Wirtschaftswissenschaftliches Zentrum (WWZ) der Universität Basel

WWZ

December 2014

Visualizing Count Data Regressions Using Rootograms

Christian Kleiber, Achim Zeileis

UNI
BASEL

Wirtschaftswissenschaftliches Zentrum (WWZ) der Universität Basel

WW|Z

Corresponding author:

Prof. Dr. Christian Kleiber
Wirtschaftswissenschaftliches Zentrum (WWZ)
Abteilung Quantitative Methoden
Peter Merian-Weg 6
Postfach, CH-4002 Basel
Tel. +41 (0)61 267 33 67
Fax +41 (0)61 267 23 97
christian.kleiber@unibas.ch

———

A publication of the Center of Business and Economics (WWZ), University of Basel.

UNI
BASEL

Visualizing Count Data Regressions Using Rootograms

Christian Kleiber
Universität Basel

Achim Zeileis
Universität Innsbruck

Abstract

The rootogram is a graphical tool associated with the work of J. W. Tukey that was originally used for assessing goodness of fit of univariate distributions. Here we show that rootograms are also useful for diagnosing and treating issues such as overdispersion and/or excess zeros in regression models for count data. We also introduce a weighted version of the rootogram that can be applied out of sample or to (weighted) subsets of the data, e.g., in finite mixture models. Two empirical illustrations are included, one from ethology, the other from public health. The former employs a negative binomial hurdle regression, the latter a two-component finite mixture of negative binomial models. A further illustration involving underdispersion and an R implementation of our tools are available in the R package **countreg**.

Keywords: rootogram, visualization, goodness of fit, count data, Poisson regression, negative binomial regression, hurdle model, finite mixture.

1. Introduction

The area of count data regression has experienced rapid growth over the last two decades. More often than not, the standard Poisson model from the generalized linear model (GLM) toolbox does not suffice in empirical work. Specifically, many data sets are plagued by some form of overdispersion, often resulting from unobserved heterogeneity that can potentially be handled by, e.g., models with additional shape parameters such as the negative binomial distribution or from an excess of zeros for which hurdle and zero-inflation models are available (Mullahy 1986; Lambert 1992). While various diagnostic tests of dispersion are also available – see, e.g., Cameron and Trivedi (1990) or Dean (1992) for some popular tests and Cameron and Trivedi (2013) for an overview – they typically only identify general issues with model fit and rarely provide clear indications regarding the source of the problems. Suitable graphical tools can guide the search for more appropriate specifications, thereby supplementing and enhancing more formal approaches.

If count data regressions are visualized at all, this is currently mainly done in the form of barplots of observed and expected frequencies; see, e.g., Figures 3.1 and 6.4 in Cameron and Trivedi (2013) for examples. In the present paper, we explore the use of rootograms for assessing the fit. Rootograms are associated with the work of John W. Tukey on exploratory data analysis (EDA) and statistical graphics, culminating in Tukey (1977). However, rootograms do not figure prominently there. Instead, early applications, all confined to continuous data,

appear in selected contributions to collected volumes and conference proceedings (Tukey 1965, 1972), which were often not easily available prior to the publication of Tukey's collected works in the 1980s. Nonetheless, the ideas pertaining to rootograms were known in some circles at an early stage (Healy 1968), and an early paper popularizing the concept is Wainer (1974). For further information on the history of statistical graphics we refer to Friendly and Denis (2001).

The following section introduces a generalized version of the rootogram allowing for weights that can be applied to new data or (weighted) subsamples of a data set. This is useful for assessing in-sample fits as well as out-of-sample predictions and also for situations with survey weights or model-based weights. The latter is required below for an empirical illustration involving finite mixture models. Several styles of the rootogram, namely standing, hanging, and suspended versions, are briefly described. We also provide some guidelines for interpretation using simulated data. Section 3 provides two empirical examples. The first presents a case where a hurdle model adjusts for excess zeros. The second considers overdispersion remaining after fitting a negative binomial model. It emerges that a finite mixture model with two negative binomial components provides an improved fit.

All analyses are run in R (R Core Team 2014), and we briefly describe an implementation of our tools in the R package **countreg** in an appendix.

2. Rootograms

Given observations y_i $(i = 1, \ldots, n)$ we want to assess the goodness of fit of some parametric model $F(\cdot; \alpha_i)$, with corresponding density or probability mass function $f(\cdot; \alpha_i)$. The parameter vector α_i could be the same for all observations $i = 1, \ldots, n$ (as considered by Friendly 2000, Chapter 2) but is often observation-specific – typically through dependence on some covariates x_i, a leading case being the GLM with $\alpha_i = g(x_i^\top \beta)$ for some monotonic function $g(\cdot)$. In practice, these parameters are typically unknown and have to be estimated from data. Hence, in the following we assume that we have fitted parameters $\hat{\alpha}_i$ where estimation may have been carried out on the same observations $i = 1, \ldots, n$ (i.e., corresponding to an in-sample assessment) or on a different data set (i.e., out-of-sample evaluation). The estimation procedure itself may be fully parametric or semiparametric etc. as long as it yields fitted parameters $\hat{\alpha}_i$ for all observations of interest.

To judge the goodness of fit of a model with estimated parameters $\hat{\alpha}_i$ to observations y_i $(i = 1, \ldots, n)$, a natural idea is to assess whether observed frequencies match expected frequencies from the model. In the case of discrete observations frequencies for the observations themselves could be considered while somewhat more generally frequencies for intervals of observations may be used. Tukey's original work often considered goodness of fit to the normal distribution on the basis of binned observations, see, e.g., his example involving the heights of 218 volcanos (Tukey 1972). In this paper, we focus on discrete distributions.

For assessing the goodness of fit in regression models, practitioners routinely check some type of residuals, i.e., (weighted) deviations of the observations y_i from the corresponding predicted means. However, this focuses on the first moment of the fitted distribution only while for count data, which are non-negative and typically skewed, further aspects of the distribution are also of interest. Relevant aspects include the amount of (over-)dispersion, skewness (or more general aspects of shape), and whether there are excess zeros. Hence, it is natural to

consider observed and expected values for a range of counts $0, 1, 2, \ldots$ to assess the entire fitted distribution.

Specifically, in the case of count data with possible outcomes $j = 0, 1, 2, \ldots$, the observed and expected frequencies for each integer j are given by

$$\text{obs}_j = \sum_{i=1}^{n} I(y_i = j),$$

$$\text{exp}_j = \sum_{i=1}^{n} f(j; \hat{\alpha}_i),$$

where $I(\cdot)$ is an indicator variable. More generally, one can use a set of breaks $b_0, b_1, b_2, \ldots$ that span (a suitable subset of) the support of y. Here, we additionally also allow for observation-specific weights w_i $(i = 1, \ldots, n)$, the observed and expected frequencies are then given by

$$\text{obs}_j = \sum_{i=1}^{n} w_i \, I(y_i \in (b_j, b_{j+1}]),$$

$$\text{exp}_j = \sum_{i=1}^{n} w_i \left\{ F(b_{j+1}; \hat{\alpha}_i) - F(b_j; \hat{\alpha}_i) \right\}.$$

The weights are needed for survey data and also for situations with model-based weights. For example, the latter may represent class membership in mixture models, a case that is relevant in one of our applications below.

2.1. Styles of rootograms

The rootogram compares observed and expected values graphically by plotting histogram-like rectangles or bars for the observed frequencies and a curve for the fitted frequencies, all on a square-root scale. The square roots rather than the untransformed observations are employed to approximately adjust for scale differences across the j values or intervals. Otherwise, deviations would only be visible for j's with large observed/expected frequencies.

Different styles of rootograms have been suggested, see Figure 1:

- *Standing:* The standing rootogram simply shows rectangles/bars for $\sqrt{\text{obs}_j}$ and a curve for $\sqrt{\text{exp}_j}$. To assess deviations across the j's, the expected curve needs to be followed as the deviations are not aligned.

- *Hanging:* To align all deviations along the horizontal axis, the rectangles/bars are drawn from $\sqrt{\text{exp}_j}$ to $\sqrt{\text{exp}_j} - \sqrt{\text{obs}_j}$ so that they are "hanging" from the curve representing expected frequencies, $\sqrt{\text{exp}_j}$.

- *Suspended:* To emphasize mainly the deviations (rather than the observed frequencies), a third alternative is to draw rectangles/bars for the differences between expected and observed frequencies, $\sqrt{\text{exp}_j} - \sqrt{\text{obs}_j}$ (some authors use $\sqrt{\text{obs}_j} - \sqrt{\text{exp}_j}$ instead).

The basic version, the standing rootogram, is perhaps the least useful among the three: it simply plots rectangles/bars and a curve representing the model, but the fit is not easily assessed. The other versions both make use of a horizontal reference line, a detail often
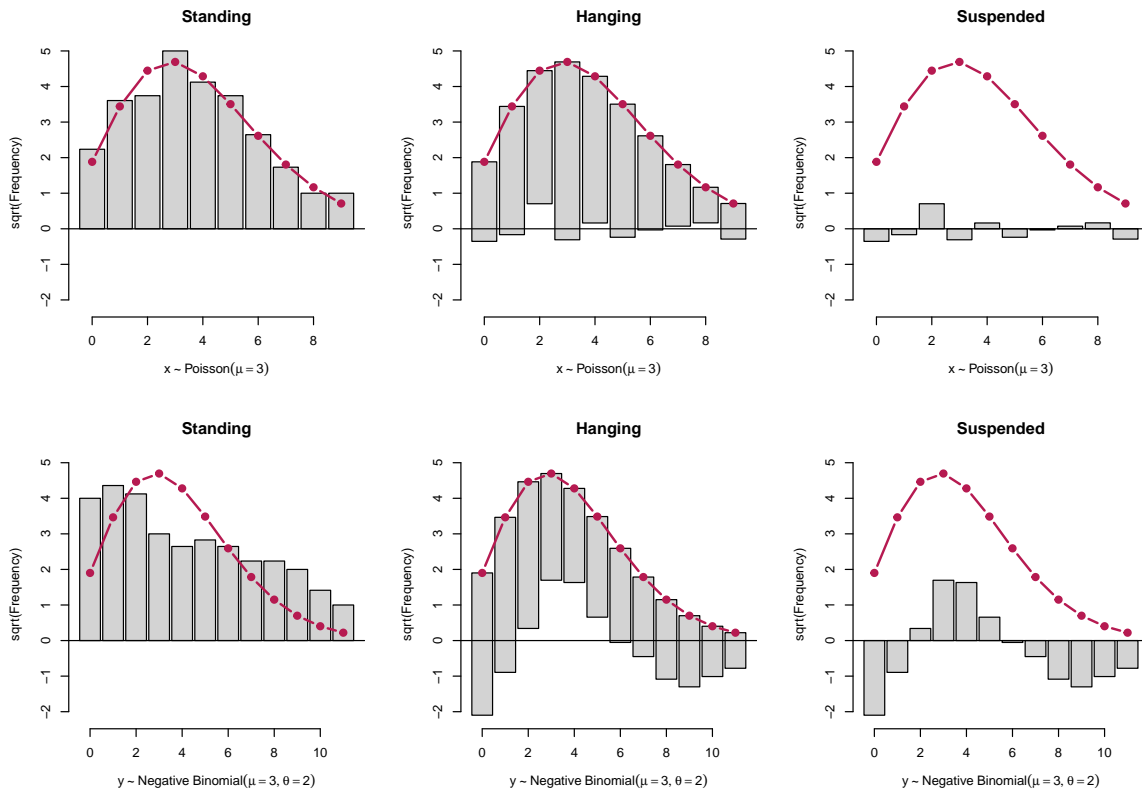
Figure 1:   Styles of rootograms for two Poisson models fitted to 100 artificial observations
from a Poisson (top row) and negative binomial (bottom row) distribution. Top row: The
Poisson model fit ($\hat{\mu} = 3.34$) captures the true mean ($\mu = 3$) as well as the distributional form
with small deviations only. Bottom row: The Poisson model fit ($\hat{\mu} = 3.32$) does not capture
the underlying distribution well ($\mu = 3$, $\theta = 2$), leading to clear deviations in the rootogram.

emphasized by Tukey (e.g, Tukey 1972). Here, it highlights the discrepancy between observed
and expected frequencies. In a sense, hanging rootograms emphasize the fitted values and
suspended rootograms the corresponding residuals. We recommend the hanging version as the
default as long as residuals are not of main concern, and hence employ hanging rootograms
below.

2.2. Interpreting rootograms

In analyses employing rootograms, one is often interested in detecting patterns such as runs
of positive or negative deviations, which highlight aspects of the model fit that might require
further attention. For example, Figure 1 presents rootograms for a Poisson model fitted
to two simulated data sets from a Poisson (top row) and negative binomial (bottom row)
distribution. Both underlying distributions have mean $\mu = 3$, the negative binomial has a
shape parameter $\theta = 2$ (while the Poisson is formally a negative binomial with $\theta = \infty$).

When fitting a Poisson model to the Poisson data, all three versions of the rootogram in the
top row exhibit only small deviations: The standing version shows that the curve representing

expected frequencies closely tracks the histogram representing observed frequencies, there are also no clear patterns in the hanging and suspended versions. All this indicates that the model fits well.

In contrast, when fitting a Poisson model to the negative binomial data in the bottom row there are substantial departures of the model from the data: in the standing version, the curve representing expected frequencies does not track the observed frequencies, there are also discernible patterns in both the hanging and suspended variants. Specifically, for the latter the rootogram bars form a 'wave-like' pattern around the horizontal reference line: the data exhibit too many small counts, notably zeros, as well as too many large counts for a Poisson model to provide an adequate fit. In summary, the patterns encountered in the bottom row of Figure 1 reflect a substantial amount of overdispersion that is not captured by the fitted Poisson distribution.

The patterns seen in the bottom panel of Figure 1 are theoretically supported by results presented by Mullahy (1997), who shows that Poisson mixtures exhibit a larger number of zeros (compared with a Poisson null model) as well as more mass in the upper tails and less mass in the center of the distribution. Mullahy's results rely on earlier work of Shaked (1980), who shows that mixing will generally spread out a distribution (from the exponential family) towards its tails. The negative binomial distribution is a gamma mixture of the Poisson distribution, hence these arguments are directly relevant in the case at hand.

While excess zeros are strictly implied by overdispersion (compare Mullahy 1997, Prop. 1), there also exist situations in practice where the number of zeros is so large that merely correcting for overdispersion via, e.g., a negative binomial model does not solve the problem. These tend to exhibit a spike at zero in graphical displays and are often best treated by fitting a two-part model. We shall encounter an example below.

3. Examples

In this section we present two empirical illustrations. The first revisits a well-known data set from ethology, for which excess zeros require treatment, the second a somewhat larger data set from health economics exhibiting a substantial amount of unobserved heterogeneity, for which finite mixture models are employed. In these illustrations, we select models using information criteria, notably the BIC, and use rootograms for highlighting deficiencies of fitted models.

3.1. Horseshoe Crab Mating

Brockmann (1996) investigates horseshoe crab mating. The crabs arrive on the beach in pairs to spawn. Furthermore, unattached males also come to the beach, crowd around the nesting couples and compete with attached males for fertilizations. These so-called satellite males form large groups around some couples while ignoring others. Brockmann (1996) shows that the groupings are not driven by environmental factors but by properties of the nesting female crabs. Larger females that are in better condition attract more satellites.

Agresti (2013, Chapter 4.3) reanalyzes these data, modeling the number of satellites using count data regression techniques. The main explanatory variable is the female crab's carapace width, but its color and spine condition are also considered in some analyses – with the ordered factors for color and spine condition often treated as numeric variables. In his analysis, Agresti (2013) starts out from a Poisson model with the standard log link and then goes on to consider
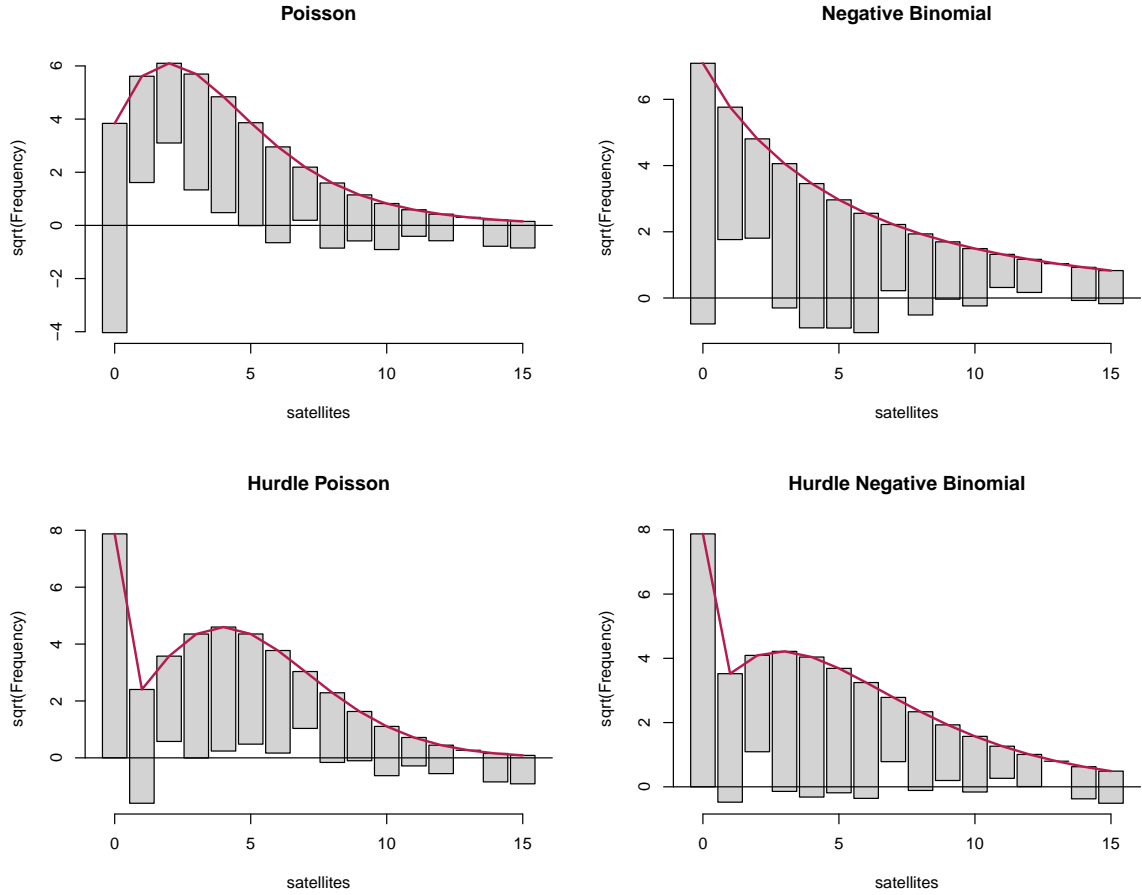
Figure 2:   Hanging rootograms for crab satellite models (counts $0, \ldots, 15$).

both Poisson and negative binomial models with both log and identity links. He finds that among these the negative binomial model fits best but also notes that further refinements might be possible, e.g., by allowing for zero inflation.

To illustrate how rootograms can help in judging the goodness of fit of various count regression models for this data, we extend the analysis of Agresti (2013) in the following way: we consider both Poisson and negative binomial regressions (with log link) and hurdle versions of these (with a logit-type binary part) to allow for excess zeros. The carapace width and a numeric coding of the color variable are used as regressors in all (sub-)models. To compare the relative performances of the four models, we employ the Bayesian information criterion (BIC), yielding: Poisson (BIC = 931.0, df = 3), negative binomial (BIC = 769.5, df = 4), hurdle Poisson (BIC = 755.1, df = 6), and hurdle negative binomial (BIC = 736.8, df = 7). These results already suggest that the hurdle negative binomial model fits best. However, a look at the corresponding hanging rootograms (for counts $0, \ldots, 15$) in Figure 2 provides much more insight into the pros and cons of the various models:

- *Poisson:* The wave-like pattern in the rootogram bars in the top left panel shows that the counts $1, \ldots, 4$ are overfitted while 0 and most counts from 6 onwards are underfitted.

Table 1:   Negative binomial hurdle models for crab satellites. Coefficient estimates (and standard errors in parentheses).

| | Hurdle NB, model 1 | | Hurdle NB, model 2 | |
	count	zero	count	zero
(Intercept)	0.429	−10.071	1.465	−10.071
	(0.941)	(2.806)	(0.068)	(2.806)
width	0.038	0.458		0.458
	(0.033)	(0.104)		(0.104)
color	0.007	−0.509		−0.509
	(0.091)	(0.224)		(0.224)
Log(theta)	1.527		1.495	
	(0.353)		(0.349)	
N	173		173	
Log-likelihood	−350.363		−351.033	
AIC	714.726		712.066	
BIC	736.799		727.832	

This indicates a substantial amount of overdispersion in the data, the clear lack of fit for 0 could be an additional indication of excess zeros.

- *Negative binomial:* The rootogram does no longer exhibit the wave-like pattern of the Poisson model, showing that the overdispersion is accounted for much better in this model. However, the underfitting of the count 0 and clear overfitting for counts 1 and 2 is typical for data with excess zeros. Note also that the fitted negative binomial model implies a decreasing probability mass function, which is not in line with the data structure.

- *Hurdle Poisson:* The rootogram now shows a perfect fit for the count 0 (by design of the hurdle model). However, there is still overdispersion in the remaining positive counts that is again reflected by a wave-like pattern, note also the clear underfitting of the count 1.

- *Hurdle negative binomial:* The rootogram shows that this model fits the data quite well. There are no clear patterns of departure anymore and the deviations between observed and predicted frequences are very small for most of the counts.

The parameter estimates (standard errors in parentheses) for the negative binomial hurdle model are reported in the first two columns of Table 1. Interestingly, this reveals that the female crab's carapace width and color both clearly affect the probability of having any satellites (binary zero hurdle part of the model). Specifically, larger crabs are much more likely to have satellites. However, given that there is at least one satellite neither carapace width nor color are individually significant (zero-truncated count part of the model). In fact, both variables could also be omitted from the count part, resulting in the model labeled 'hurdle NB, model 2' in Table 1. This improves the fit in terms of both AIC and BIC. As the fitted models do not differ significantly, the rootogram of the simplified model is essentially identical to the rootogram of the full hurdle model and is therefore omitted here.
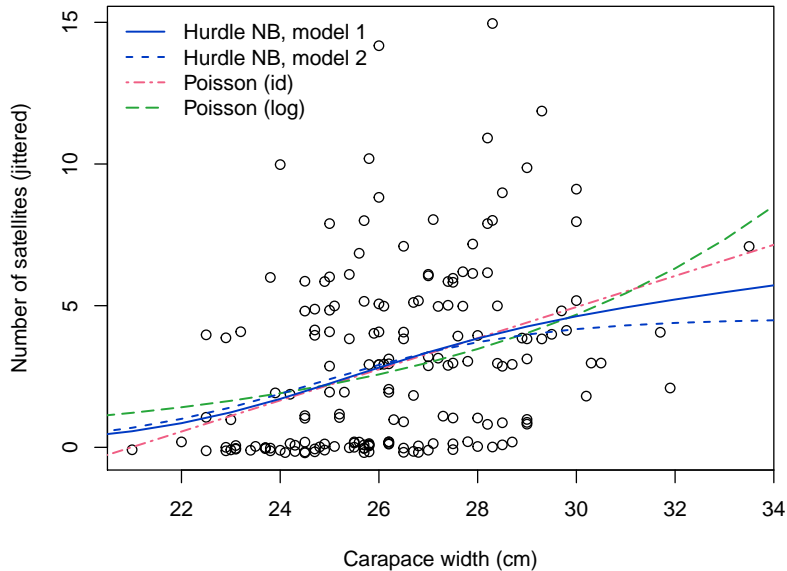
Figure 3: Predicted effect for the mean number of satellite at increasing carapace width and mean color.

Additionally, identity (rather than log) links or a zero-inflation (rather than hurdle) specification could also be employed but are omitted here for compactness. Both lead to qualitatively identical insights and similar patterns in the rootograms while neither leads to improvements over the negative binomial hurdle model. Hence, these are not presented in more detail here. Instead, we conclude with a comparison of predicted effects for the mean function from several models. Figure 3 shows the effects on the mean number of satellites for increasing carapace width at the mean color ($= 2.5$ in the center of the scale $1, \ldots, 4$). This shows that, compared to the identity link model preferred by Agresti (2013), the hurdle model leads to very similar predictions at average widths while avoiding negative predictions for small widths and at the same time increasing even more slowly for large widths. This complements the findings from the rootograms and underlines that the hurdle model fits the data rather well. (The fitted mean function for the hurdle Poisson model looks virtually identical to that of the hurdle negative binomial model and is hence omitted from Figure 3.)

3.2. Demand for Medical Care

Our second example uses cross-sectional data originating from the US National Medical Expenditure Survey (NMES) conducted in 1987 and 1988. The NMES is based upon a representative, national probability sample of the civilian non-institutionalized population and of individuals admitted to long-term care facilities during 1987. The subsample used here comprises only individuals aged 66 and over, all of whom are covered by Medicare (a public insurance program providing substantial protection against health-care costs). For R users, these data are conveniently available from the **AER** package supplementing Kleiber and Zeileis
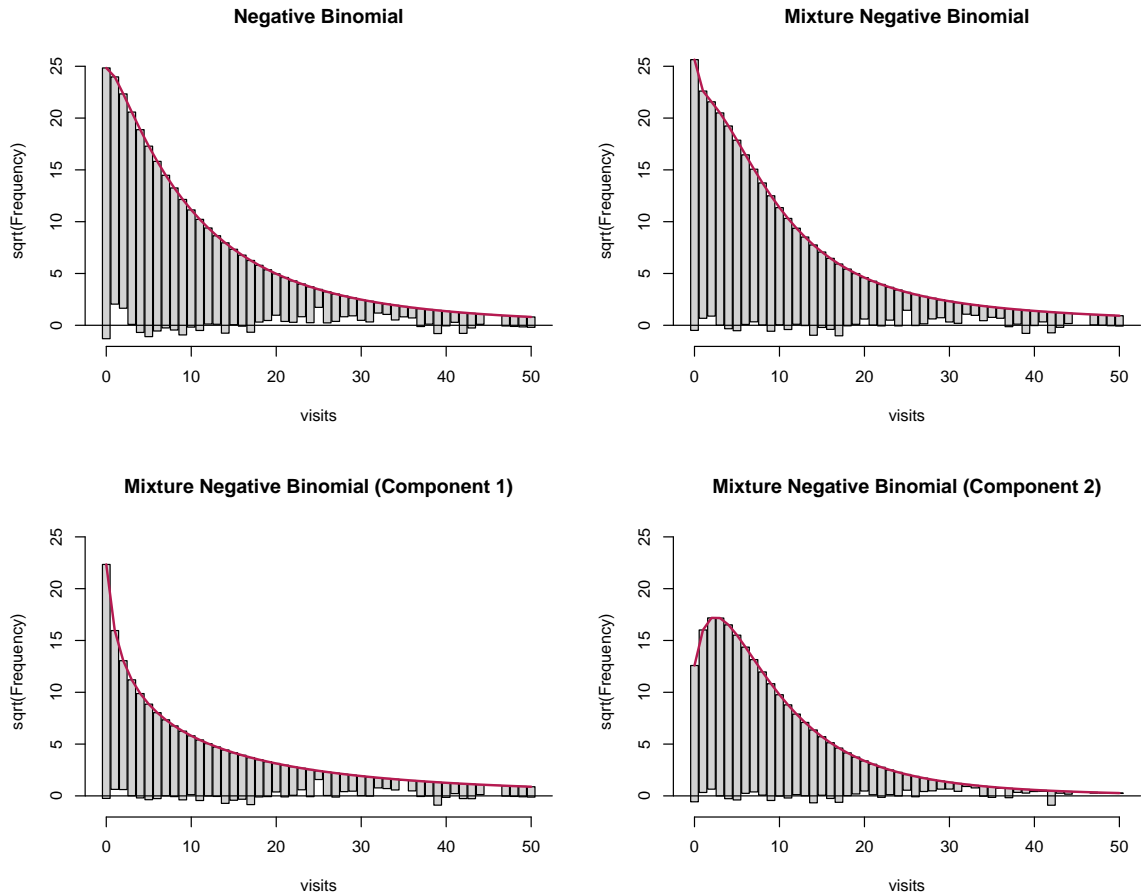
Figure 4: Hanging rootograms for NMES 1988 models.

(2008) under the name `NMES1988`. They have been explored originally by Deb and Trivedi (1997) using finite mixtures of count data regressions. Zeileis, Kleiber, and Jackman (2008) employ the data for illustration of hurdle and zero-inflation models while Cameron and Trivedi (2013) reinvestigate finite mixtures. Here, we follow the latter approach but employ a slightly reduced set of regressors to facilitate interpretation while still obtaining reasonably good fits.

Figure 4 displays the rootograms for a single negative binomial regression as well as for a finite mixture of two negative binomial regressions. For the latter, the mixture model (upper panel, right) as well as both components (lower panel) are given. The corresponding parameter estimates (standard errors in parentheses) as well as the sums of the posterior weights (denoted N) are reported in Table 2.

The single NB regression clearly misfits, especially for the low counts $0, 1, 2$, while the mixture NB provides an improved fit. It is possible to study the mixture model in more detail by decomposing observed and expected frequencies into the individual components and visualizing them separately. To this end, the observed and expected frequencies are computed as weighted sums using the posterior probabilities for each component. Figure 4 (bottom panels) highlights nicely that both components fit rather well. It also brings out the different means and variances in the two components. Specifically, the first component contains a fraction of

Table 2:   Negative binomial regression models (single and 2-component finite mixture) for NMES 1988 physician office visits. Coefficient estimates (and standard errors in parentheses).

	Single	Component 1	Component 2
(Intercept)	0.801	−0.961	1.464
	(0.062)	(0.395)	(0.145)
health: poor/average	0.345	0.402	0.286
	(0.049)	(0.143)	(0.061)
health: excellent/average	−0.379	−0.219	−0.452
	(0.062)	(0.162)	(0.099)
chronic	0.192	0.272	0.160
	(0.012)	(0.041)	(0.019)
gender: male/female	−0.094	−0.160	−0.076
	(0.032)	(0.095)	(0.048)
school	0.030	0.065	0.010
	(0.004)	(0.020)	(0.009)
insurance: yes/no	0.353	1.700	−0.097
	(0.044)	(0.316)	(0.097)
medicaid: yes/no	0.307	0.805	0.171
	(0.062)	(0.257)	(0.080)
Log(theta)	0.159	−0.406	0.899
N	4406	1744.866	2661.134
Log-likelihood	−12215.009	−12149.842	
AIC	24448.019	24337.684	
BIC	24505.535	24459.108	

$0.396 = 1744.866/4406$ of all observations and is characterized by a zero-modal rootogram. On average, the corresponding individuals have fewer physician office visits but at the same time a rather high variance. The parameter estimates are mostly larger (in absolute values) than in the second component, especially for the insurance and medicaid parameters. In contrast, the second component is characterized by a unimodal rootogram with comparatively lighter tails. On average, the corresponding individuals have more physician office visits but at the same time a smaller variance. The first group may, therefore, be seen as the group of occasional users, for which the number of visits likely depends on the severity of the issues, while the second group may be seen as the group of regular users, for which the number of visits often results from the presence of chronic conditions. Indeed, when splitting the patients into two clusters (with hard assignment to the clusters according to the highest posterior probability), it can be seen that the second cluster has a lower proportion of persons with excellent health status (10.1% vs. 7.1%), or without chronic diseases (34.4% vs. 19.8%), and a higher proportion of insured persons (64.7% vs. 81.7%). Moreover, further unobserved factors such as the type of diseases and medication might be captured by the two latent components.

4. Conclusions

Various flavors of rootograms are discussed as graphical diagnostic tools for visualizing complex regression models for count data, such as two-part hurdle models or finite mixture models. They combine exploratory data analysis with model-based inference by bringing out discrepancies between observed and fitted distributions (see also Gelman 2004, for a Bayesian framework employing posterior predictive checks). Unlike other model-based graphics that often focus on effects on the mean of the fitted distribution (e.g., effect displays, Fox 2003; Fox and Hong 2009), rootograms capture deviations across the support of the entire distribution and hence can help diagnosing misfit regarding scatter and/or shape. This is particularly relevant for count data models, which are often affected by problems such as overdispersion and/or excess zeros.

We have presented two examples involving excess zeros and unobserved heterogeneity and, guided by features of rootograms corresponding to various fitted models, settled on a hurdle and a finite mixture model, respectively. A further example involving the less frequent case of underdispersion is available from the R package **countreg** (Zeileis and Kleiber 2014). See the `TakeoverBids` data there and the appendix below for more information on the package.

Computational Details

Our results were obtained using R 3.1.2 (R Core Team 2014) with the packages **countreg** 0.1-2 (Zeileis and Kleiber 2014; Zeileis *et al.* 2008), **MASS** 7.3-35 (Ripley 2014; Venables and Ripley 2002), and **flexmix** 2.3-12 (Leisch 2004; Grün and Leisch 2008). See `help("CrabSatellites", package = "countreg")` and `help("FLXMRnegbin", package = "countreg")` for more details and replication code.

References

Agresti A (2013). *Categorical Data Analysis.* 3rd edition. John Wiley & Sons, Hoboken, NJ.

Brockmann HJ (1996). "Satellite Male Groups in Horseshoe Crabs, *Limulus polyphemus.*" *Ethology*, **102**(1), 1–21.

Cameron AC, Trivedi PK (1990). "Regression-Based Tests for Overdispersion in the Poisson Model." *Journal of Econometrics*, **46**, 347–364.

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

Dean CB (1992). "Testing for Overdispersion in Poisson and Binomial Regression Models." *Journal of the American Statistical Association*, **87**(418), 451–457.

Deb P, Trivedi PK (1997). "Demand for Medical Care by the Elderly: A Finite Mixture Approach." *Journal of Applied Econometrics*, **12**(3), 313–336.

Fox J (2003). "Effect Displays in R for Generalised Linear Models." *Journal of Statistical Software*, **8**(15), 1–27. URL http://www.jstatsoft.org/v08/i15/.

Fox J, Hong J (2009). "Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the **effects** Package." *Journal of Statistical Software*, **32**(1), 1–24. URL http://www.jstatsoft.org/v32/i01/.

Friendly M (2000). *Visualizing Categorical Data.* SAS Insitute, Cary, NC. URL http://www.datavis.ca/books/vcd/.

Friendly M, Denis DJ (2001). "Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization." URL http://www.datavis.ca/milestones/.

Gelman A (2004). "Exploratory Data Analysis for Complex Models." *Journal of Computational and Graphical Statistics*, **13**(4), 755–779.

Grün B, Leisch F (2008). "FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters." *Journal of Statistical Software*, **28**(4), 1–35. URL http://www.jstatsoft.org/v28/i04/.

Healy MJR (1968). "The Disciplining of Medical Data." *British Medical Bulletin*, **24**(3), 210–214.

Kleiber C, Zeileis A (2008). *Applied Econometrics with R.* Springer-Verlag, New York.

Lambert D (1992). "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics*, **34**(1), 1–14.

Leisch F (2004). "FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R." *Journal of Statistical Software*, **11**(8), 1–18. URL http://www.jstatsoft.org/v11/i08/.

Mullahy J (1986). "Specification and Testing of Some Modified Count Data Models." *Journal of Econometrics*, **33**(3), 341–365.

Mullahy J (1997). "Heterogeneity, Excess Zeros, and the Structure of Count Data Models." *Journal of Applied Econometrics*, **12**(3), 337–350.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org/.

Rigby RA, Stasinopoulos DM (2005). "Generalized Additive Models for Location, Scale and Shape." *Journal of the Royal Statistical Society C*, **54**(3), 507–554.

Ripley BD (2014). **MASS**: *Support Functions and Datasets for Venables and Ripley's MASS.* R package version 7.3-33, URL http://CRAN.R-project.org/package=MASS.

Shaked M (1980). "On Mixtures from Exponential Families." *Journal of the Royal Statistical Society B*, **42**(2), 192–198.

Stasinopoulos DM, Rigby RA (2007). "Generalized Additive Models for Location, Scale and Shape (GAMLSS) in R." *Journal of Statistical Software*, **23**(7), 1–46. URL http://www.jstatsoft.org/v23/i07/.

Tukey JW (1965). "The Future of Processes of Data Analysis." In *Proceedings of the 10th Conference on the Design of Experiments in Army Research, Development and Testing*, pp. 691–729. Army Research Office, Durham, NC. Reprinted in Lyle V. Jones (ed.) *The Collected Works of John W. Tukey, Volume IV. Philosophy and Principles of Data Analysis: 1965–1986*, Wadsworth & Brooks/Cole, Monterey, CA, 1986.

Tukey JW (1972). "Some Graphic and Semigraphic Displays." In TA Bancroft (ed.), *Statistical Papers in Honor of George W. Snedecor*, pp. 293–316. Iowa State University Press, Ames, IA. Reprinted in William S. Cleveland (ed.): *The Collected Works of John W. Tukey, Volume V. Graphics: 1965–1985*, Wadsworth & Brooks/Cole, Pacific Grove, CA, 1988.

Tukey JW (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, MA. ISBN 0-201-07616-0.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S.* 4th edition. Springer-Verlag, New York.

Wainer H (1974). "The Suspended Rootogram and Other Visual Displays: An Empirical Validation." *The American Statistician*, **28**(4), 143–145.

Wickham H (2009). ***ggplot2**: Elegant Graphics for Data Analysis.* Springer-Verlag, New York.

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

Zeileis A, Kleiber C (2014). ***countreg**: Count Data Regression.* R package version 0.1-2/r88, URL http://R-Forge.R-project.org/projects/countreg/.

Zeileis A, Kleiber C, Jackman S (2008). "Regression Models for Count Data in R." *Journal of Statistical Software*, **27**(8), 1–25. URL http://www.jstatsoft.org/v27/i08/.

A. R Implementation

For an overview of count data regression models in R we refer to Zeileis *et al.* (2008), where R implementations of hurdle and zero-inflation models are described in some detail. The corresponding fitting functions have now been moved to the **countreg** package, a new package that is currently under development by the authors of the present paper. First versions are already available from http://R-Forge.R-project.org/projects/countreg/.

The current implementation of rootograms in **countreg** provides a generic function `rootogram(object, ...)` along with several methods for different types of models/data. The methods all proceed in the same way: They first compute the observed and expected frequencies, obs_j and exp_j respectively (see Section 2), and then call the default method that computes all required coordinates for drawing the rootograms. The latter has the following arguments:

```
rootogram(object, fitted, breaks = NULL,
  style = c("hanging", "standing", "suspended"),
  scale = c("sqrt", "raw"), plot = TRUE,
  width = NULL, xlab = NULL, ylab = NULL, main = NULL, ...)
```

The arguments `object` and `fitted` need to provide the tables/vectors of observed and fitted frequencies. (The first argument is called `object` rather than `observed` for consistency with the generic function that only takes one required `object` argument and `....`) The `breaks` need to be specified if a continuous distribution is employed while for a discrete distribution one may want to set the `width` of the bars to leave small gaps between the bars (as in our examples). Additionally, one of three `style`s can be specified: `"hanging"` (default), `"standing"`, or `"suspended"`. The object returned is then a 'data.frame' with all the coordinates needed for plotting, and this is also drawn directly by default (`plot = TRUE`) along with the specified graphical arguments (`xlab`, `ylab`, `main`, `...`). By default, the base graphics `plot()` method is used for drawing rootograms. In addition, there is also an `autoplot()` method for drawing rootograms using the **ggplot2** package (Wickham 2009).

Above we used methods for objects of classes 'glm' and 'hurdle'. There are further methods available, currently for univariate distributions fitted via `fitdistr()` (to objects of class 'numeric', Venables and Ripley 2002), zero-inflated models (objects of class 'zeroinfl', Zeileis *et al.* 2008), zero-truncated models (objects of class 'zerotrunc', as fitted by the `zerotrunc()` function in **countreg**), generalized additive models (objects of class 'gam', Wood 2006), and for selected count distributions falling within the framework of generalized additive models for location, scale and shape (objects of class 'gamlss', Rigby and Stasinopoulos 2005; Stasinopoulos and Rigby 2007).

Affiliation:

Christian Kleiber
Faculty of Business and Economics
Universität Basel
Peter Merian-Weg 6
4002 Basel, Switzerland
E-mail: Christian.Kleiber@unibas.ch
URL: http://wwz.unibas.ch/kleiber/

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