A diagnostic plot for assessing model fit in count data models

Jochen Einbeck[1], Paul Wilson[2]

[1] Department of Mathematical Sciences, Durham University, UK
[2] School of Mathematics and Computer Science, University of Wolverhampton, UK

E-mail for correspondence: jochen.einbeck@durham.ac.uk

Abstract: Whilst many numeric methods, such as AIC and deviance, exist for assessing model fit, diagrammatic methods are few. We present here a diagnostic plot, to which we refer as 'Christmas tree plot' due its characteristic shape, that may be used to visually assess the suitability of a given count data model.

Keywords: Diagnostic plot, model fit, count data.

1 Introduction

Consider univariate count data $Y_1, \ldots, Y_n$, which are supposedly distributed according to some count distribution $F(\mu_i, \theta)$, with mean parameters $\mu_i = E(Y_i|x_i)$ possibly depending on covariates x_i (which may be vector–valued). We assume that a routine to obtain estimates $\hat{\mu}_i = \hat{E}(Y_i|x_i)$ and $\hat{\theta}$ is readily available, and we are interested in assessing graphically the quality of the resulting model fit. The idea is to check whether, for each count k, the number $N(k)$ of observed counts k is consistent with the suspected count distribution F. More precisely, denote $p_i(k) = P(k|\mu_i, \theta)$ the probability of observing the count k under covariate x_i and model F, which can be estimated by $\hat{p}_i(k) = P(k|\hat{\mu}_i, \hat{\theta})$ from the fitted model. For instance, in the special case that $F(\mu_i, \theta)$ corresponds to Pois(μ_i), one has $\hat{p}_i(k) = \exp(-\hat{\mu}_i)\hat{\mu}_i^k/k!$. This scenario is discussed in Wilson and Einbeck (2015, 2016) with focus on the case $k = 0$. This abstract generalizes those ideas to general k and F and proposes a generic diagrammatic tool.
The random variable $N(k)$ follows a Poisson–Binomial distribution with parameters $p_1(k), \ldots, p_n(k)$ (Chen and Liu, 1997). Hence, for any choice of k and F, a range of plausible values of $N(k)$ can be obtained by confidence

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016.

intervals from this distribution, which can be computed using the R package `poibin` (Hong, 2013). By doing this for a range of values of k, one can draw diagrams which give envelopes for plausible values of $N(k)$ which can then be compared to the true values. Since these diagrams resemble Christmas trees, we refer to them as 'Christmas tree plots' from now on. We explain the construction of the diagram in systematic form in the next section, and give examples in the final sections.

2 The Christmas tree plot

For count data $Y = (Y_1, \dots Y_n)$, we will typically be interested in the range of counts $K = [0, \max(Y)]$, though in some applications, where very small counts are not to be expected, one may prefer using $K = [\min(Y), \max(Y)]$. Denote the chosen range by $K = [k_a, k_b]$. We construct the diagnostic plot as follows.

(i) Fit the model $F(\mu_i, \theta)$ to the data Y.

(ii) For k in $k_a ... k_b$, obtain estimates $\hat{p}_i(k)$. Use a Poisson-Binomial distribution to estimate the median $m(k) = \text{med}(N(k))$ under count data model F, as well as lower and upper limits, say $\underline{c}_\alpha(k)$ and $\bar{c}_\alpha(k)$ of a $(1-\alpha)\%$ confidence interval for $N(k)$.

(iii) Compute the median–adjusted bounds $\underline{b}_\alpha(k) = \underline{c}_\alpha(k) - m(k)$ and $\bar{b}_\alpha(k) = \bar{c}_\alpha(k) - m(k)$.

(iv) Plot the functions $\underline{b}_\alpha(k)$ and $\bar{b}_\alpha(k)$ versus k.

(v) Add to the plot the observed adjusted counts, $A(k) = N(k) - m(k)$ of the observed data Y.

If the data is consistent with the distribution fitted, the curve $A(k)$ should (largely) stay within the adjusted bands $\underline{b}_\alpha(k)$ and $\bar{b}_\alpha(k)$. If the data is *not* consistent with the distribution fitted then $A(k)$ is likely not stay within these bands. Additionally, when interpreting the bands as a measure of typical variation of $N(k)$, we can use this plot to diagnose whether the counts exhibit less random variation than expected under model F.
One may argue that due to the consideration of a sequence of confidence intervals for $k_a ... k_b$ one has to account for multiple testing issues. It should be stressed, however, that we do not consider the proposed plot as a *testing* procedure, but as a simple diagrammatic tool which supports the data analyst in identifying potential model inadequacies, similar in spirit to a QQ plot.

TABLE 1. Simulated data with upper and lower confidence intervals for $N(k)$ and $A(k)$.

k	$N(k)$	$\underline{c}_{0.1}(k)$	$\bar{c}_{0.1}(k)$	$m(k)$	$A(k)$	$\underline{b}_{0.1}(k)$	$\bar{b}_{0.1}(k)$
0	38	19	33	26	12	-7	7
1	28	27	43	35	-7	-8	8
2	15	17	31	24	-9	-7	7
3	7	6	16	10	-3	-4	6
4	8	1	7	3	5	-2	4
5	1	0	3	1	0	-1	2
6	2	0	1	0	2	0	1
7	1	0	0	0	1	0	0

3 Simulation example

Consider a covariate–free data set of size $n = 100$ drawn from a zero-inflated Poisson (ZIP) distribution with Poisson parameter 1.5 and zero-inflation parameter 0.2, that is overall mean equal to 1.2. The data are given in terms of $N(k)$ in the 2nd column of Table 1. Following the procedure outlined in Section 2 with $F \sim \text{Pois}(\mu)$ yields 90% confidence intervals for $N(k)$ (displayed in the 3rd and 4th column of Table 1), resulting in the Christmas tree plot displayed in the left hand panel of Figure 1. This plot indicates that the Poisson model is not suitable, as visible by the number of zero-observations falling well above the upper confidence band, as well as by the adjusted count $A(2)$ falling below the lower band. The right hand plot is constructed similar to that of the left, except that here the zero-inflated Poisson (ZIP) model serves as model F. Clearly this plot indicates that a ZIP model is suitable for the data.

4 Application on biodosimetry data

We consider data consisting of $n = 14430$ chromosome aberration counts previously studied by Oliveira et al. (2016). The covariate *dose*, with values between 0 and 4.5Gy, gives the radiation dose applied to blood sample cells, causing DNA damage in form of double–strand breaks. When incorrectly repaired by the cellular DNA–damage response mechanism, this can lead to dicentric chromosomes which can be counted under a microscope. That is, each examined blood sample cell contributes, for known covariate dose, exactly one count observation. For this data set, the counts take values in the range from 0 to 5. Data of this type have been fitted traditionally through Poisson regression models, though the presence of excess zero counts has been regularly reported in the literature.

FIGURE 1. Christmas tree plots for simulated covariate–free data. The dashed curve corresponds to $A(k)$ and the dotted curves give the median–adjusted bounds.

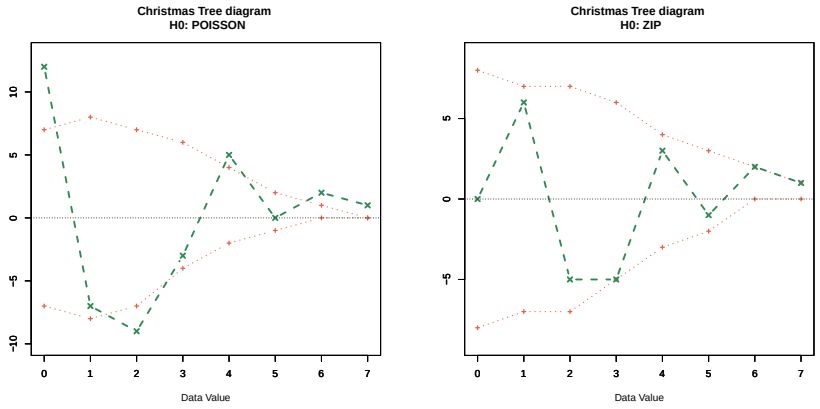


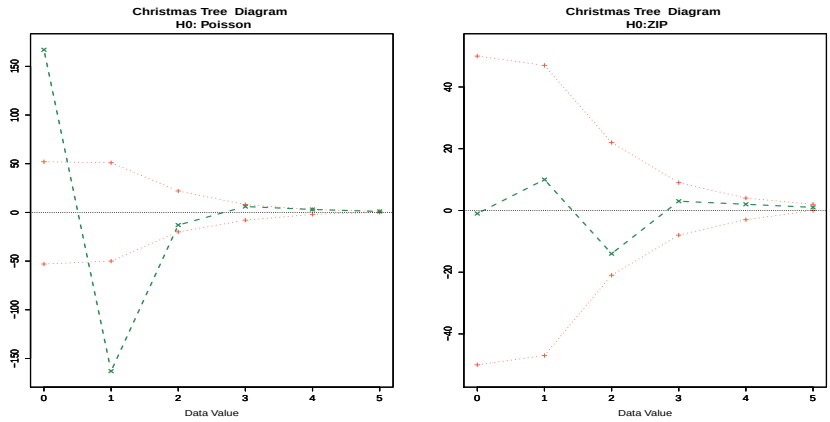
Table 2 displays the data under investigation, and Figure 2 contains the Christmas Tree diagrams obtained when Poisson and zero–inflated Poisson models, using a log–link and quadratic polynomial for dose, are fitted to these data. The left hand plot clearly indicates the unsuitability of the Poisson model, whereas the right hand plot indicates that ZIP is suitable. Oliveira et al. (2016) carried out an extensive analysis of this data set, applying several statistical tests and model selection criteria in order to decide for an adequate modelling strategy. Specifically, they found that a negative binomial type 2 model returned the lowest AIC (7489.1), closely followed by a ZIP model (AIC=7490.4). Other models considered included the Poisson as reference model (AIC=7504.7), and a Poisson Inverse Gaussian (AIC=7495.2).

The two plots in Figure 3 corresponding to the NB2 and PIG models, respectively, illustrate cases where the adjusted observed data line, $A(k)$, remains close to the centre line. For the NB2, all observations lie between the 43rd and 57th quantiles of their respective Poisson–Binomial distribution. Hence, there is less random variation amongst observed counts than would be expected under NB2, most likely indicating that the variance of the fitted model is inflated in order to accommodate the number of observed zeros. A similar effect is observed for the PIG model. In summary, these plots suggest that the ZIP model is the most adequate model for these data, deviating from what would be concluded by looking at a single–number model selection criterion such as AIC.

TABLE 2. Frequency of dicentric chromosomes after acute whole body *in vitro* exposure to doses between 0 and 4.5Gy of Cobalt-60 γ-rays. (This corresponds to data set A1 in the notation of Oliveira et al. (2016), where also the reference for the data source is provided.)

	Frequency of counts					
dose	0	1	2	3	4	5
0.00	2591	1	0	0	0	0
0.25	2185	8	0	0	0	0
0.75	2550	44	1	0	0	0
1.00	2231	54	2	0	0	0
1.50	1712	96	3	0	0	0
2.50	1196	123	7	1	0	0
3.00	1070	320	41	6	1	0
4.50	895	360	110	25	5	1

FIGURE 2. Christmas tree plots for biodosimetry data, with the hypothesized distribution F corresponding to Poisson and ZIP, respectively.



References

Chen, S.X. and Liu, J.S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875–892.

Hong, Y. (2013). poibin: The Poisson Binomial Distribution. R package version 1.2. https://CRAN.R-project.org/package=poibin

Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P. and Rothkamm, K.

(2016). Zero-inflated regression models for radiation–induced chromosome aberration data: A comparative study. *Biometrical Journal* **58**, 259-279.

Wilson, P. and Einbeck, J. (2015). A simple and intuitive test for number–inflation or number–deflation. In: Wagner, H. and Friedl, H. (Eds). Proc's of the 30th IWSM, Linz, Austria, Vol 2, pages 299–302.

Wilson, P. and Einbeck, J. (2016). On statistical testing and mean parameter estimation for zero–modification in count data regression. Proc's of the 31st IWSM, Rennes, France, *to appear*.

FIGURE 3. Christmas tree plots for biodosimetry data, with the hypothesized distribution F corresponding to NB2 and PIG, respectively.