EXACT LOGISTIC REGRESSION | R DATA ANALYSIS EXAMPLES

From: <https://stats.idre.ucla.edu/r/dae/exact-logistic-regression/>

Note: This is to get a binary response either 0 or 1, not a value between 0 and 1. Do not use this for your insurance (Project 2 homework)

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 If you do not have a package installed, run: install.packages("packagename"), or if you see the version is out of date, run: update.packages().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[**rep**(1:**nrow**(dat), dat$num), -4]

*## look at various tables*

**xtabs**(~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.

* Exact logistic regression – This technique is appropriate because the outcome variable is binary, the sample size is small, and some cells are empty.
* Regular logistic regression – Due to the small sample size and the presence of cells with no subjects, regular logistic regression is not advisable, and it might not even be estimable.
* Two-way contingency tables – with Fisher’s exact test due to small expected values.

(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[2, ], 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 CIs*

**summary**(m.female)

##

## Call:

## [[1]]

## 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

##

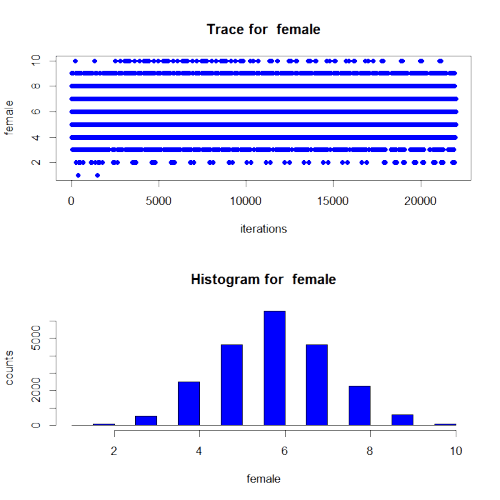
## lower upper

## female -1.13 2.31

*## trace plot and histogram of sampled values from the sufficient*

*## statistic*

**plot**(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 CIs*

**summary**(m.apcalc)

##

## Call:

## [[1]]

## 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

##

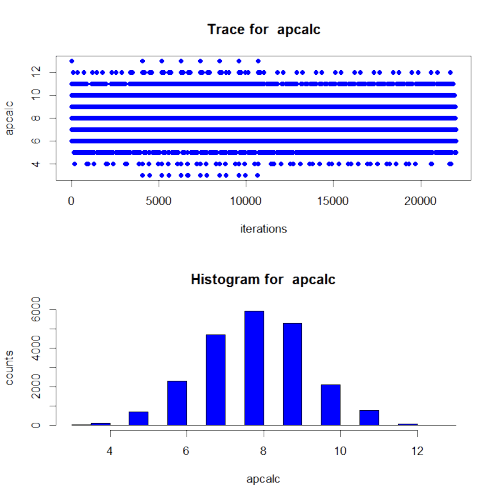
## lower upper

## apcalc 1.08 Inf

*## trace plot and histogram of sampled values from the sufficient*

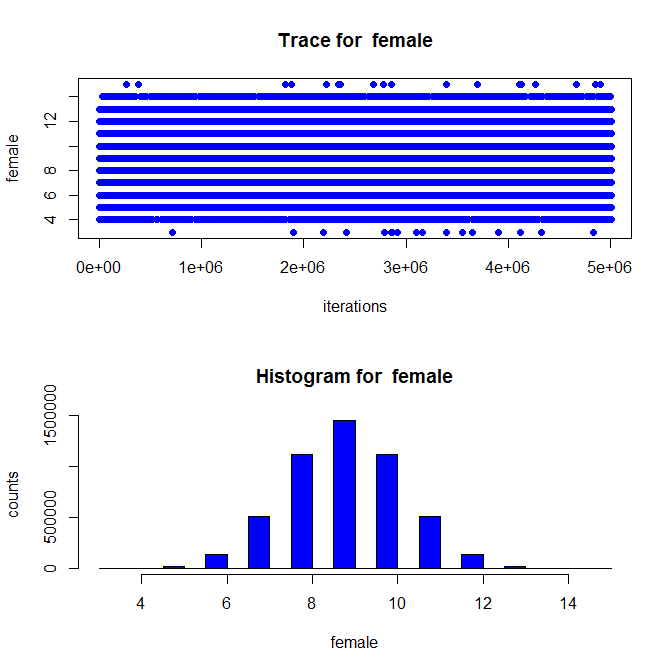
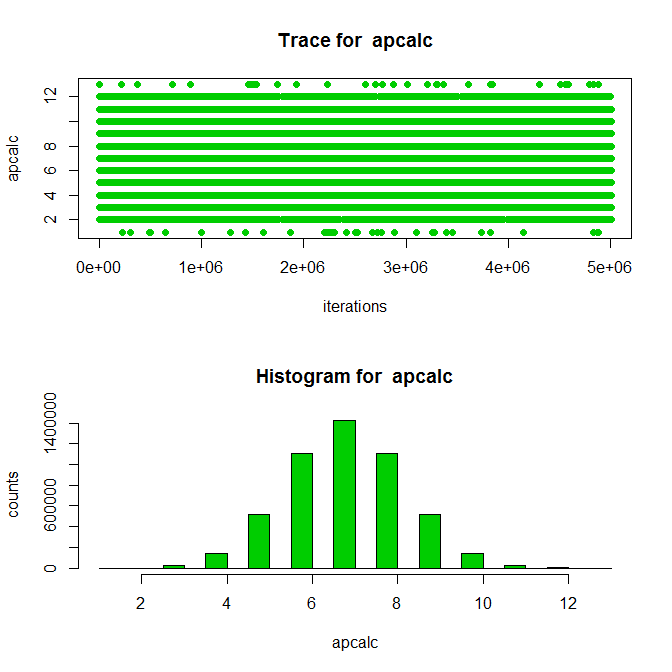
*## statistic*

**plot**(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](https://stats.idre.ucla.edu/stata/dae/exact-logistic-regression/) from Stata.

* The first part of the output echoes the call used to create the model. The results has three rows, one for the joint distribution, and one for each parameter.
* The estimate column has the parameter estimates (joint is NA). The next column gives the p-values. For the joint row, the p-value is that both effects are simultaneously zero. For female and apcalc, it is the p-value for testing that the individual parameter estimate is zero. Next is the Monte Carlo standard errors for the p-value. Finally, mc\_size is the length of the Markov chain of sampled values of sufficient statistics used for each parameter estimate. Note that all the length of the chain for the joint test was 5 million, a mere 1,739 for the female parameter. Because each Markov chain needs to be sufficiently long for stable inference, we needed to increase the total chain size to stably estimate female. For apcalc, the chain length is larger than necessary (over 1 million).
* below the main results are the confidence intervals for each parameter estimate.
* We can also plot each model, to obtain trace plots of the Markov chains.

Things to consider

* Firth logit may be helpful if you have separation in your data. This can be done in R using the logistf package.
* Exact logistic regression is an alternative to conditional logistic regression if you have stratification, since both condition on the number of positive outcomes within each stratum. The estimates from these two analyses will be different because a conditional logistic conditions only on the intercept term, exact logistic conditions on the sufficient statistics of the other regression parameters as well as the intercept term.

References

* Zamar, D., McNeney, B., and Graham, J. (2007). *elrm: Software Implementing Exact-like Inference for Logistic Regression Models*. <http://www.jstatsoft.org/v21/i03/paper>
* Collett, D. *Modeling Binary Data, Second Edition*. Boca Raton: Chapman and Hall.
* Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data, Second Edition*. Boca Raton: Chapman and Hall.
* Hirji, K. F. (2005). *Exact Analysis of Discrete Data*. Boca Raton: Chapman and Hall.