = 22000,
## dataset = cdat, burnIn = 2000)
##
##
## Results:
## estimate p-value p-value_se mc_size
## female 0.522 0.486 0.00428 20000
##
## 95% Confidence Intervals for Parameters
##
## lower upper
## female -1.13 2.31

## trace plot and histogram of sampled values from the sufficient## statisticplot(m.female)
```





model with apcalc predictor only

m.apcalc <- elrm(formula = admit/ntrials ~ apcalc, interest = ~apcalc, iter = 22000, dataset = cdat, burnIn = 2000)

Progress: 0% Progress: 5% Progress: 10% Progress: 15% Progress: 20% Progress: 25% Progress: 30% Progress: 35% Progress: 40% Progress: 45% Progress: 50% Progress: 55% Progress: 60% Progress: 65%

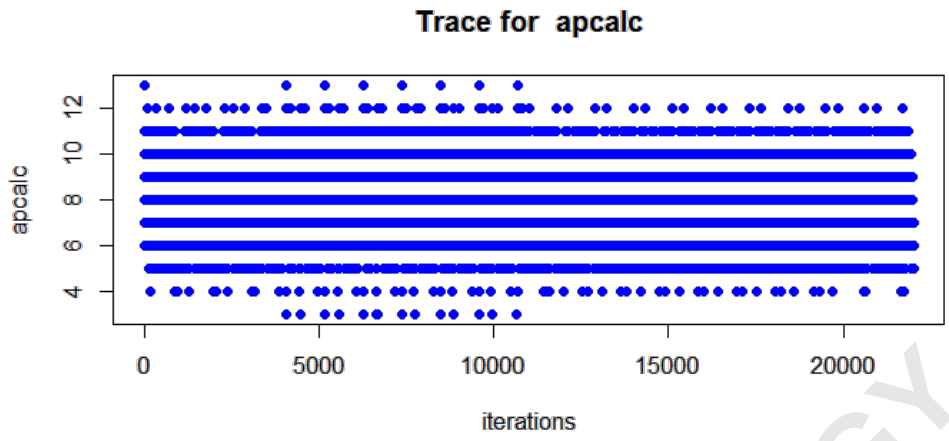
Progress: 70% Progress: 75% Progress: 80% Progress: 85% Progress: 90% Progress: 95% Progress: 100%

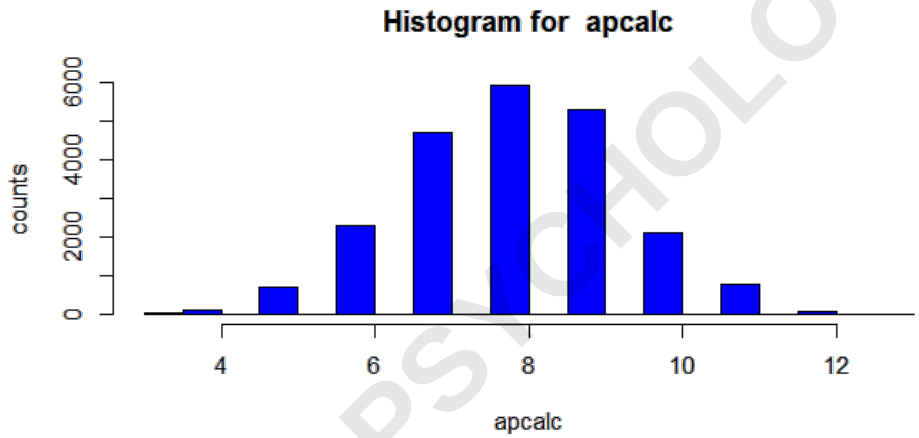
summary of model including estimates and CIssummary(m.apcalc)

##
Call:
]
elrm(formula = admit/ntrials ~ apcalc, interest = ~apcalc, iter = 22000,
dataset = cdat, burnIn = 2000)
##
##
Results:
estimate p-value p-value_se mc_size
apcalc 2.86 0.00035 0.00013 20000
##
95% Confidence Intervals for Parameters
##
lower upper
apcalc 1.08 Inf

trace plot and histogram of sampled values from the

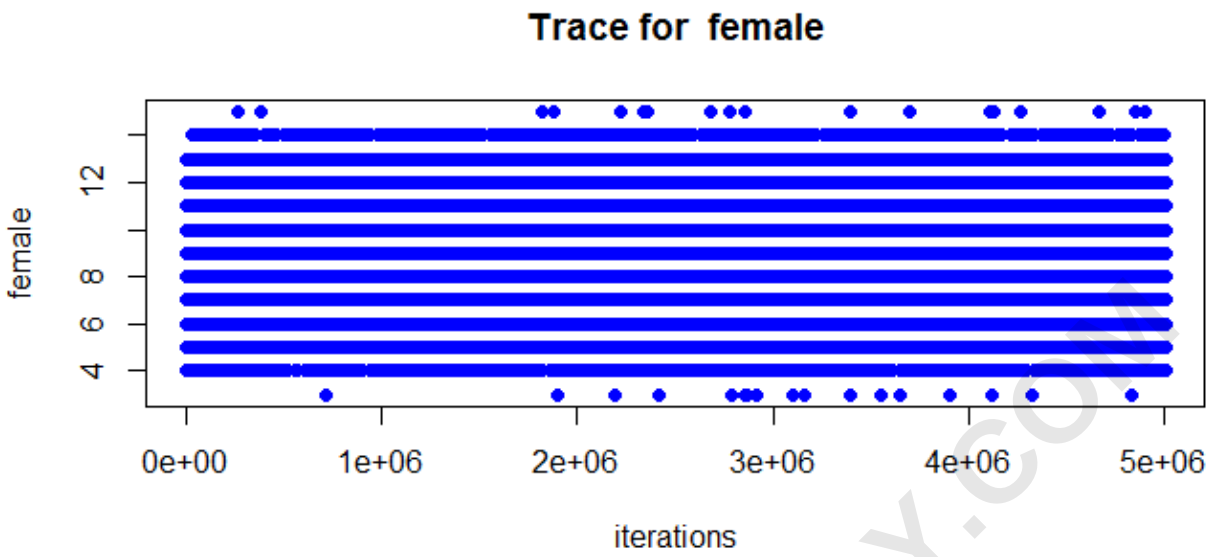
sufficient## statisticplot(m.apcalc)

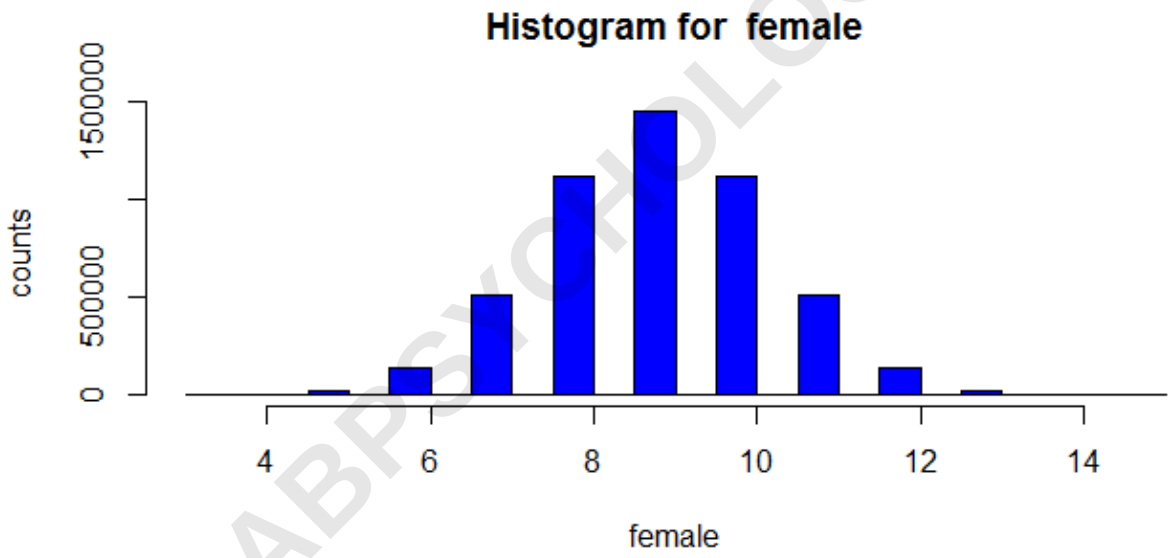


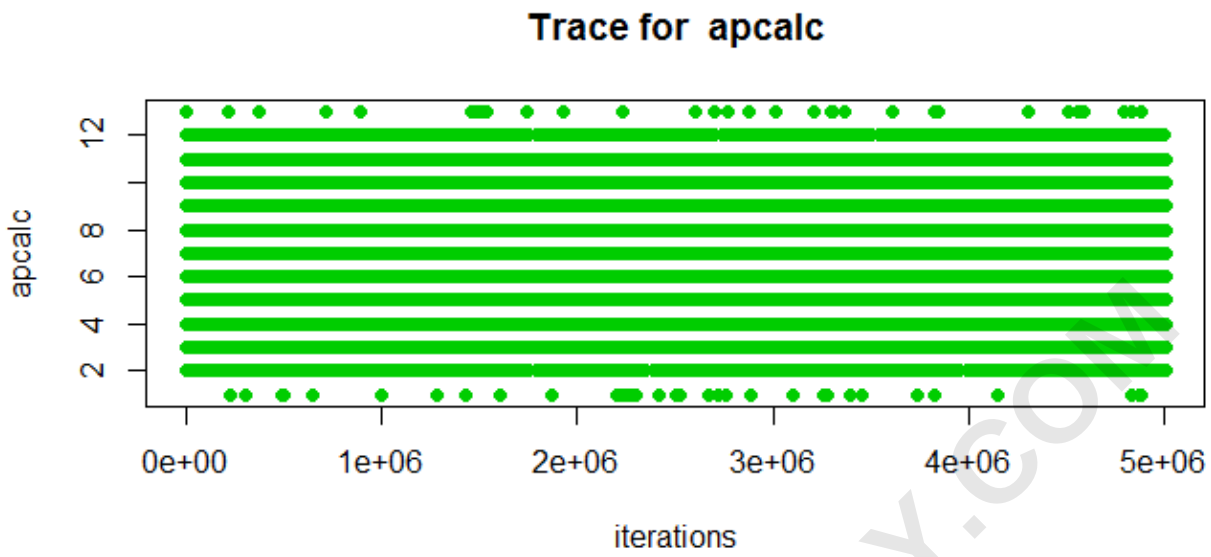


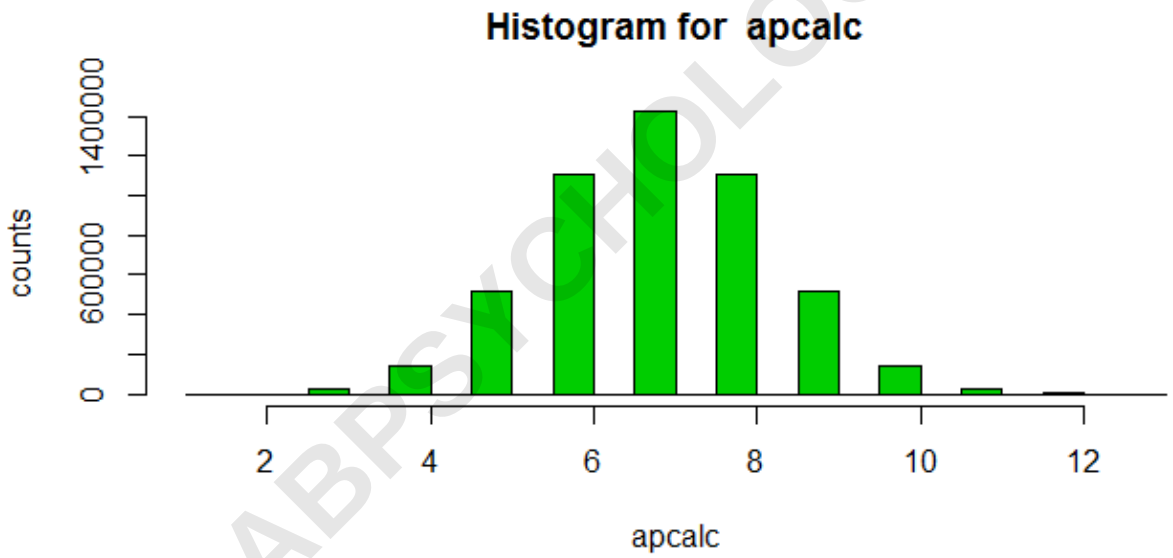
run not automated for time purposes

results









Note that this approximate technique with sufficient burnin and iterations is quite similar with the exact logistic estimates from Stata.

Things to consider

References