Package ltable 2.0.3. Part 4.

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 interval censored survival data

In the setting of survival analysis, interval censored data occur when an event time is known only up to an interval. It covers majority of situations with mixed case censoring, that can include left censored, right censored, uncensored and observations that are censored but neither right nor left censored. The last type of censoring can occur if a subject is regularly inspected and all that is known is that the event of interest occurred between check-ups. The standard assumption is that this observation time is independent of the event of interest, although the observation time may be random or fixed by design. A classic example of mixed case interval censored datasets is retrospective study presented by Klein and Moeschberger (1997) [1]. Study was carried out to compare the cosmetic effects of radiotherapy alone versus radiotherapy and adjuvant chemotherapy on women with early breast cancer. To compare the two treatments, a retrospective study of 46 radiation only and 48 radiation plus chemotherapy patients was conducted. Patients was observed initially every 4-6 months, but, as their recovery progressed, the interval between visits lengthened. The event of interest was the time to first appearance of moderate or severe breast retraction. As the patients were observed only at some random times, the exact time of breast retraction is known only to fall within the interval between visits. Data is retrievable from package *interval* (data *bcdeter*). I described the study to put user in a picture of real world research setups.

To see how package *ltable* deals with such setups I simulated interval censored survival data. First I used Weibull r.n. generator to sample 50 values: 10 with logscale 1.5+1, 10 with logscale 1.5, 10 with logscale 1 and 20 with logscale 0. 1.5 and 1 are regression effects of exposures T (treatment) and C (comorbidity free status). Shape=1.5 in all groups. Generated values fall in a range from 0.2731 to 26.8083. It's verifiable given seed=1966. Let's assume generated values are months. Afterward generated values transformed to interval censored with year interval width and data grouped with $table\_f()$ and $tableToData()$ functions. Indicator variables are created for years to estimate baseline hazard rates h0(year). Finally offset variable is calculated as person-years of survival for each profile, that is, for each row of the final table. The code is following:

```
require(ltable)
set.seed(1966)
shape<-1.5
scale11<-exp(1.5+1)
scale10<-exp(1.5)
scale01<-exp(1)
scale00<-exp(0)
simData1<-data.frame(time=rweibull(n=10, shape=shape, scale = scale11), T=1, C=1)
simData2<-data.frame(time=rweibull(n=10, shape=shape, scale = scale10), T=1, C=0)
simData3<-data.frame(time=rweibull(n=10, shape=shape, scale = scale01), T=0, C=1)
simData4<-data.frame(time=rweibull(n=20, shape=shape, scale = scale00), T=0, C=0)
simData<-rbind(simData1,rbind(simData2, (rbind(simData3,simData4))))
```

[1] KLEIN, J. P, MOESCHBERGER, M. Survival Analysis. New York: Springer Verlag, 1997.

```
simData$Year<-round(simData$time/12)+1
simGroupData<-simData[,-1]
tab<-table_f(simGroupData, "T,C,Year")
```

```
tab_p<-tableToData(tab)
tab_s<-tab_p[tab_p$Counts>0,]
tab_s$Year2<-ifelse(tab_s$Year>=2,1,0)
tab_s$Year3<-ifelse(tab_s$Year>=3,1,0)
tab_s$offset<-c(rep(50,4),rep(10*2,2),6*3)
tab_s
```

```
   T C Year Counts Year2 Year3 offset
1  0 0    1     20     0     0     50
2  0 1    1     10     0     0     50
3  1 0    1      9     0     0     50
4  1 1    1      1     0     0     50
7  1 0    2      1     1     0     20
8  1 1    2      3     1     0     20
12 1 1    3      6     1     1     18
```

Data supplied with file $offsetdata.rda$ (data *SimData*). The call to function $MCLogLin()$ is as follows:

```
require(ltable)
data(SimData)
res<-MCLogLin(formula = Counts ~ Year2 + Year3 + T + C,
              data = SimData, offset = offset, draw = 5000)
```

```
Call:
MCLogLin(formula = Counts ~ Year2 + Year3 + T + C, data = SimData,
    offset = offset, draw = 5000)


Coefficients:
                              Estimate  Std.Error  |z-score|   Pr(>|z|)
(Intercept)                 -5.100e-01  5.028e-01  1.014e+00  3.104e-01
Year2                        1.160e-01  8.249e-01  1.406e-01  8.882e-01
Year3                        1.906e+00  9.589e-01  1.988e+00  4.681e-02
T1                          -1.305e+00  6.741e-01  1.936e+00  5.283e-02
C1                          -7.537e-01  5.922e-01  1.273e+00  2.031e-01
phi                          3.559e+00  1.946e-01  1.829e+01  9.563e-75
```

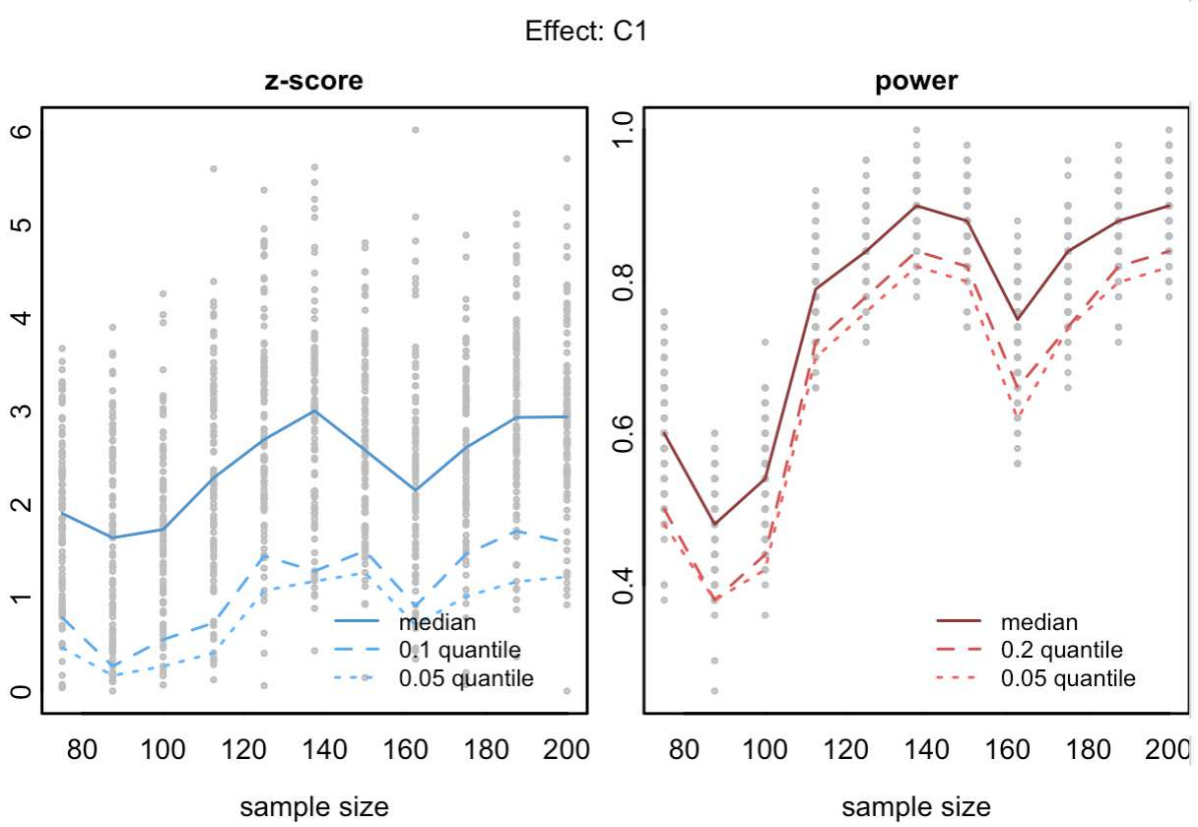
```
Model fit:
MCMC fitting
Samplers : Gibbs for expected counts, Slice for regr. coeff. and inv.var.par. phi
Language: R
Jacobian reciprocal condition number =  0.1203108
chisq/n =  0.09574534
Deviance=  0.0007742469
NULL Deviance=  0.400395
Log.likelihood=  -18.10798
AIC(1) =  46.21597
AIC(n) =  6.602281
BIC =  45.94552


Residuals report ( R is individual risk):
Row Ovserved R Predicted R Raw Residual Pearson Residual Anscombe Residual
 1    0.40000     0.40074     -0.00074          -0.0011          -0.0020
 2    0.20000     0.20096     -0.00096          -0.0021          -0.0033
 3    0.18000     0.16271      0.01729           0.0419           0.0647
 4    0.02000     0.03687     -0.01687          -0.0874          -0.1254
 5    0.05000     0.08745     -0.03745          -0.1251          -0.1931
 6    0.15000     0.11070      0.03930           0.1163           0.1673
 7    0.33333     0.32762      0.00571           0.0095           0.0164
```

From the output we can deduce that covariance matrix of model parameters is not stable enough and sensitive due to the paucity of profiles. It engenders problems discussed below. Fit to the data is good as chi-square test is less than 1 per degree of freedom with actual value of 0.09. That is supported by residuals report.

Next, regression effects are of expected directions and magnitudes. Weibull model regression effects of variables T and C are positive (increase survival), that correspond to negative effects in NB2 model (both variables reduce the risk of event). Magnitude of T effect is larger than that of C, that also agrees with true generation scenario.
Intercept is confluent with baseline hazard rates for first year. Effects Year2 and Year3 depict augmented baseline hazards in these years against previous. Therefore one can see that each consecutive year baseline hazard grows.
All effects are not significant. Let's consider power analysis to elicit whether it's due to insufficient sample size or the model just can't substantiate underlying mechanism. Let's do power analysis for effect of C, using scale_min=1.5, scale_max=4:

```
require(ltable)
load("offsetdata.rda")
res<-MCPower(formula = Counts ~ Year2 + Year3 + T + C, effect="C1",
data = SimData, offset = offset, draw = 5000, scale_min=1.5, scale_max=4)
plot(res, stencil=3)
```



Note, that we put effect quoted with the name that appears in design matrix and output of $MCLogLin()$ by adding 1 to C to show that it is contrast of C=1 against C=0.

With the growth of sample size the regression effect of C obviously gains significance. The irregularity of curves explained by over-sensitivity of covariance matrix to data with Jacobian reciprocal condition number less than 1.

So power analysis helps to illustrate the makings of NB2 to reveal and substantiate true data generation mechanism of survival data.

Joint hypotheses testing

In most studies there is a need to test several dependent hypotheses. Dependency may be structural or sequential. Log-linear model can be put to the task with tabulated data. Generally speaking if there is a system of equations modelling several outcomes *y_1...y_K* :

$$y_1 = f_1(y_{set_1}, X_{set_1}) \tag{1}$$
$$y_2 = f_2(y_{set_2}, X_{set_2}) \tag{2}$$
$$... \tag{3}$$
$$y_K = f_K(y_{set_K}, X_{set_K}) \tag{4}$$

and data is of tabulated format we can apply log-linear model to test several related hypotheses. Let's have two dependent hypotheses based on *Titanic* data. First concerns the checking of policy "safety to child and woman" which can be formulated as class of passengers accommodation regressed on age and gender. Another related hypothesis is that probability of survival depends on class, age, and their combination (second order effect). To test these hypotheses jointly we include all relevant effects into linear predictor. Code is as given:

```
require(ltable)
TitanicData <- as.data.frame(datasets::Titanic)
names(TitanicData)[5] <- "Counts"
TitanicData$Class <- factor(TitanicData$Class,
                            ordered=FALSE)
set.seed(1966)
res<-MCLogLin(formula = Counts ~ Class * Age + Class * Sex + Survived *
    Class * Age, data = TitanicData)
```

```
Call:
MCLogLin(formula = Counts ~ Class * Age + Class * Sex + Survived *
    Class * Age, data = TitanicData)


Coefficients:
                              Estimate  Std.Error  |z-score|   Pr(>|z|)
(Intercept)                -1.348e+01  7.589e+00  1.776e+00  7.580e-02
Class2nd                   -6.185e+00  1.364e+01  4.534e-01  6.503e-01
Class3rd                    1.739e+01  7.620e+00  2.282e+00  2.250e-02
ClassCrew                  -4.860e+01  1.705e+01  2.850e+00  4.370e-03
```

```
AgeAdult                          1.798e+01  7.524e+00  2.389e+00  1.688e-02
SexFemale                        -1.181e+00  1.359e+00  8.685e-01  3.851e-01
SurvivedYes                       1.532e+01  7.564e+00  2.026e+00  4.279e-02
Class2nd:AgeAdult                 6.522e+00  1.361e+01  4.791e-01  6.319e-01
Class3rd:AgeAdult                -1.589e+01  7.592e+00  2.094e+00  3.630e-02
ClassCrew:AgeAdult                5.062e+01  1.715e+01  2.952e+00  3.153e-03
Class2nd:SexFemale                1.023e+00  1.844e+00  5.547e-01  5.791e-01
Class3rd:SexFemale                6.154e-01  1.624e+00  3.789e-01  7.047e-01
ClassCrew:SexFemale              -2.709e+00  2.089e+00  1.296e+00  1.948e-01
Class2nd:SurvivedYes              7.323e+00  1.367e+01  5.358e-01  5.921e-01
Class3rd:SurvivedYes             -1.596e+01  7.650e+00  2.087e+00  3.691e-02
ClassCrew:SurvivedYes            -7.716e+01  2.681e+01  2.878e+00  4.008e-03
AgeAdult:SurvivedYes             -1.386e+01  7.635e+00  1.816e+00  6.940e-02
Class2nd:AgeAdult:SurvivedYes    -9.235e+00  1.374e+01  6.723e-01  5.014e-01
Class3rd:AgeAdult:SurvivedYes     1.357e+01  7.808e+00  1.737e+00  8.230e-02
ClassCrew:AgeAdult:SurvivedYes    7.599e+01  2.675e+01  2.841e+00  4.503e-03
phi                               1.007e+00  7.319e-02  1.375e+01  4.795e-43


Model fit:
MCMC fitting
Samplers : Gibbs for expected counts, Slice for regr. coeff. and inv.var.par. phi
Language: R
Jacobian reciprocal condition number =  0.001705213
chisq/n =  0.002269504
Deviance=  0.1260396
NULL Deviance=  2.510553
Log.likelihood=  -118.1038
AIC(1) =  276.2076
AIC(n) =  8.631486
BIC =  305.5223


Residuals report (Y denotes Counts):
Row Ovserved Y Predicted Y Raw Residual Pearson Residual Anscombe Residual
 1          0       0.009       -0.009           -0.093           -0.160
 2          0       0.001       -0.001           -0.035           -0.056
 3         35      34.893        0.107            0.003            0.020
 4          0       0.000       -0.000           -0.000           -0.000
 5          0       0.006       -0.006           -0.075           -0.126
 6          0       0.002       -0.002           -0.047           -0.077
```

```
 7         17      17.096       -0.096          -0.005          -0.028
 8          0       0.000       -0.000          -0.000          -0.000
 9        118     117.051        0.949           0.008           0.079
10        154     153.304        0.696           0.005           0.049
11        387     386.438        0.562           0.001           0.021
12        670     669.435        0.565           0.001           0.015
13          4       4.648       -0.648          -0.127          -0.447
14         13      13.755       -0.755          -0.053          -0.260
15         89      89.331       -0.331          -0.004          -0.033
16          3       3.544       -0.544          -0.136          -0.442
17          5       4.912        0.088           0.016           0.057
18         11      11.073       -0.073          -0.006          -0.029
19         13      13.305       -0.305          -0.022          -0.107
20          0       0.000       -0.000          -0.000          -0.000
21          1       1.183       -0.183          -0.114          -0.274
22         13      12.822        0.178           0.013           0.063
23         14      13.677        0.323           0.023           0.110
24          0       0.000       -0.000          -0.000          -0.000
25         57      57.748       -0.748          -0.013          -0.100
26         14      14.634       -0.634          -0.042          -0.209
27         75      75.320       -0.320          -0.004          -0.036
28        192     192.549       -0.549          -0.003          -0.033
29        140     139.123        0.877           0.006           0.065
30         80      79.351        0.649           0.008           0.070
31         76      75.802        0.198           0.003           0.022
32         20      19.451        0.549           0.028           0.149

Warning in MCLogLin(formula = Counts ~ Class * Age + Class * Sex + Survived * :
MCMC based errors are used
```

Based on output we can accept both hypotheses except first on part of gender.