countreg: Tools for count data regression

Christian Kleiber $^1$ Achim Zeileis $^2$

$^1$Universität Basel
$^2$Universität Innsbruck

useR!2017, Brussels, 2017-07-06
Outline

1. Introduction

2. Overview of the countreg Package

3. Zero-altered Count Data Regression

4. Example: Doctor Visits

5. Open issues / To do

6. References
Introduction

Count data are everywhere:

actuarial science, ecology, econometrics, engineering, epidemiology, fisheries research, political science, psychology, . . .

Count data regression:

• classical models: Poisson, negative binomial (incl. geometric)
• frequent issues: overdispersion, excess zeros
• some extensions
  ▶ hurdle models aka two-part models (Mullahy 1986)
  ▶ zero-inflation models (Lambert 1992)
  ▶ finite mixtures
  ▶ Negbin-H (multiple index model, think gamlss)

Count data regression in R:

• Poisson and negative binomial regressions in stats and MASS
• many further methods, but widely scattered across packages
• countreg: zero-altered models, unification, diagnostics, . . .
Overview of the `countreg` package

Fitting functions:

- `hurdle()`, `zeroinfl()` (extended versions of functions in `pscl`)
- `zerotrunc()` (zero-truncated Poisson, Negbin, geometric)

Model visualization:

- `rootogram()`
  Rootograms for regressions, via aggregation of conditional distributions
  Methods for objects of classes “glm”, “glm.nb”, “hurdle”, “zeroinfl”, “zerotrunc”, also for “gam”, “gamlss”

- `pit()`, `pithist()`
  Probability integral transform (PIT) and associated histograms

Diagnostics:

- `qresiduals()` (randomized quantile residuals, generic function)
Overview of the countreg package

Data sets:

- VolcanoHeights, SerumPotassium (from Tukey papers)
- CodParasites (ecology)
- CrabSatellites (ethology)
- OralHealthNL (dentistry)
- TakeoverBids (finance)

Infrastructure for zero-altered / zero-modified distributions:

- hurdle / zeroinfl / zeroetrunc versions of d/p/q/rdist
- sdist (score function) for count distributions

Miscellanea:

- driver for finite mixtures of negative binomial regressions, for use with flexmix
- (enhanced) negative binomial, zero-truncated Poisson / negative binomial families for use with mboost
Zero-altered count data regression

**Two main models** for data with excess zeros:

- **Zero-inflation models** (two-component finite mixtures)
  \[ f_{\text{zeroinfl}}(y; x, z, \beta, \gamma) = f_{\text{zero}}(0; z, \gamma) \cdot I\{0\}(y) + (1 - f_{\text{zero}}(0; z, \gamma)) \cdot f_{\text{count}}(y; x, \beta) \]

- **Hurdle models** (aka two-part models)
  \[ f_{\text{hurdle}}(y; x, z, \beta, \gamma) = \begin{cases} 
  f_{\text{zero}}(0; z, \gamma), & y = 0, \\
  (1 - f_{\text{zero}}(0; z, \gamma)) \cdot f_{\text{count}}(y; x, \beta) / \{1 - f_{\text{count}}(0; x, \beta)\}, & y > 0 
\end{cases} \]

Hurdle has many computational advantages: Likelihood separates, etc.

**Example:**

geometric hurdle \( = \) logit + zero-truncated geometric

In R:

\[ \text{hurdle}(..., \text{dist} = "geo") \]
\[ = \text{glm}(..., \text{family} = \text{binomial}) + \text{zerotrunc}(..., \text{dist} = "geo") \]
Example: Doctor visits

**Data:** NMES1988 from *AER* (Kleiber and Zeileis 2008).

Cross-sectional data originating from the US National Medical Expenditure Survey (NMES), 1987 and 1988. The data are a subsample of individuals ages 66 and over, all of whom are covered by Medicare.

Originally \( n = 4406 \) observations, here only \( n = 666 \). Some variables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>visits</td>
<td>Number of doctor visits.</td>
</tr>
<tr>
<td>health</td>
<td>Factor indicating self-perceived health status, levels are &quot;poor&quot;, &quot;average&quot; (reference category), &quot;excellent&quot;.</td>
</tr>
<tr>
<td>chronic</td>
<td>Number of chronic conditions (an integer).</td>
</tr>
<tr>
<td>gender</td>
<td>Factor indicating gender.</td>
</tr>
<tr>
<td>school</td>
<td>Number of years of education.</td>
</tr>
<tr>
<td>insurance</td>
<td>Factor. Is the individual covered by private insurance?</td>
</tr>
<tr>
<td>medicaid</td>
<td>Factor. Is the individual covered by Medicaid?</td>
</tr>
</tbody>
</table>


Also used in count data book by Cameron and Trivedi (2013).
NMES1988 doctor visits: response
NMES1988 doctor visits

Models for doctor visits: (not all shown below)

- (Poisson regression)
- Negative binomial (NB2) regression
- Negative binomial hurdle regression
- Zero-inflated negative binomial regression
- Finite mixture of NB2 regressions (2 components)
- Heterogeneous negative binomial regression (Negbin-H)
NMES1988 doctor visits: rootograms

Poisson

Negative Binomial

countreg: Tools for count data regression
NMES1988 doctor visits: another rootogram

![Graph showing a rootogram for NMES1988 doctor visits. The x-axis represents visits with a range from 0 to 20, and the y-axis represents the square root of frequency with a range from -5 to 10. The graph includes a fitted line and bars representing frequency.]
NMES1988 doctor visits: PIT histograms

Poisson

Negative Binomial
NMES1988 doctor visits: NB2-FM2 regression

Negative Binomial

Mixture Negative Binomial

Mixture Negative Binomial (Component 1)

Mixture Negative Binomial (Component 2)
NMES1988 doctor visits: NB-H regression

NegBin

NegBin−H

sqrt(Frequency)

visits

sqrt(Frequency)

visits
NMES1988 doctor visits: Comparison

Information criteria for all fitted models:

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>nmes.poi</td>
<td>8</td>
<td>5940</td>
<td>5976</td>
</tr>
<tr>
<td>nmes.nb</td>
<td>9</td>
<td>3732</td>
<td>3772</td>
</tr>
<tr>
<td>nmes.h</td>
<td>17</td>
<td>3715</td>
<td>3792</td>
</tr>
<tr>
<td>nmes.zi</td>
<td>17</td>
<td>3717</td>
<td>3794</td>
</tr>
<tr>
<td>nmes.nbh</td>
<td>12</td>
<td>3711</td>
<td>3765</td>
</tr>
</tbody>
</table>
NMES1988 doctor visits: Quantile residuals

Q–Q residuals plot: Hurdle NB

Q–Q residuals plot: NB
NMES1988 doctor visits: Quantile residuals

**One Random Sample**

**Median**

**Five Random Samples and Range**

**Ten Random Samples and Shading**
Open issues / To do

- More predict types?
  
  Currently, `predict.hurdle()` offers
  
  ```r
  type = c("response", "prob", "count", "zero")
  ```
  
  In principle, could also offer
  
  "density", "probability", "quantile", ...

- more methods for `pit()`, `pithist()`

- diagnostic tests: some (score) tests against over/underdispersion

Package currently still on R-Forge at

  `http://R-Forge.R-project.org/projects/countreg/

...“soon” on CRAN!
References


