Package ltable 2.0.3. Part 3.

Ocheredko Oleksandr

Content:

FUNCTIONALITY

1. Constructs tables of counts and proportions out of data sets.
2. Inserts table into Excel and Word documents using clipboard, into LaTeX, HTML, Markdown and reStructuredText documents by the knitr::kable agency.
3. Molds table into acceptable for log-linear modeling data.frame, co.
4. Performs log-linear modeling.
5. Performs power analysis.

- This version is coded in R language exclusively to support across-systems portability.
- Log-linear and power analyses are enhanced with ability to model risks (rates) and relative risks (standardized ratios). Modelling survival data with interval censoring is also supported.

Modelling risks, relative risks, standardized ratios.

Every so often exposure can't be grouped without great loss of information, e.g., number of pills used by patient, time that passed after treatment, accumulated exposure to pesticide, person-years under treatment. Moreover, in such circumstances researcher is more inclined to model risks or rates that generate counts. Using par $offset$ one can model underlying risks. Offset is a variable from data set that contains size of exposure. Say, if there were 150 cases of cancer observed in population A of 10000 in 5 years, the exposure is 50000 person-years, so that rate is 3/1000. If we collect such data across different populations the question may arise whether the rate depends on provision with oncologists. So dependent variable is still counts of cancer cases observed in different populations, independent variable is provision with oncologists, while offset is exposure measured in person-years. Exposure covers different populations of different size and different periods of observation. This feature is auspicious for clinical trials data. Here we have personal records. If we measure exposure as number of pills taken by patient and response (outcome measure) is number of side effects under different regiments of treatment then dependent variable is number of side effects developed by patient, independent variable is regiment of treatment, and offset is number of pills.

Example 1.

Let's consider breast cancer rates in Iceland by year of birth (11 cohorts from 1840-1849 to 1940-1949) and by age (13 groups from 20-24 to 80-84 years), analysed by Breslow and Clayton (1993). Data is used also in BUGS Example *"Ice: non-parametric smoothing in an age-cohort model"* [1]. Data supplied with package and include variables age (Age group: 1-13), $year$ (Birth cohort: 1-11), $cases$ (Breast cancer counts), pyr (Person-years of risk). Let's model effects of age and year of birth on the rates of breast cancer with function $MCLogLin()$.

```
require(ltable)
data(BCdata)
res<-MCLogLin(cases~age+year, offset=pyr,
              draw=5000, data=BCdata)
```

File $offsetdata.rda$ includes 3 data sets (BCdata, SMdata, SimData), each used in examples with offset. The result is given in table below:

[1] BUGS. Examples Volume 2. https://www.mrc-bsu.cam.ac.uk/wp-content/uploads/WinBUGS_Vol2.pdf.

Effects	Estimate	Std.Error	z-score	Pr(>z)
(Intercept)	-1.110e+01	4.439e-01	2.500e+01	0.000e+00
age	3.388e-01	2.997e-02	1.130e+01	1.254e-29
year	1.984e-01	4.367e-02	4.543e+00	5.541e-06
phi	3.212e+00	8.740e-03	3.675e+02	0.000e+00

All effects are significant. ψ equals 3.212 that indicates overdispersion. Let's change ψ to value 0.01. It's not possible to do for user, but just for didactic purpose let's key up overdispersion. The output is as follows:

Effects	Estimate	Std.Error	z-score	Pr(>z)
(Intercept)	-1.022e+01	6.713e+00	1.523e+00	1.278e-01
age	4.427e-01	4.572e-01	9.684e-01	3.329e-01
year	2.819e-01	6.743e-01	4.180e-01	6.759e-01
phi	1.000e-02	1.807e-04	5.535e+01	0.000e+00

As one can see strong heterogeneity dissipates significance. Note, that ψ is parameter to be sampled with MCMC that evaluates heterogeneity. I just displayed what influence heterogeneity renders to statistical tests of regression effects. Now, let's change ψ to value 100. It means it is not sampled any more and is taken as it put. Here what we have:

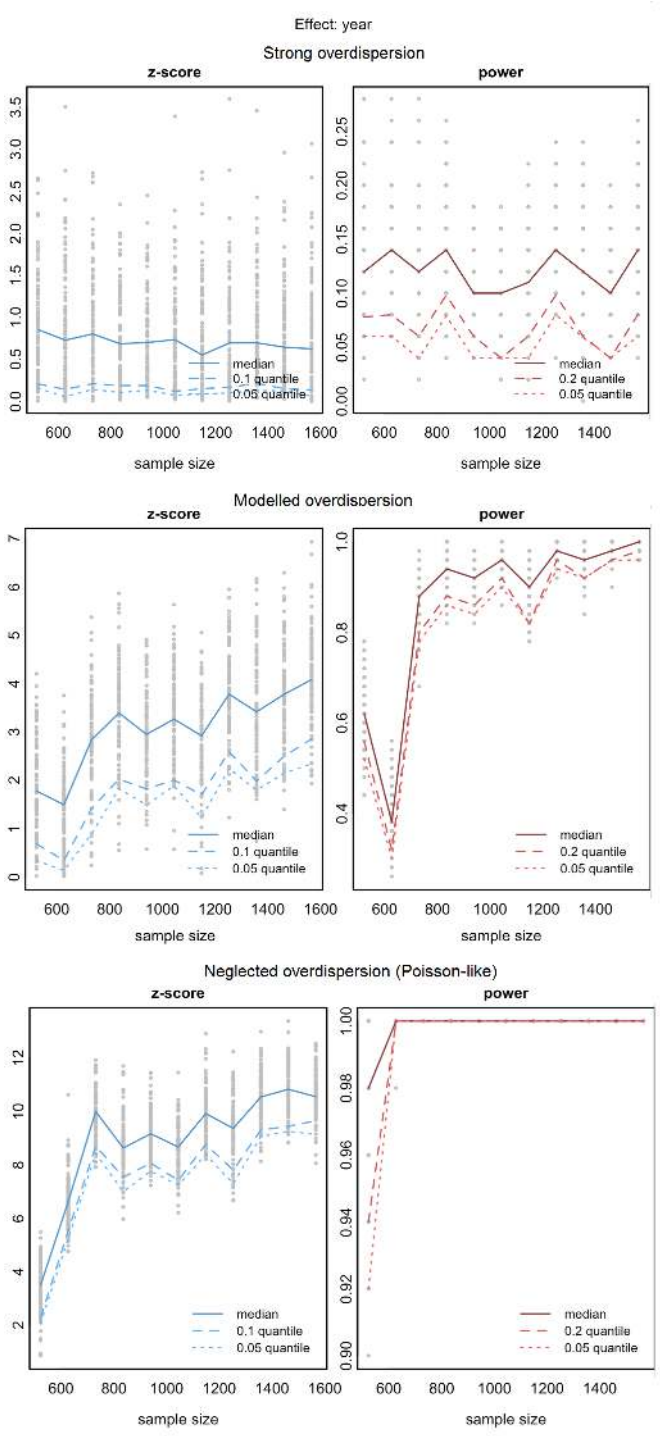
Effects	Estimate	Std.Error	z-score	Pr(>z)
(Intercept)	-1.079e+01	2.101e-01	5.134e+01	0.000e+00
age	3.060e-01	1.316e-02	2.326e+01	0.000e+00
year	1.907e-01	1.990e-02	9.583e+00	9.418e-22
phi	1.000e+02	8.500e-02	1.177e+03	0.000e+00

This acts to the opposite effect. Neglect of present overdispersion leads to overoptimistic tests of regression effects. This is exactly what poisson model does. Let's compare with glm $\{stats\}$ modelling that uses poisson distribution:

```
glm(formula = cases ~ age + year + offset(log(pyr)), family = poisson,cdata = BCdata)
```

Effects	Estimate	Std.Error	z-score	Pr(>z)
(Intercept)	-10.80955	0.19519	-55.38	<2e-16 ***
age	0.30637	0.01194	25.66	<2e-16 ***
year	0.19490	0.01838	10.60	<2e-16 ***

The output almost the same to negligible differences of algorithms used, MLE in case of $glm()$ and MCMC in case of $MCLogLin()$. It's important to note that fit of the model and residuals are remarkably better while sampling ψ as parameter.



To support deductions with power curves I demonstrate power analysis of this data with 3 values of ψ which are 0.01 (strong overdispersion simulated), sampled by MCMC as model parameter, and 100 (overdispersion neglected, poisson like approach). Call is the same:

```
res<-MCPower(cases~age+year, offset=pyr, draw=5000, burnin=1000, effect="year",
scale_min=0.4, scale_max=1.2, data=BCdata)
```

but to model extremes in 2 scenarios I put 0.01 and 100 as ψ values and doing so prevented this par from sampling.
Power curves with ψ={0.01, modelled, 100} displayed above.
Exemplary power curves support the fact, that overdispersed data are more required as to sample size.

In example independent variables are of numeric class. If one needs to inspect non-linear relationships, class should be changed either to unordered factor class to examine effect of each birth cohort and age group, or to ordered factor to elicit non-linear relationships as quadratic, cubic, and higher order polynomials.

Conclusion: to be on the safe side *better use NB2* both for modelling regression effects and power curve.

Example 2.

It's ubiquity in medicine, health administration, epidemiology to use standardized indexes. For example, effectiveness of hospital treatment in US and Canada is assessed by comparison of observed number of lethal cases against expected. Latter are calculated by applying logistic regression with coefficients that describe nationwide technology of treatment. Independent variables are comorbidity, stage of disease in question, gender, and age of patient. Having each treated patient's characteristics logistic regression produces risk of passing on for each patient given nationwide technology. Pulling the expected individual risks together they have expected number of deceased for each department/hospital/union. If observed number of lethal cases significantly less than expected, the technology of treatment applied in given unit is better than nationwide. Of course, if one interests in conducive factors it's possible to model their influence using offset. This example is based on counts of patient deaths following heart transplant surgery in 131 hospitals in the US between October 1987 and December 1989. These were analysed by Christiansen and Morris (1996, 1997)[2]. Data is also analysed by Peter D. Congdon [3]. There are two variables only: Number of Deaths (*y*) and Number of Expected Deaths (*o*) across 131 hospitals. The emphasis is maid to demonstrate functionality to cope with standardized ratios. Here is the call to function $MCLogLin()$ with excerpts of output:

Call: MCLogLin(formula = y ~ 1, data = SMdata, offset = o, DIC = TRUE, draw = 1500, burnin = 500)

[2]Christiansen C, Morris C (1996) Fitting and checking a two-level Poisson model: modeling patient mortality rates in heart transplant patients, pp 467-501, in Bayesian Biostatistics, eds D Berry, D Stangl. Marcel Dekker, New York.

[3]Peter D. Congdon. Bayesian Hierarchical Models With Applications Using R (2020) Second Edition. Example 4.5 Hospital Mortality, p.125-26.

Effects	Estimate	Std.Error	z-score	Pr(>z)
(Intercept)	4.382e-02	7.049e-02	6.216e-01	5.342e-01
phi	7.262e+00	2.646e-02	2.745e+02	0.000e+00

DIC related components:
DIC = 464.7117
pD = 1.836479
meanDev = 462.8752
Devmean = 461.0387

Linear predictor includes intercept only which is correctly assessed to be close to 1 (1.0448) upon exponentiation. Peter D. Congdon tried Poisson-gamma mixture to model the data with JAGS (package *jagsUI*). The DIC reported for this model is 475 which is close.