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)

Complementos online: http://www.leg.ufpr.br/rmcd
Contato: rmcd@leg.ufpr.br

XV EMR - Escola de Modelos de Regressão
26 a 19 de março de 2017
Goiânia - GO - Brasil

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).

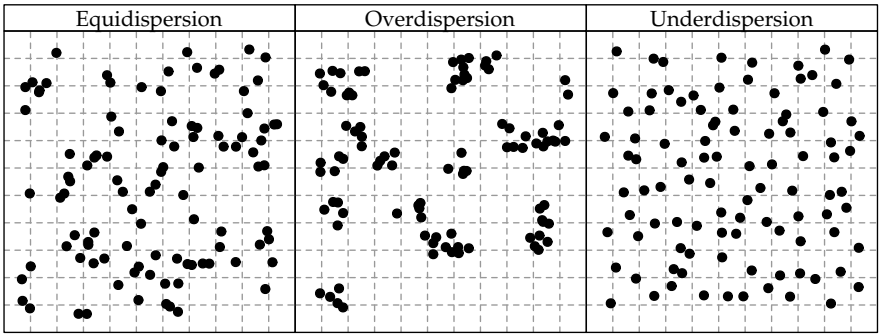

Figure 1.1: Illustration of generator process for under, over and equidispered count data.

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 analyses

Chapter 6

Discussion

Bibliography

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.