A Short Demo Article: Regression Models for Count Data in R

Achim Zeileis
Universität Innsbruck

Second Author
Plus Affiliation

Abstract

This short article illustrates how to write a manuscript for the *Journal of Statistical Software* (JSS) using its LaTeX style files. Generally, we ask to follow JSS's style guide and FAQs precisely. Also, it is recommended to keep the LaTeX code as simple as possible, i.e., avoid inclusion of packages/commands that are not necessary. For outlining the typical structure of a JSS article some brief text snippets are employed that have been inspired by Zeileis, Kleiber, and Jackman (2008), discussing count data regression in R. Editorial comments and instructions are marked by vertical bars.

Keywords: JSS, style guide, comma-separated, not capitalized, R.

1. Introduction: Count data regression in R

The introduction is in principle "as usual". However, it should usually embed both the implemented *methods* and the *software* into the respective relevant literature. For the latter both competing and complementary software should be discussed (within the same software environment and beyond), bringing out relative (dis)advantages. All software mentioned should be properly `\cite{}`d. (See also Appendix B for more details on BibTeX.)

For writing about software JSS requires authors to use the markup `\proglang{}` (programming languages and large programmable systems), `\pkg{}` (software packages), `\code{}` (functions, commands, arguments, etc.). If there is such markup in (sub)section titles (as above), a plain text version has to be provided in the LaTeX command as well. Below we also illustrate how abbrevations should be introduced and citation commands can be employed. See the LaTeX code for more details.

Modeling count variables is a common task in economics and the social sciences. The classical Poisson regression model for count data is often of limited use in these disciplines because empirical count data sets typically exhibit overdispersion and/or an excess number of zeros. The former issue can be addressed by extending the plain Poisson regression model in various directions: e.g., using sandwich covariances or estimating an additional dispersion parameter (in a so-called quasi-Poisson model). Another more formal way is to use a negative binomial (NB) regression. All of these models belong to the family of generalized linear models (GLMs). However, although these models typically can capture overdispersion rather well, they are in many applications not sufficient for modeling excess zeros. Since Mullahy (1986) there is increased interest in zero-augmented models that address this issue by a second model component capturing zero counts. An overview of count data models in econometrics, including hurdle and zero-inflated models, is provided in Cameron and Trivedi (2013).

In R (R Core Team 2017), GLMs are provided by the model fitting functions `glm()` in the **stats** package and `glm.nb()` in the **MASS** package (Venables and Ripley 2002, Chapter 7.4) along with associated methods for diagnostics and inference. The manuscript that this document is based on (Zeileis *et al.* 2008) then introduced hurdle and zero-inflated count models in the functions `hurdle()` and `zeroinfl()` in the **pscl** package (Jackman 2015). Of course, much more software could be discussed here, including (but not limited to) generalized additive models for count data as available in the R packages **mgcv** Wood (2006), **gamlss** (Stasinopoulos and Rigby 2007), or **VGAM** (Yee 2010).

2. Models and software

The basic Poisson regression model for count data is a special case of the GLM framework McCullagh and Nelder (1989). It describes the dependence of a count response variable y_i ($i = 1, \dots, n$) by assuming a Poisson distribution $y_i \sim \mathrm{Pois}(\mu_i)$. The dependence of the conditional mean $\mathsf{E}[y_i \,|\, x_i] = \mu_i$ on the regressors x_i is then specified via a log link and a linear predictor

$$\log(\mu_i) \quad = \quad x_i^\top \beta, \tag{1}$$

where the regression coefficients β are estimated by maximum likelihood (ML) using the iterative weighted least squares (IWLS) algorithm.

> Note that around the `{equation}` above there should be no spaces (avoided in the LaTeX code by `%` lines) so that "normal" spacing is used and not a new paragraph started.

R provides a very flexible implementation of the general GLM framework in the function `glm()` (Chambers and Hastie 1992) in the **stats** package. Its most important arguments are

```
glm(formula, data, subset, na.action, weights, offset,
  family = gaussian, start = NULL, control = glm.control(...),
  model = TRUE, y = TRUE, x = FALSE, ...)
```

where `formula` plus `data` is the now standard way of specifying regression relationships in R/S introduced in Chambers and Hastie (1992). The remaining arguments in the first line (`subset`, `na.action`, `weights`, and `offset`) are also standard for setting up formula-based regression models in R/S. The arguments in the second line control aspects specific to GLMs while the arguments in the last line specify which components are returned in the fitted model object (of class '`glm`' which inherits from '`lm`'). For further arguments to `glm()` (including alternative specifications of starting values) see `?glm`. For estimating a Poisson model `family = poisson` has to be specified.

> As the synopsis above is a code listing that is not meant to be executed, one can use either the dedicated `{Code}` environment or a simple `{verbatim}` environment for this. Again, spaces before and after should be avoided.
>
> Finally, there might be a reference to a `{table}` such as Table 1. Usually, these are placed at the top of the page (`[t!]`), centered (`\centering`), with a caption below the table, column headers and captions in sentence style, and if possible avoiding vertical lines.

Type	Distribution	Method	Description
GLM	Poisson	ML	Poisson regression: classical GLM, estimated by maximum likelihood (ML)
		Quasi	"Quasi-Poisson regression": same mean function, estimated by quasi-ML (QML) or equivalently generalized estimating equations (GEE), inference adjustment via estimated dispersion parameter
		Adjusted	"Adjusted Poisson regression": same mean function, estimated by QML/GEE, inference adjustment via sandwich covariances
	NB	ML	NB regression: extended GLM, estimated by ML including additional shape parameter
Zero-augmented	Poisson	ML	Zero-inflated Poisson (ZIP), hurdle Poisson
	NB	ML	Zero-inflated NB (ZINB), hurdle NB

Table 1: Overview of various count regression models. The table is usually placed at the top of the page (`[t!]`), centered (`centering`), has a caption below the table, column headers and captions are in sentence style, and if possible vertical lines should be avoided.

3. Empirical Illustration

This section introduces the main features of the package **SmoothCurves** employing some simulated datasets and the Canadian daily temperature data (Ramsay and Silverman 2005), that are available e.g. in the **fda** package (Ramsay, Wickham, Graves, and Hooker 2018). It contains daily observations of temperature (in řC) for $N = 35$ Canadian weather stations averaged over the years 1960 to 1994.

From within R, the latest version of **SmoothCurves** can be installed from the Comprehensive R Archive Network (CRAN) (maybe one day) using the following

```
R> install.packages("SmoothCurves")
```

and the latest development version of the package can be installed from Github using the following

```
R> devtools::install_github("StevenGolovkine/SmoothCurves")
```

Once the package has been installeed, it can be attached by typing:

```
R> library(SmoothCurves)
```

Additionally, the package help manual can be accessed via:

```
R> help(package="SmoothCurves")
```

3.1. A note on the format of the data

3.2. Brownian motion

The data are generated from the following model:

$$Y_{n,j} = X_n(T_{n,j}) + \varepsilon_{n,j}, \quad 1 \leq n \leq N, \tag{2}$$

where $T_{n,j} \sim f$, with f a probability density function with support $[0,1]$, and $\varepsilon_{n,j} \sim \mathcal{N}(0,\sigma^2)$ are independant variables. Denote by H the regularity of X and define X as a standard Brownian motion $(H = 0.5). The fractional\_brownian data from the package Smooth Curves is used.$ The data can be loaded by

```
R> data("fractional_brownian", package = "SmoothCurves")
```

The smoothing of the curves can be easyly performed using one line

```
R> smooth_data <- smooth.curves(fractional_brownian)
R> smooth_data$parameter

$sigma
[1] 0.07080734

$H0
[1] 0.4956724

$L0
[1] 0.9788827

$b
[1] 0.00718395
```

We discuss alternative choices for implementation and inference further below, but first we plot the results obtained above.

```
R> df <- tibble(t = fractional_brownian[[1]]$t,
+               x = fractional_brownian[[1]]$x,
+               x_true = fractional_brownian[[1]]$x_true,
+               x_hat = smooth_data$smooth[[1]]$x)
R> ggplot(df) +
+    geom_point(aes(x = t, y = x), alpha = 0.5) +
+    geom_line(aes(x = t, y = x_true), col = 'red') +
+    geom_line(aes(x = t, y = x_hat), col = 'blue') +
+    labs(x = '', y = '') +
+    theme_custom()
```

4. Summary and discussion

▌ As usual . . .

Computational details

▌ If necessary or useful, information about certain computational details such as version numbers, operating systems, or compilers could be included in an unnumbered section. Also, auxiliary packages (say, for visualizations, maps, tables, . . .) that are not cited in the main text can be credited here.

The results in this paper were obtained using R 3.6.2 with the **MASS** 7.3.51.5 package. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

Acknowledgments

▌ All acknowledgments (note the AE spelling) should be collected in this unnumbered section before the references. It may contain the usual information about funding and feedback from colleagues/reviewers/etc. Furthermore, information such as relative contributions of the authors may be added here (if any).

References

Cameron AC, Trivedi PK (2013). *Regression Analysis of Count Data*. 2nd edition. Cambridge University Press, Cambridge.

Chambers JM, Hastie TJ (eds.) (1992). *Statistical Models in S*. Chapman & Hall, London.

Jackman S (2015). ***pscl**: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. R package version 1.4.9, URL https://CRAN.R-project.org/package=pscl.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. 2nd edition. Chapman & Hall, London. doi:10.1007/978-1-4899-3242-6.

Mullahy J (1986). "Specification and Testing of Some Modified Count Data Models." *Journal of Econometrics*, **33**(3), 341–365. doi:10.1016/0304-4076(86)90002-3.

Ramsay J, Silverman BW (2005). *Functional Data Analysis*. Springer Series in Statistics, 2 edition. Springer-Verlag, New York. ISBN 978-0-387-40080-8. URL //www.springer.com/fr/book/9780387400808.

Ramsay JO, Wickham H, Graves S, Hooker G (2018). *fda: Functional Data Analysis*. R package version 2.4.8, URL https://CRAN.R-project.org/package=fda.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Stasinopoulos DM, Rigby RA (2007). "Generalized Additive Models for Location Scale and Shape (GAMLSS) in R." *Journal of Statistical Software*, **23**(7), 1–46. `doi:10.18637/jss.v023.i07`.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York. `doi:10.1007/978-0-387-21706-2`.

Wood SN (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton.

Yee TW (2010). "The **VGAM** Package for Categorical Data Analysis." *Journal of Statistical Software*, **32**(10), 1–34. `doi:10.18637/jss.v032.i10`.

Zeileis A, Kleiber C, Jackman S (2008). "Regression Models for Count Data in R." *Journal of Statistical Software*, **27**(8), 1–25. `doi:10.18637/jss.v027.i08`.

A. More technical details

Appendices can be included after the bibliography (with a page break). Each section within the appendix should have a proper section title (rather than just *Appendix*).

For more technical style details, please check out JSS's style FAQ at `https://www.jstatsoft.org/pages/view/style#frequently-asked-questions` which includes the following topics:

- Title vs. sentence case.
- Graphics formatting.
- Naming conventions.
- Turning JSS manuscripts into R package vignettes.
- Trouble shooting.
- Many other potentially helpful details…

B. Using BibTEX

References need to be provided in a BIBTEX file (`.bib`). All references should be made with `\cite`, `\citet`, `\citep`, `\citealp` etc. (and never hard-coded). This commands yield different formats of author-year citations and allow to include additional details (e.g., pages, chapters, …) in brackets. In case you are not familiar with these commands see the JSS style FAQ for details.

Cleaning up BIBTEX files is a somewhat tedious task – especially when acquiring the entries automatically from mixed online sources. However, it is important that informations are complete and presented in a consistent style to avoid confusions. JSS requires the following format.

- JSS-specific markup (`\proglang`, `\pkg`, `\code`) should be used in the references.
- Titles should be in title case.
- Journal titles should not be abbreviated and in title case.
- DOIs should be included where available.
- Software should be properly cited as well. For R packages `citation("pkgname")` typically provides a good starting point.

Affiliation:

Achim Zeileis
Journal of Statistical Software
and
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: https://eeecon.uibk.ac.at/~zeileis/