XV EMR - Escola de Modelos de Regressão

SHORT COURSE

Regression Models for Count Data

March 26 to 29, 2017

Wagner Hugo Bonat
Walmes Marques Zeviani
Eduardo Elias Ribeiro Jr

Regression Models for Count Data

Wagner Hugo Bonat
Walmes Marques Zeviani
Eduardo Elias Ribeiro Jr

Regression Models for Count Data

Wagner Hugo Bonat
www.leg.ufpr.br/~wagner

Walmes Marques Zeviani
www.leg.ufpr.br/~walmes

Eduardo Elias Ribeiro Jr
www.leg.ufpr.br/~eduardojr

Laboratório de Estatística e Geoinformação (LEG)
http://www.leg.ufpr.br
Departamento de Estatística
Universidade Federal do Paraná (UFPR)

Supplementary content: http://www.leg.ufpr.br/rmcd
Contact: rmcd@leg.ufpr.br

XV EMR - Escola de Modelos de Regressão
Goiânia - Goiás, Brasil
March 26 to 29, 2017

Contents

Preface

The analysis of normal and non-normal data are mostly based on the class of generalized linear models (Nelder and Wedderburn, 1972). The class offers a very attractive statistical modelling framework which includes the Gaussian, logistic and Poisson regression models for the analysis of continuous, binomial and count data, respectively. The theoretical background for the GLM is based on the exponential dispersion models (Jørgensen, 1987, 1997) as a generalization of the exponential family of distributions. Furthermore, the whole class of models can be fitted by a simple Newton score algorithm relying only on second-moment assumptions for estimation and inference. Despite of the flexibility of the GLM class, the Poisson distribution is the only choice for analysis of count data. For this reason, in practice there is probably an over-emphasis on the use of the Poisson distribution. A well known limitation of the Poisson distribution is the mean and variance relationship, referred to as equidispersion. In practice, however, count data can present other features, namely underdispersion and overdispersion that is often related to zero-inflation, heavy tail or absence of important explanatory variables. These features can make the Poisson distribution unsuitable for the analysis of count data. The main goal of this course is to present a wider range of statistical models to deal with count data. In particular, we focus on parametric and second-moments specified models. We shall present the model specification along with strategies for model fitting and the associated R code. Furthermore, a book-course and supplementary material as R (R Core Team, 2016) code and data sets will be made available for the students. We intend to keep the course in a level suitable for bachelor students who already attended a course on generalized linear models. However, since the course also covers updated topics, it can be of interest of postgraduate students and researches in general. In what follows, we describe the course structure as well as the main bibliography references on which the course is based. The subject covered and the Expected learning outcome.

Chapter 1

Introduction

Figure 1.1 illustrates the generator process for under, over and equidispered counts in two dimensions context. The grid lines in this figure indicate fixed regions for which events are counted and the counts within each interval are displayed. For the equidispersed case the distribution of events is random. In overdispersed case, the events are clustered. This behaviour can be explained by a contamination process (e.g. count contagious disease). The underdispersed case, in contrast of overdispersion, shows the events distribution is nearly regular and the counts have smaller variances. The natural process that explains underdispersion is repulsion, exactly the opposite of overdispersion, that means, an event occurence inhibits others near (e.g. count territorialistas animals).

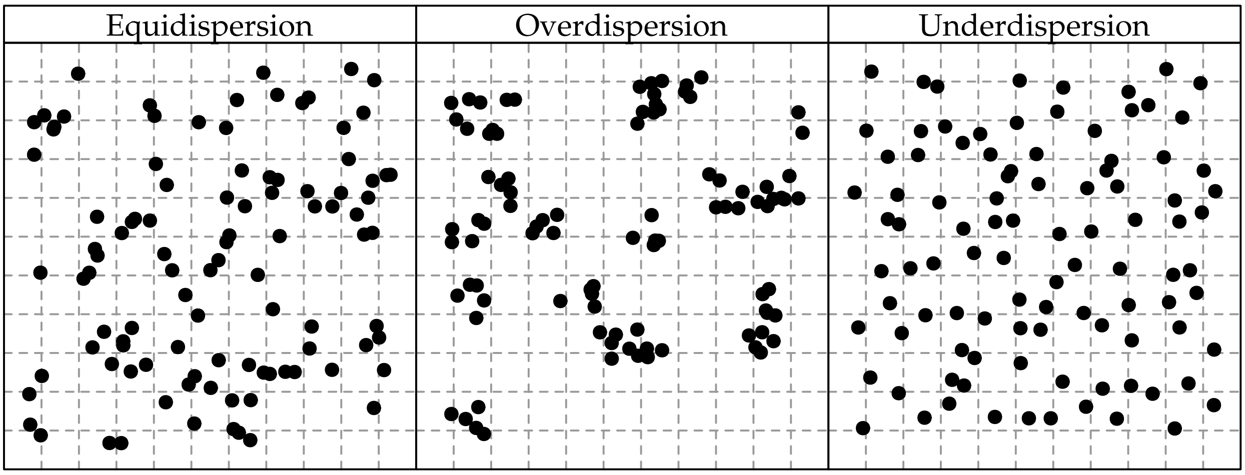

Figure 1.1: Illustration of generator process for under, over and equidispered count data.

Chapter 2

Background

Chapter 3

Full Parametric Approach

3.1 Models for count data

3.2 Regression models

3.3 Estimation and inference

3.4 Computational implementation

Chapter 4

Second Moments Based Especification

4.1 Mean and variance relationship

4.2 Estimation functions approach

4.3 Extended Poisson-Tweedie regression models

4.4 Computational implementation

Chapter 5

Data analysis

5.1 Ovedispersed case

5.2 Underdispersed case

5.2.1 Cotton bolls greenhouse experiment

The data set in this section come from a greenhouse experiment with cotton plants (Gossypium hirsutum) obtained under a completely randomized design with five replicates. The experiment was aimed to assess the effects of five defoliation levels (0%, 25%, 50%,75% and 100%) on the observed number of bolls produced by plants at five growth stages: vegetative, flowerbud, blossom, fig and cotton boll. The experimental unity was a vase with two plants. The number of cotton bolls was recorded at the each culture cycle. This data was analysed by Bonat et al. (2016), with extended Poisson-Tweedie model and Zeviani et al. (2014) with Gamma-Count model. In this section the results of articles are reproduced and compared jointly others alternatives for analysis (COM-Poisson model and Generalized-Poisson model).

This code below shows how to get the data and how it is structured in R. The data set contains 125 records and 4 variables, described below:

- `defol` A numeric factor with 5 levels that represents the (artifitial) levels of defoliation (percent in leaf area removed with scissors) applied for all leaves of the plants.

- phenol A categorical ordered factor with 5 levels that represents the (phenological) growth stages of the cotton plants in which the levels of defoliation was applied.

- rept Integer variable that indexes each experimenal unit in each treatment cell.

- bolls The number of bolls produced (count variable) evaluated at harvest of cotton.

```
## Read the data via package book
cotton <- read.table("./data/cotton.csv", header = TRUE,
                     sep = ";")  ## Remove this
## data(cotton, package = "rmcd")

## ## or read the data via url address
## url <- "http://cursos.leg.ufpr.br/rmcd/data/cotton.csv"
## cotton <- read.table(url, header = TRUE, sep = ";")

str(cotton)

## 'data.frame':    125 obs. of  4 variables:
##  $ phenol: Factor w/ 5 levels "blossom","boll",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ defol : num  0 0 0 0 0 0.25 0.25 0.25 0.25 0.25 ...
##  $ rept  : int  1 2 3 4 5 1 2 3 4 5 ...
##  $ bolls : int  10 9 8 8 10 11 9 10 10 10 ...
```

Figure 5.1 shows the number of cotton bolls recorded for each combination of defoliation level and growth stage.

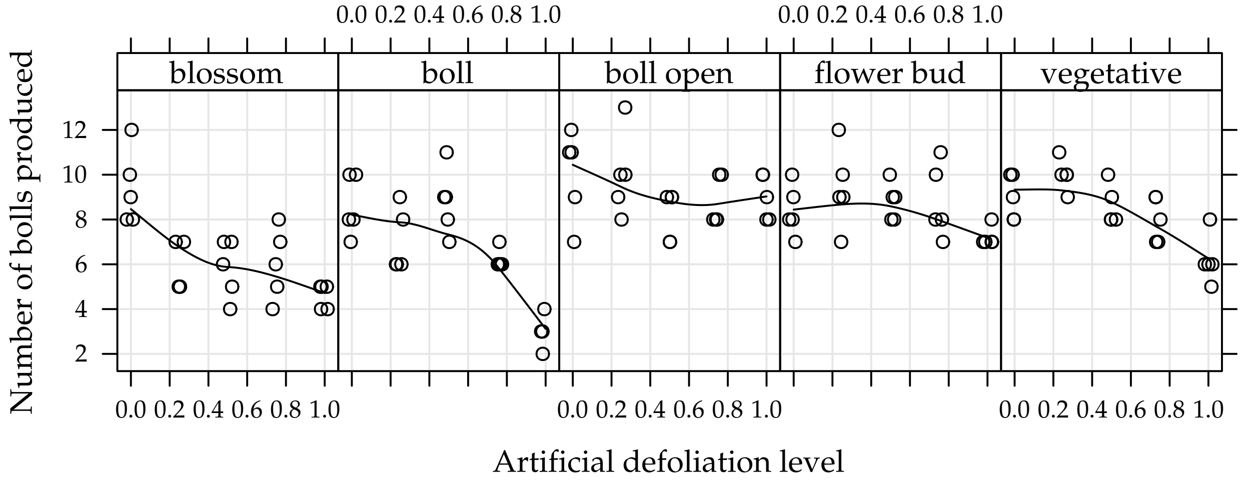

Figure 5.1: Number of bolls produced for each artificial defoliation level and each growth stage.

5.3 Equidispersed case

5.4 Zero-Inflated case

Chapter 6

Discussion

Bibliography

Bonat, W. H., Jørgensen, B., Kokonendji, C. C., Hinde, J., and Demétrio, C. G. B. (2016). Extended poisson-tweedie: properties and regression models for count data.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(2):127–162.

Jørgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall, London.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

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

Zeviani, W. M., Ribeiro Jr, P. J., Bonat, W. H., Shimakura, S. E., and Muniz, J. A. (2014). The gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, 41(12):2616–2626.