UDC 37.013:535.328:004
DOI https://doi.org/10.32782/2786-7684/2024-2-27

Ocheredko Oleksandr Mykolayovich,
Doctor of Medical Sciences, Professor,
Head of the Department of Social Medicine
and Administration of Medical Services,
Vinnytsia National Pirogov Medical University
ORCID ID: 0000-0002-4792-8581
Vinnytsia, Ukraine

Rudenko Anastasiia Abdukarimivna,
PhD, Associate Professor at the Department of Social Medicine
and Health Services Administration Department,
Vinnytsia National Pirogov Medical University
ORCID ID: 0000-0003-0444-1869
Vinnytsia, Ukraine

Modelling interval over-dispersed censored survival data

Introduction. Unknown exact timing of events and quite possible heterogeneity of counts distributions across time intervals together with profile insufficiencies pose challenges to data modelling. it calls either for mixture of latent variables distributions or generalised distributions that accommodate heterogeneity. Our goal was to examine capacity of Negative Binomial Type 2 (NB2) regression to deal with presented challenges. Methods and data. We suggested application of NB2 regression with support of power analysis implemented in R package «ltable». We examined efficacy by simulated example to demonstrate the capacity to unveil the data generation mechanism. Results confirmed that counts strongly over-dispersed which is usual situation. Covariance matrix of model parameters was not stable enough and sensitive due to the paucity of profiles that is also typical. Fit to the data was good as chi-square test is less than 1 per degree of freedom with actual value of 0.09. That was supported by residuals report. Regression effects were of expected directions and magnitudes and revealed data generation mechanism. Power analysis validated the output and substantiated true data generation mechanism of interval censored survival data. We suggest that tentative example supports the effectiveness of modelling. In its turn power analysis helps to validate the output and to reveal and substantiate true data generation mechanism of interval censored survival data. Conclusions. Modelling interval over-dispersed censored survival data is still a challenge due to heterogeneity and profile shortage that underpowers hypotheses tests. We suggest NB2 based regression to process such data with support of power analysis. Simulated data analysis supports the effectiveness of modelling.

Key words: over-dispersion, interval censoring, survival data, NB2 regression.

Очередько Олександр Миколайович, доктор медичних наук, професор, завідувач кафедри соціальної медицини та організації охорони здоров'я, Вінницький національний медичний університет імені М. І. Пирогова, ORCID ID: 0000-0002-4792-8581, м. Вінниця, Україна

Руденко Анастасія Абдукарімівна, PhD, доцент кафедри соціальної медицини та організації охорони здоров'я, Вінницький національний медичний університет імені М. І. Пирогова ORCID ID: 0000-0003-0444-1869, м. Вінниця, Україна

Моделювання інтервально-цензурованих надмірно дисперсійних даних процесу виживання

Вступ. Невідомий точний час подій і можлива неоднорідність розподілу частот подій за часовими інтервалами разом із недоліками профілю створюють проблеми для моделювання даних. Це вимагає або міксту розподілів латентних змінних, або узагальнених розподілів, які враховують неоднорідність. Нашою **метою** було вивчити здатність негативної біноміальної регресії типу 2 (NB2) справлятися з представленими проблемами.

Методи та дані. Ми запропонували застосування регресії NB2 із підтримкою аналізу потужності, реалізованого в пакеті R «ltable». Ми перевірили ефективність на штучно змодельованих даних, щоб продемонструвати здатність до розкриття механізму генерації даних.

Результати підтвердили, що частоти гетерогенні, що є звичайною ситуацією. Коваріаційна матриця параметрів моделі була недостатньо стабільною та чутливою через брак профілів, що також характерно. Відповідність даним була гарною, оскільки критерій хі-квадрат був менше 1 на ступінь свободи з фактичним значенням 0,09. Це було підтверджено аналізом залишків моделі. Ефекти регресії мали очікувані напрямки та величини та розкрили механізм генерації даних. Аналіз потужності підтвердив результати моделювання та підтримав справжній механізм генерації інтервально цензурованих даних. Ми припускаємо, що даний приклад підтверджує ефективність моделювання. У свою чергу, аналіз потужності допомагає підтвердити вихідні дані, виявити, та обґрунтувати справжній механізм генерації інтервальних цензурованих даних.

Висновки. Моделювання інтервальних цензурованих даних з гетерогенністю все ще є проблемою через наддисперсність і дефіцит профілів, що зменшують потужність тестування гіпотез. Ми пропонуємо NB2 регресію за підтримки аналізу потужності для аналізу таких даних. Результати аналізу штучно згенерованих даних підтримує ефективність моделювання.

Ключові слова: надмірна дисперсія, інтервальне цензурування, дані про виживання, NB2 регресія.

Introduction. Modelling interval censored survival data requires complex algorithms with latent variables and is not readily obtainable. Even more complexity added up with over-dispersed interval censored survival data for it calls either for mixture of latent variables distributions or generalised distributions that accommodate heterogeneity. The latter approach pursued in the paper. 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. Lengthening visits is one of typical source of heterogeneity in event counts per interval. We described the study to put reader in a picture of real world research setups.

Methodology and methods. Given unknown exact timing of events and quite possible heterogeneity of counts distributions across intervals we applied Negative Binomial Type 2 (NB2) distribution that effectively accommodates heterogeneity. Due to complexity of the model estimation is processed with MCMC Gibbs&Slice samplers. Algorithm was developed and implemented in R package «ltable» by Ocheredko Oleksandr [2].

To see how package «ltable» deals with aforementioned setup we simulated interval censored survival data. First we 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 of package «ltable». Indicator variables are created for years (Year2 and Year3) 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 R code is following (Fig.1):

Data structure with content displayed in Fig.2. Profiles are composed by variables T, C, Year. Data supplied with package «ltable» (Simdata) and processed with function MCLogLin() of «ltable». Call to function MCLogLin() is also displayed. Offset variable supplied to argument offset, in given situation describing person-years of risk.

Results. Results are given in Table 1 and Figure 3.

From the output we can deduce that counts are strongly over-dispersed (heterogeneity coefficient $\psi < 1$) [3]. 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 [4]. That is supported by residuals report.

Next, regression effects are of expected directions and magnitudes. Weibull model regression effects of variables

```
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))))
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
```

Fig. 1. R code of interval censored heterogenous data generation

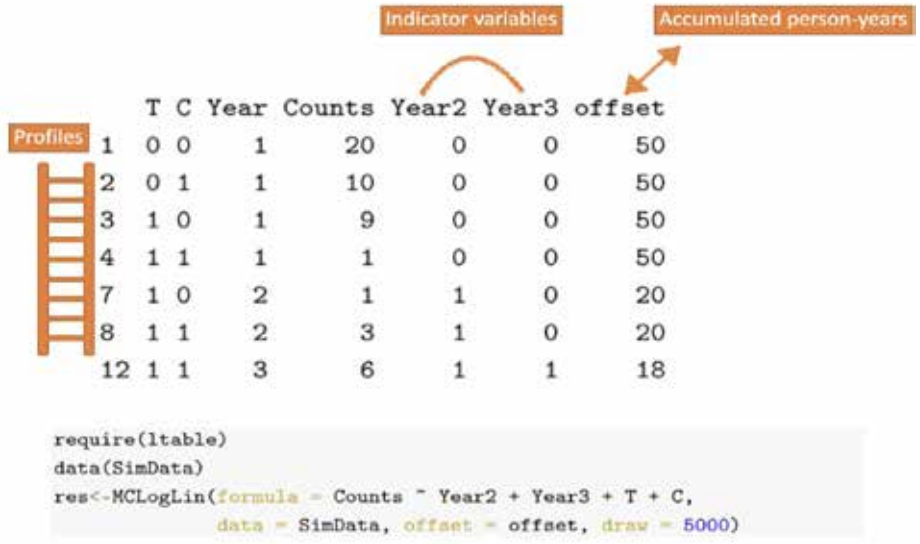

Fig. 2. Data structure and processing with MCLogLin function call

Table 1
MCLogLin coefficients output, SimData

Coefficients	Estimate	Error	Z value	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
ψ	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
```

Fig. 3. Model fit to synthetic data

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 sensible by sign and magnitude but are not significant. Given obvious over-dispersion and small sample size we can question test validity.

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 using function MCPower() of package «ltable», call is following:

resP <- MCPower(formula = Counts ~ Year2 + Year3 + T + C, effect="C1", data = SimData, offset = offset, draw = 5000, scale_min=1.5, scale_max=4)

plot(resP, stencil=3)

We abstained from printing results, power curves suffice (Fig.4).

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.

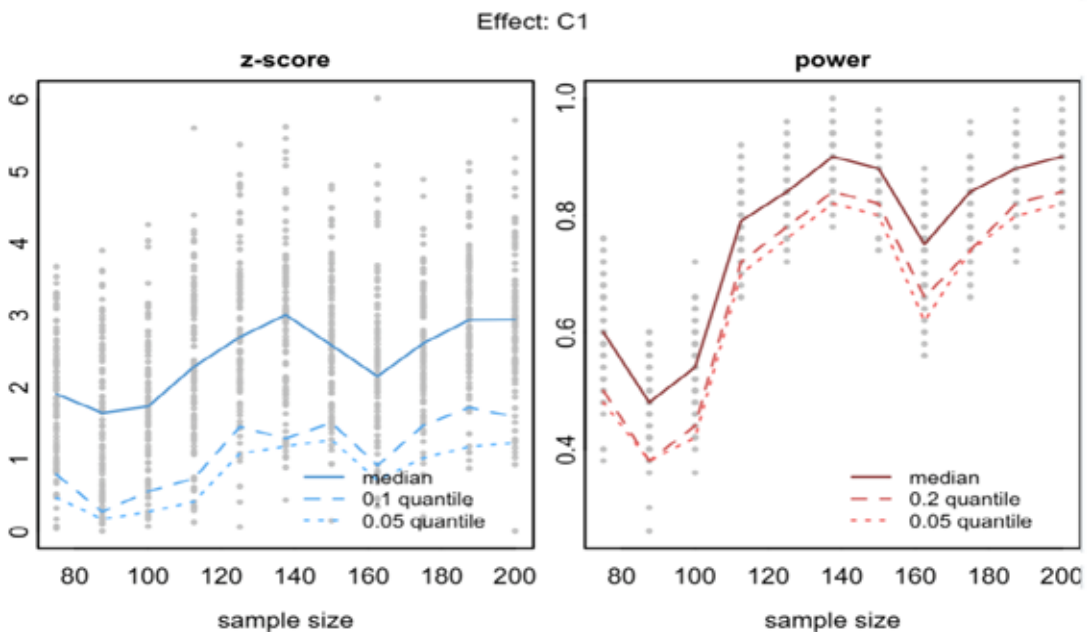

Fig. 4. Z-score and Power curves

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

Discussion. The combined problem of heterogeneity and absence of precise timing calls for particular attention and solution for such data are ubiquitous. The solutions are usually dished up with generalised poisson regression [5] or to that reason with mean-parametrized Conway–Maxwell–Poisson regression [6]. First tries to relax coincidental equality of first two moments of poisson distribution, while the second generalizes the poisson distribution by adding a parameter to model overdispersion. The NB2 distribution is most flexible in that its joint poisson and gamma distributions, the last accommodates random effects, namely gamma–Poisson random variable is a Poisson random variable with a random parameter μ which has the gamma distribution with parameters α and β. Given that NB2 can be suitably managed with MCMC samplers with random effects prior distribution rendered by gamma prior [3]. This is the most efficient way to tackle over dispersed counts that is rendering it with flexible priors. There are no to our knowledge publications shedding the light on heterogenous counts distribution based regressions power analysis, theory and implementation of which developed by Ocheredko Oleksandr [7]. Without power support it's impossible to test hypotheses based on interval censored data unambiguously. Firstly, we showed that heterogeneity greatly undermines power of the test in manual to R library «ltable» [8]. Secondly, as appeared in this case one may lack sample power. We referred to synthetic simulated data approach to examine the capacity of NB2 approach to uncover true generation mechanism under heterogeneity and absence of precise timing. Data shortage (again widespread situation) should be supported by congruous power analysis the benefit is also demonstrated. We suggest that tentative example supports the effectiveness of modelling. In its turn power analysis helps to validate the output and to reveal and substantiate true data generation mechanism of interval censored survival data.

Conclusions. Modelling interval over-dispersed censored survival data is still a challenge due to heterogeneity and profile shortage that underpowers hypotheses tests.

We suggest NB2 based regression to process such data with support of power analysis.

Simulated data analysis supports the effectiveness of modelling.

REFERENCES

1. Klein JP, Moeschberger M. Survival Analysis. New York: Springer Verlag, 1997.
2. CRAN R package ltable. https://cran.r-project.org/package=ltable
3. Congdon P. Bayesian models for categorical data. 2005. John Wiley & Sons Ltd, England. 415 p.
4. Agresti A. Categorical Data Analysis. 3rd ed. (Wiley series in prob. and stat.; 792). 2013.
5. Consul PC., Famoye F. Generalized poisson regression model. *Communications in Statistics – Theory and Methods.* 1992;21(1);89-109 https://doi.org/10.1080/03610929208830766
6. Alan Huan. Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling*. 2017;17(6):359–380. https://doi.org/10.1177/1471082X176977
7. Ocheredko OM. MCMC Bootstrap Based Approach to Power and Sample Size Evaluation. /New Developments in Data Science and Analytics. Proceedings of the 2019 Meeting of International Society for Data Science and Analytics. Zhiyong Zhang Ke-Hai Yuan Yong Wen Jiashan Tang. ISDSA Press · Granger, IN. p.67-87
8. Ocheredko Oleksandr. Library «ltable». The Comprehensive R Archive Network. https://cran.r-project.org/package=ltable