Model selection and comparison

an example with package `Countr`

Tarak Kharrat[1] and Georgi N. Boshnakov[2]

[1]Salford Business School, University of Salford, UK.
[2]School of Mathematics, University of Manchester, UK.

August 15, 2019

Abstract

This document describes a strategy to choose between various possible count models. The computation described in this document is done in `R` (R Core Team, 2016) using the contributed package `Countr` (Kharrat and Boshnakov, 2017) and the `quine` data shipped with the `MASS` package (Venables and Ripley, 2002). The ideas used here are inspired by the demand for medical care example detailed in Cameron and Trivedi (2013, Section 6.3).

This vignette is part of package `Countr` (see Kharrat et al., 2019).

1 Prerequisites

We will do the analysis of the data with package `Countr`, so we load it:

```
library(Countr)
library("MASS") # for glm.nb()
```

Packages `dplyr` (Wickham and Francois, 2016) and `xtable` (Dahl, 2016) provide usefull facilities for data manipulation and presentation:

```
library(dplyr)
library(xtable)
```

2 Data

The dataset used in this example is the `quine` data shipped with package `MASS` (Venables and Ripley, 2002) and first analysed in Aitkin (1978). The data can be loaded in the usual way:

```
data(quine, package = "MASS")
```

The dataset gives the number of days absent from school (`Days`) of 146 children in a particular school year. A number of explanatory variables are available describing the children's ethnic background (`Eth`), sex (`Sex`), age (`Age`) and learner status (`Lrn`). The count variable `Days` is characterised by large *overdispersion* — the variance is more than 16 times larger the mean, 264.2 versus 16.46. Table 1 gives an idea about its distribution. The entries in the table were calculated as follows:

```
breaks_ <- c(0, 1, 3, 5:7, 9, 12, 15, 17, 23, 27, 32)
freqtable <-
    count_table(count = quine$Days, breaks = breaks_, formatChar = TRUE)
```

	0	1-2	3-4	5	6	7-8	9-11
Frequency	9	10	7	19	8	10	13
Relative frequency	0.062	0.068	0.048	0.13	0.055	0.068	0.089

	12-14	15-16	17-22	23-26	27-31	>= 32
Frequency	13	6	14	6	6	25
Relative frequency	0.089	0.041	0.096	0.041	0.041	0.17

Table 1: quine data: Frequency distribution of column `Days`.

3 Models for `quine` data

Given the extreme over-dispersion observed in the `quine` data, we do not expect the Poisson data to perform well. Nevertheless, we can still use this model as a starting point and treat it as a benchmark (any model worse than Poisson should be strongly rejected). Besides, the negative binomial (as implemented in `MASS:glm.nb()`) and 3 renewal-count models based on weibull, gamma and generalised-gamma inter-arrival times are also considered giving in total 5 models to choose from. The code used to fit the 5 models is provided below:

```
quine_form <- as.formula(Days ~ Eth + Sex + Age + Lrn)
pois <- glm(quine_form, family = poisson(), data = quine)
nb <- glm.nb(quine_form, data = quine)

## various renewal models
wei <- renewalCount(formula = quine_form, data = quine, dist = "weibull",
                            computeHessian = FALSE, weiMethod = "conv_dePril",
                            control = renewal.control(trace = 0,
                                method = c("nlminb",
                                     "Nelder-Mead","BFGS")),
                            convPars = list(convMethod = "dePril")
                            )

gam <- renewalCount(formula = quine_form, data = quine, dist = "gamma",
                        computeHessian = FALSE, weiMethod = "conv_dePril",
                        control = renewal.control(trace = 0,
                                                  method = "nlminb"),
                            convPars = list(convMethod = "dePril")
                        )

gengam <- renewalCount(formula = quine_form, data = quine, dist = "gengamma",
                           computeHessian = FALSE, weiMethod = "conv_dePril",
                           control = renewal.control(trace = 0,
                                                     method = "nlminb"),
                               convPars = list(convMethod = "dePril")
                           )
```

Note that it is possible to try several optimisation algorithms in a single call to `renewalCount()`, as in the computation of `wei` above for the weibull-count model. The function will choose the best performing one in terms of value of the objective function (largest log-likelihood). It also

reports the computation time of each method, which may be useful for future computations (for example, bootstrap).

4 Model Selection and Comparison

The different models considered here are fully parametric. Therefore, a straightforward method to use to discriminate between models is a likelihood ratio test (LR). This is possible when models are nested and in this case the LR statistic has the usual $\chi^2(p)$ distribution, where p is the difference in the number of parameters in the model. In this current study, we can compare all the renewal-count models against Poisson, negative-binomial against Poisson, weibull-count against generalised-gamma and gamma against the generalised-gamma.

For non-nested models, the standard approach is to use information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). This method can be applied to discriminate between weibull and gamma renewal count models and between these two models and the negative binomial.

Therefore, a possible strategy can be summarised as follows:

- Use the LR test to compare Poisson with negative binomial.
- Use the LR test to compare Poisson with weibull-count.
- Use the LR test to compare Poisson with gamma-count.
- Use the LR test to compare Poisson with generalised-gamma-count.
- Use the LR test to compare weibull-count with generalised-gamma-count.
- Use the LR test to compare gamma-count with generalised-gamma-count.
- Use information criteria to compare gamma-count with weibull-count.
- Use information criteria to compare weibull-count to negative binomial.

	Alternative.model	Chisq	Df	Critical_value
1	negative-binomial	1192.03	1.00	3.84
2	weibull	1193.21	1.00	3.84
3	gamma	1193.36	1.00	3.84
4	generalised-gamma	1194.46	2.00	5.99

Table 2: LR results against Poisson model. Each row compares an alternative model vs the Poisson model. All alternatives are preferable to Poisson.

As observed in Table 2, the LR test rejects the null hypothesis and all the alternative models are preferred to Poisson. This due to the large over-dispersion discussed in the previous section. Next, we use the LR test to discriminate between the renewal count models:

	Model	Chisq	Df	Critical_value
1	weibull	1.25	1.00	3.84
2	gamma	1.10	1.00	3.84

Table 3: LR results against generalised-gamma model

The results of Table 3 suggest that the weibull and gamma models should be preferred to the generalised gamma model.

Finally, we use information criteria to choose the best model among the weibull and gamma renewal models and the negative binomial:

```
ic <- data.frame(Model = c("weibull", "gamma", "negative-binomial"),
                 AIC = c(AIC(wei), AIC(gam), AIC(nb)),
```

```
                BIC = c(BIC(wei), BIC(gam), BIC(nb)),
                stringsAsFactors = FALSE
                )

print(xtable(ic, caption = "Information criteria results",
             label = "tab:ic_models"))
```

	Model	AIC	BIC
1	weibull	1107.98	1131.84
2	gamma	1107.83	1131.70
3	negative-binomial	1109.15	1133.02

Table 4: Information criteria results

It is worth noting here that all models have the same degree of freedom and hence comparing information criteria is equivalent to comparing the likelihood value. According to Table 4, the gamma renewal model is slightly preferred to the weibull model.

5 Goodness-of-fit

We conclude this analysis by running a formal chi-square goodness of fit test (Cameron and Trivedi, 2013, Section 5.3.4) to the results of the previously selected model.

```
  gof <- chiSq_gof(gam, breaks = breaks_)
  print(gof)

chi-square goodness-of-fit test

Cells considered 0 1-2 3-4 5 6 7-8 9-11 12-14 15-16 17-22 23-26 27-31 >= 32
  DF  Chisq Pr(>Chisq)
1 12 17.502     0.1317
```

The null hypothesis cannot be rejected at standard confidence levels and conclude that the selected model presents a good fit to the data. User can check that the same test yields similar results for the weibull and negative binomial models but comfortably rejects the null hypothesis for the Poisson and generalised gamma models. These results confirm what we observed in the previous section.

We conclude this document by saving the work space to avoid re-running the computation in future exportation of the document:

```
save.image()
```

References

Aitkin, M. (1978). The analysis of unbalanced cross-classifications. *Journal of the Royal Statistical Society. Series A (General)*, pages 195–223.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

Dahl, D. B. (2016). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-2.

Kharrat, T. and Boshnakov, G. N. (2017). *Countr: Flexible Univariate Count Models Based on Renewal Processes*. R package version 3.4.0.

Kharrat, T., Boshnakov, G. N., McHale, I., and Baker, R. (2019). Flexible regression models for count data based on renewal processes: The Countr package. *Journal of Statistical Software*, 90(13):1–35.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wickham, H. and Francois, R. (2016). *dplyr: A Grammar of Data Manipulation*. R package version 0.5.0.