Biometrical Journal **0** (0) 0, zzz–zzz / DOI: 10.1002/

Beyond Mean Modelling: Bias due to Misspecification of Dispersion in Poisson-inverse Gaussian Regression

Gillian Z Heller[*,1], **Dominique-Laurent Couturier** [2], and **Stephane R Heritier** [3]

[1] Department of Statistics, Macquarie University, Sydney, Australia

[2] Cancer Research UK – Cambridge Institute, University of Cambridge, Cambridge, UK

[3] School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

Received zzz, revised zzz, accepted zzz

In clinical trials one traditionally models the effect of treatment on the mean response. The underlying assumption is that treatment affects the response distribution through a mean location shift on a suitable scale, with other aspects of the distribution (shape/dispersion/variance) remaining the same. This work is motivated by a trial in Parkinson's disease patients in which one of the endpoints is the number of falls during a ten-week period. Inspection of the data reveals that the Poisson-inverse Gaussian (PiG) distribution is appropriate, and that the experimental treatment reduces not only the mean, but also the variability, substantially. The conventional analysis assumes a treatment effect on the mean, either adjusted or unadjusted for covariates, and a constant dispersion parameter. On our data, this analysis yields a non-significant treatment effect. However, if we model a treatment effect on both the mean and dispersion parameters, both effects are highly significant. A simulation study shows that if a treatment effect exists on the dispersion and is ignored in the modelling, estimation of the treatment effect on the mean can be severely biased. We show further that if we use an orthogonal parametrization of the PiG distribution, estimates of the mean model are robust to misspecification of the dispersion model. We also discuss inferential aspects that are more difficult than anticipated in this setting. These findings have implications in the planning of statistical analyses for count data in clinical trials.

Key words: Poisson-inverse Gaussian regression, dispersion modelling, parameter orthogonality, profile likelihood confidence interval, count data

1 Introduction

This work was motivated by the analysis of a clinical trial in which a secondary outcome was the number of falls that participants experienced while undergoing a drug treatment or usual care (control group) (Hauser et al., 2016). The study is described in Section 3. Based on a blinded inspection of the data, the Poisson-inverse Gaussian (PiG) distribution provided the best marginal fit, so was chosen as the response distribution. On unblinding it became clear that dispersion of falls varied considerably with treatment group, and accordingly a model for dispersion was appropriate. Regression models which extend modelling to dispersion and other shape parameters, and to distributions beyond the exponential family, have become popular in the last decade (Rigby and Stasinopoulos, 2005; Kneib, 2013). While interest is generally still focussed on modelling the mean, these models allow the flexibility of modelling shape parameters as a function of covariates. The PiG, parametrized in terms of its mean μ and a dispersion parameter σ, is available as a response distribution in existing regression software (Stasinopoulos et al., 2017) and parameter estimates are easily obtained. However, in our study, estimates of the treatment effect on the mean μ were found to be sensitive to specification of the model for σ. This sensitivity of the mean model to the dispersion model is of particular concern in the context of clinical trials, as statistical analysis plans do not in general specify modelling of a dispersion parameter. A regression model using an alternative parametrization of the PiG distribution, in which the shape parameter (α) is orthogonal to the mean, was

*Corresponding author: e-mail: gillian.heller@mq.edu.au, Phone: +61-2-98508541

considered. For our study, the estimate of the treatment effect on the mean was robust to the model for α. In a simulation study it was confirmed that, for the (μ, σ) parametrization, estimates in the μ model can be severely biased if the σ model is misspecified. However, using the (μ, α) parametrization, μ model estimates are robust to misspecification of the α model. This potentially has implications not only for PiG regression, but for regression models for any response distribution, in which the shape parameter(s) being modelled are not orthogonal to the mean.

An incidental but important finding in this work was that inference for PiG regression is problematic, in that standard confidence intervals (Wald, sandwich, bootstrap) do not reach their nominal coverage even for large sample sizes, while the profile likelihood confidence interval, based on inversion of the likelihood ratio test, produces accurate coverage.

In Section 2 we introduce the PiG distribution in non-orthogonal and orthogonal parametrizations, and corresponding regression models; in Section 3 we describe and analyse the data set that motivated this study; in Section 4 we discuss inference for PiG regression; and in Section 5 we describe two simulation studies which investigated parameter estimation and inference.

2 Models for overdispersed count data

Accounting for overdispersion in count data by applying a mixing distribution to the Poisson mean parameter dates back to Greenwood and Yule (1920), who used the gamma as Poisson mixing distribution to obtain the negative binomial (NB) distribution. Several other Poisson mixture distributions have been created in this way, including the Poisson-generalized inverse Gaussian distribution (Sichel, 1971), which is particularly flexible in modeling long-tailed discrete data. This distribution is, however, rather unattractive computationally. Its two-parameter special case, the Poisson-inverse Gaussian (PiG) distribution, is more tractable and typically useful in accommodating data with tails longer than the negative binomial (Stein et al., 1987; Dean et al., 1989). Rigby et al. (2008) give a number of other Poisson mixture distributions. In their framework, all of the distributions have parameters μ and σ, where $E(Y) = \mu$ and $Var(Y) = \mu(1 + \sigma\mu)$. Thus σ may conveniently be interpreted as a Poisson overdispersion parameter. Using this parametrization, the inverse Gaussian mixing distribution for the Poisson parameter λ is

$$f_\lambda(\lambda \,|\, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2\lambda^3}} \exp\left\{ -\frac{(\lambda-\mu)^2}{2\mu^2\sigma^2\lambda} \right\} \qquad \lambda > 0\,,$$

which we denote as $\lambda \sim \text{IG}(\mu, \sigma)$. This gives the $\text{PiG}(\mu, \sigma)$ probability function

$$f(y \,|\, \mu, \sigma) = \sqrt{\frac{2}{\pi\sigma}}(1+2\mu\sigma)^{\frac{1}{4}} e^{\frac{1}{\sigma}} \frac{\left(\mu/\sqrt{1+2\mu\sigma}\right)^y}{y!} K_{y-0.5}\left(\sqrt{1+2\mu\sigma}/\sigma\right) \qquad y = 0, 1, 2, \ldots \tag{1}$$

where $K_\nu(x)$ is a modified Bessel function of the third kind. Dean et al. (1989) originally used this parametrization of the PiG (with $\tau = \sigma$). The Poisson is the limiting distribution as σ approaches zero. An advantage of this parametrization is that it can be expressed as a multiplicative random effect model. Indeed, if

$$Y \sim \text{Poisson}(t\mu\delta) \quad \text{and} \quad \delta \sim \text{IG}(1, \sigma), \tag{2}$$

then $Y \sim \text{PiG}(t\mu, \sigma)$ where t denotes the model offset.

The PiG has appeared in the literature in other parametrizations. The distribution was originally proposed by Sichel (1971) with parameters α and λ, where $\alpha > 0$, $0 < \lambda < 1$, $E(Y) = \alpha\lambda/2\sqrt{1-\lambda}$ and $Var(Y) = \alpha\lambda(2-\lambda)/(4(1-\lambda)^{1.5})$. This parametrization has the disadvantage that the asymptotic correlation of the MLEs $\hat{\alpha}$ and $\hat{\lambda}$ is strongly negative; to overcome this, Stein et al. (1987) proposed an orthogonal parametrization (μ, α):

$$f(y\,|\,\mu,\alpha) = \sqrt{\frac{2\alpha}{\pi}}\exp\left(\sqrt{\mu^2+\alpha^2}-\mu\right)\frac{\left(\mu\left(\sqrt{\mu^2+\alpha^2}-\mu\right)/\alpha\right)^y}{y!}K_{y-0.5}(\alpha) \tag{3}$$

having $\mu > 0$, $\alpha > 0$, $E(Y) = \mu$ and $Var(Y) = \mu\left(1+\mu/\left((\mu^2+\alpha^2)^{0.5}-\mu\right)\right)$. In this form, the variance has an inverse relationship with α and the Poisson is the limiting case as α goes to infinity. To our knowledge, this parametrization cannot be expressed as a multiplicative random effect model as seen in (2). It should be noted that the relationship between α and σ is

$$\alpha = \frac{\sqrt{1+2\mu\sigma}}{\sigma} \qquad \text{or} \qquad \sigma = \frac{1}{\sqrt{\mu^2+\alpha^2}-\mu}\,.$$

The notion of parameter orthogonality (Huzurbazar, 1950; Cox and Reid, 1987) means, for a two-parameter distribution $f(y\,|\,\mu,\alpha)$,

$$E\left(\frac{\partial^2}{\partial\mu\,\partial\alpha}\log f\right) = 0\,, \tag{4}$$

resulting in the MLEs $\hat{\mu}$ and $\hat{\alpha}$ being asymptotically independent. In the regression context, it is usual to express response distributions in terms of the mean μ and a shape parameter; however, the choice of shape parameter is by no means unique. In the exponential family, the mean and exponential dispersion parameter ϕ are orthogonal (Barndorff-Nielsen, 1978), with the result that generalized linear modelling is based on orthogonal parametrization, as long as the shape parameter is taken as the exponential dispersion parameter, or a differentiable function of it. For example, the Gamma distribution has the exponential family parametrization:

$$f(y\,|\,\mu,\phi) = \frac{1}{\Gamma(1/\phi)y}\left(\frac{y}{\mu\phi}\right)^{1/\phi}\exp\left(-\frac{y}{\mu\phi}\right)$$

where $E(Y) = \mu$, $Var(Y) = \phi\mu^2$ and ϕ is the exponential dispersion parameter. In the R `glm` function (R Core Team, 2014) and SPSS (IBM Corp, 2013), the dispersion parameter ϕ is estimated; in SAS `proc genmod` (SAS Institute Inc., 2011), $\nu = 1/\phi$ is the shape parameter; and in the R package `gamlss` (Stasinopoulos et al., 2017), $\sigma = \sqrt{\phi}$ is the shape parameter. As ν and σ are both differentiable functions of ϕ, they retain orthogonality to the mean μ.

Regression models on the mean μ and a generic shape parameter η take the form

$$g(\mu) = x^\top\beta; \qquad h(\eta) = w^\top\gamma \tag{5}$$

where the two covariate vectors x and w of respective dimension p and q may be distinct or overlapping; and $g(\cdot)$ and $h(\cdot)$ are suitable link functions. If μ and η are orthogonal parameters, then simple application of the chain rule to (4) and (5) yields orthogonality between the elements of β and the elements of γ:

$$E\left(\frac{\partial^2}{\partial\beta_j\,\partial\gamma_k}\log f\right) = 0\,.$$

It is this aspect of PiG regression that we will investigate. PiG regression models on the mean parameter have been specified, in their different parametrizations, by Dean et al. (1989), Stein and Juritz (1988) and

Table 1 Summary statistics: droxidopa trial

	Treatment	Control
n	92	105
Number of falls		
Mean	3.35	8.65
Variance	62.0	1388.1
Maximum	49	358

Jørgensen (1987). All of these models assume a constant shape parameter. More recently, Generalized Additive Models for Location, Scale and Shape (GAMLSS) (Rigby and Stasinopoulos, 2005) enable the specification of regression models for the mean and up to three shape parameters, on a wide range of response distributions. The GAMLSS for the PiG response specifies (1) as response distribution and

$$g(\mu) = x^\top \beta; \qquad h(\sigma) = w^\top \gamma \, .$$

Estimation is available in the R package `gamlss` (Stasinopoulos et al., 2017). In Section 3 we compare estimation for PiG regression models for our data using the GAMLSS model and a model based on the orthogonal parametrization (3) for which, by analogy, we assume:

$$g(\mu) = x^\top \beta; \qquad h(\alpha) = w^\top \delta \, .$$

3 Description of data

Neurogenic Orthostatic Hypotension (nOH) is a sudden, dangerous fall in blood pressure when standing from a sitting or lying position. This disease affects patients with primary autonomic failure, such as Parkinson's Disease (PD), multiple system atrophy and pure autonomic failure. In the study which we are analysing (Hauser et al., 2016), the aim was to demonstrate the efficacy of the drug droxidopa, a treatment for nOH, over a ten-week period. Patient-reported falls was a secondary efficacy measure. Participants were patients having PD and nOH, randomized to receive either droxidopa ($n = 92$) or placebo ($n = 105$). Summary statistics are given in Table 1. Inspection of the data reveals a reduction in the number of falls in the treatment group, and a marked reduction in the variance of falls in that group, with a few heavy fallers in the placebo arm. The incidence rate ratio (IRR) is 0.387, 95% CI=(0.133 - 0.897) using the nonparametric bootstrap. Alternatively, a treatment effect may be demonstrated using nonparametric testing. Because of the between-group heterogeneity, nonparametric tests that assume a location shift without a variability difference, such as the Wilcoxon-Mann Whitney test, are inappropriate (Neuhäuser, 2011). Nonparametric location-scale tests such as the Lepage test (Hollander et al., 2013) are suitable in this case, and the Lepage test detects a significant location-scale change ($p < 0.0001$).

In order to accommodate covariates and to adjust for differing exposure times, a regression approach was adopted. The marginal fits of the PiG and negative binomial distributions were examined, and are shown in Figure 1. The PiG appears to provide a better marginal fit than the negative binomial, and in addition has a lower AIC value (891 vs 934), so was considered an appropriate response distribution. The following two PiG regression models were fitted in `gamlss`, including and excluding a treatment effect on the dispersion parameter σ:

Model A	Model B
$\log \mu_i = \beta_0 + \beta_1 x_i + \log t_i$	$\log \mu_i = \beta_0 + \beta_1 x_i + \log t_i$
$\log \sigma = \gamma_0$	$\log \sigma_i = \gamma_0 + \gamma_1 x_i$

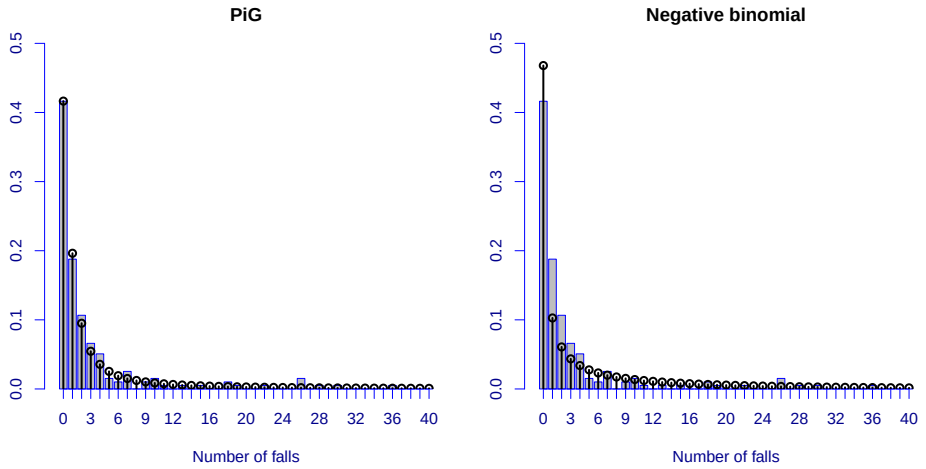

Figure 1 Barplots of the number of falls, truncated at 40 for display purposes and theoretical probabilities according to the marginal fits of PiG (left) and negative binomial (right) distributions.

Table 2 Parameter estimates for the PiG regression, using non-orthogonal (GAMLSS) model

	Model A			Model B		
Parameter	estimate	s.e.	p-value	estimate	s.e.	p-value
β_0	-1.779	0.327	<0.001	-1.417	0.541	0.009
β_1	-0.322	0.337	0.341	-1.489	0.601	0.014
γ_0	2.970	0.380	<0.001	3.461	0.592	<0.001
γ_1	-	-	-	-1.667	0.706	0.002
Global deviance		887.9			882.0	

Table 3 Parameter estimates for the PiG regression, using orthogonal model

	Model C			Model D		
Parameter	estimate	s.e.	p-value	estimate	s.e.	p-value
β_0	-0.864	0.632	0.171	-0.870	0.669	0.193
β_1	-2.077	0.687	0.003	-2.074	0.714	0.004
δ_0	-0.034	0.095	0.720	-0.093	0.124	0.453
δ_1	-	-	-	0.152	0.196	0.438
Global deviance		884.2			883.6	

for patient $i = 1, \ldots, n$, where x_i is a treatment indicator and $\log t_i$ the offset term for treatment duration t_i. Estimates for the two models are given in Table 2. In Model B, the treatment effect on σ, $\hat{\gamma}_1$, is highly significant. However, the parameter of interest is β_1, the logarithm of the IRR. A non-significant effect of droxidopa is observed in Model A, $\hat{\beta}_1 = -0.322$ (SE $= 0.337$), $p = 0.341$, whereas a different conclusion $\hat{\beta}_1 = -1.489$ (SE $= 0.601$), $p = 0.014$ stems from Model B. These inconsistent results are troubling as we are dealing with a clinical trial.

We consider the orthogonal parametrization of the PiG (3). We specify the log-linear models

Model C
$\log \mu_i = \beta_0 + \beta_1 x_i + \log t_i$
$\log \alpha = \delta_0$

Model D
$\log \mu_i = \beta_0 + \beta_1 x_i + \log t_i$
$\log \alpha_i = \delta_0 + \delta_1 x_i$

Results are given in Table 3. The estimates of β_1 are almost the same under models C and D, indicating a strong treatment effect on the number of falls which is robust to specification of the α model. The estimated treatment effects in Models A and B are attenuated by comparison, and quite different from each other. We will investigate this issue further by simulations in Section 5.1.

4 Inference

When we consider confidence intervals (CIs) for the model coefficients, we need to bear in mind that the skewness of the underlying PiG distribution may make convergence to asymptotic normality of the MLEs rather slow, rendering Wald-type confidence intervals inaccurate, particularly for small to moderately sized samples (Royston et al., 2007). We should also consider the effect of misspecification of the dispersion model on the CIs.

Let $\theta = (\beta^\top, \delta^\top)^\top$ be the overall parameter of dimension $p+q$ under the orthogonal parametrization, and $\hat{\theta} = (\hat{\beta}^\top, \hat{\delta}^\top)^\top$ the corresponding MLE. Standard maximum likelihood theory shows that under usual regularity conditions $\sqrt{n}(\hat{\theta} - \theta)$ is normally distributed with mean $\mathbf{0}$ and variance-covariance matrix:

$$V(\theta) = \begin{pmatrix} I_{\beta\beta}^{-1} & \mathbf{0} \\ \mathbf{0}^\top & I_{\delta\delta}^{-1} \end{pmatrix},$$

where $I_{\beta\beta}$ and $I_{\delta\delta}$ are the corresponding block-diagonal matrices of the Fisher information matrix, i.e. $I_{\beta\beta} = E\left(-\frac{\partial^2}{\partial\beta\partial\beta^\top}\log f\right)$ and $I_{\delta\delta} = E\left(-\frac{\partial^2}{\partial\delta\partial\delta^\top}\log f\right)$. Various asymptotic Wald confidence intervals (CIs) with nominal $(1-\alpha)$ coverage are available for any component $\theta = \theta_j$, $j = 1, \ldots, p+q$ of the overall parameter θ. They all have the familiar expression

$$\hat{\theta}_j \pm z_{\alpha/2}\sqrt{\hat{V}_{jj}/n} \tag{6}$$

where $\hat{V}_{jj}$ is the jth diagonal element of a consistent estimate of the asymptotic variance $V(\theta)$, and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the normal distribution.

A simulation study comparing coverage of the CI (6) for the parameter β_1, based on different estimates of $\hat{V}_{jj}$, is reported in Section 5.2. Three possible estimates of $\hat{V}_{jj}$ were considered: the observed variance where $\hat{V}_{\text{obs}} = \hat{I}^{-1}$ and $\hat{I}$ is minus the average second derivative of the log-likelihood over the sample computed at $\hat{\theta}$; the asymptotic variance $\hat{V}_{\text{asym}}$ where $\hat{I}_{\beta\beta}(\hat{\theta})$ and $\hat{I}_{\delta\delta}(\hat{\theta})$ replace their expectations in the block-diagonal matrix V; and a sandwich formula (Zeileis, 2006) $\hat{V}_{\text{sand}} = \hat{P}^{-1}\hat{Q}\hat{P}^{-1}$ where

$$\hat{P} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi(y_i, x_i, w_i; \theta)}{\partial\theta}\Big|_{\theta=\hat{\theta}}$$

$$\hat{Q} = \frac{1}{n}\sum_{i=1}^{n}\psi(y_i, x_i, w_i; \theta)\psi^\top(y_i, x_i, w_i; \theta)\Big|_{\theta=\hat{\theta}}$$

and $\psi(y, x, w; \theta) = \frac{\partial}{\partial\theta}\log f(y, x, w; \theta)$ is the score function. The rationale comes from M-estimation theory whereby, under mild regularity conditions given in Huber (1967), both $\hat{P}$ and $\hat{Q}$ have expectation V for the MLE defined as the solution for θ to $\sum_{i=1}^{n}\psi(y_i, x_i, w_i; \theta) = 0$.

In addition, we considered a profile likelihood CI (Venzon and Moolgavkar, 1988), created by inversion of the likelihood-ratio test (LRT). Let $L(\theta_j, \nu_j)$ denote the likelihood of the PiG model as a function of θ_j,

Table 4 Simulation scenarios

Scenario	Mean		Variance	
	β_0	β_1	Treatment group	Control group
1	2	-1	50	50
2	2	-1	50	250
3	2	-1	50	357
4	2	-1	50	1000

the parameter of interest and ν_j the nuisance parameter (all components of θ apart from θ_j), and $L_1(\theta_j) = \max_{\nu_j} L(\theta_j, \nu_j)$, the profile likelihood. The LRT statistic for $H_0 : \theta_j = \theta^*$ is simply $2(\log L_1(\hat{\theta}_j) - \log L_1(\theta^*))$. A $(1-\alpha)$ profile likelihood CI for θ_j consists of those values of θ^* for which the test cannot reject H_0 at significance level α, i.e.

$$\{\theta^* : 2(\log L_1(\hat{\theta}_j) - \log L_1(\theta^*)) \leq \chi^2_1(1-\alpha)\},$$

where $\chi^2_1(1-\alpha)$ is the $(1-\alpha)$ quantile of the χ^2 distribution with one degree of freedom. The profile likelihood CI is usually obtained using the bisection algorithm proposed by Venzon and Moolgavkar (1988).

To avoid relying on asymptotic results, we considered both parametric and non-parametric percentile Bootstrap CIs for the PiG regression parameters and used the R package `boot` (Canty and Ripley, 2016). For the parametric bootstrap, as recommended in Davison and Hinkley (1997), a simulated dataset of the same form of the original dataset was simulated $B = 1000$ times according to the PiG model with the original parameter estimate and passed to the MLE statistic to get a bootstrap replicate $\hat{\theta}^b_j$, $b = 1, \ldots, B$. The bootstrap $(1-\alpha)$ CI reported here is the percentile bootstrap CI, i.e.

$$[\hat{G}^{-1}(\alpha/2),\ \hat{G}^{-1}(1-\alpha/2)]$$

where $\hat{G}$ is the cumulative distribution of all bootstrap replicates $\hat{\theta}^b_j$.

5 Simulation studies

5.1 Parameter estimation

Our first simulation study tests estimation of the two alternative PiG regression models. We assume that we are evaluating the effect of a treatment between two groups, under differing conditions of dispersion in the two groups, and differing specifications of the dispersion model. However, we need to define the scale on which "dispersion" between the two groups differs. If dispersion differences were to be specified on the scale of either σ or α, this would give an estimation advantage to the corresponding model. In order not to bias the comparison towards either model, we rather specify the dispersion difference on the scale of the variance.

Two thousand five hundred samples of size $n = 2000$ (1000 in each group) were generated under each of the four conditions shown in Table 4. Scenario 4 was chosen to roughly replicate the falls data set; in scenario 3 the dispersion of the two groups is equal; and scenarios 1 and 2 present other combinations. Estimation was carried out under Models A and B for PiG response distribution (1); and under Models C and D for response distribution (3). Results are shown in Figure 2. (Similar results were obtained for $n = 100, n = 200$ and $n = 500$ and are not shown here.)
As expected, the treatment effect estimate $\hat{\beta}_1$ is unbiased when the full model is specified, regardless of the parametrization used. However, when restricted model A is used, i.e. when a common dispersion across

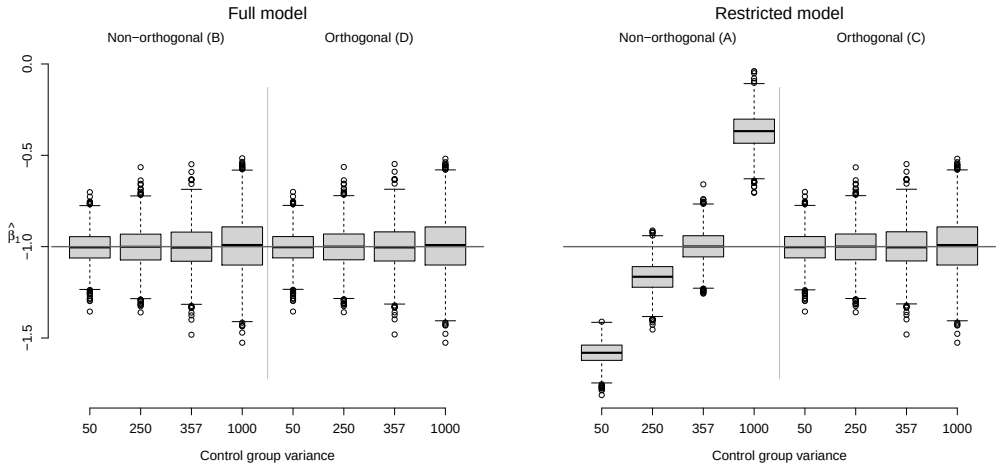

Figure 2 Simulation results: distribution of $\hat{\beta}_1$ under the full and restricted models and using non-orthogonal and orthogonal parametrizations. In all cases the treatment group variance is 50.

Table 5 Coverage of 95% CI for β_1

n	Wald Obs	Wald Asym	Wald Sand	LRT	Boot Non Par	Boot Par
200	89.9	89.9	81.4	96.4	80.9	86.8
500	91.9	91.7	87.5	95.9	86.4	90.7
1000	93.6	93.6	89.9	96.2	89.6	91.9

treatment arms is assumed using the non-orthogonal parametrization (standard analysis), severe bias in $\hat{\beta}_1$ is observed, except under scenario 3 where the difference in variance between groups leads to the same dispersion parameter for both groups.

The incorrect assumption of the same dispersion parameter σ for the two treatment groups under scenarios 1, 2 and 4 leads to biased estimates of σ for both groups and, because of the correlation between the estimates of μ and σ, this leads in turn to biased estimates of μ for both groups. In contrast, an unbiased treatment effect estimate $\hat{\beta}_1$ is obtained when fitting restricted model C based on the orthogonal parametrization. Orthogonalization has made the estimation of the mean parameter robust to the dispersion model misspecification.

5.2 Inference

We investigated inferential aspects in the PiG model using an orthogonal parametrization, starting with the well-specified model D with parameter values $(\beta_0, \beta_1, \delta_0, \delta_1) = (-1, -2, -0.1, 0.15)$ to mimic the results obtained on the trial data. The coverage of various nominal 95% CIs for the treatment effect parameter β_1 was computed for three sample sizes: $n = 200, 500$ and 1000, corresponding roughly to small, intermediate and large clinical trials. Results are displayed in Table 5. Competitors defined in Section 4 include the three Wald CIs with variance computed respectively as the observed variance (Wald Obs, also used by the `gamlss` function for inference), the asymptotic variance (Wald Asymp) and the sandwich formula (Wald Sand); the profile likelihood CI (LRT) and the nonparametric (Boot NonPar) and parametric (Boot Par) bootstraps based on $B = 1000$ replicates.

Table 6 Coverage of 95% profile likelihood CI for β_1 by δ_1, for the full (correctly specified) and restricted (misspecified) models under the orthogonal parametrization (Models D and C, respectively) for $n = 200$ (100 per group).

	Treatment effect on dispersion δ_1				
	0.15	0.25	0.40	0.60	0.8
Full model (D)	96.4	96.2	96.2	96.3	96.3
Restricted model (C)	94.9	94.3	93.5	91.6	89.4

An additional difficulty arises when computing the profile likelihood CI. Often, and particularly in small samples, only the upper bound is computable due to the likelihood flattening out for small values of parameter value δ_1. A possible remedy is to add a small quadratic penalty to the likelihood function before inverting it. If we write the penalty as $-\frac{1}{2}\lambda||\theta||^2$ with $\theta = (\beta^\top, \delta^\top)^\top$ and λ the penalty parameter chosen typically small, the penalized likelihood appears as the log-density of a posterior distribution (Cole et al., 2013) assuming a normally-distributed prior for θ with $\lambda = 1/\sigma^2$, the inverse variance of the prior. We chose $\lambda = 0.10$, which is equivalent to assuming a Gaussian prior centered at zero with variance 10. Other choices are possible, with smaller λ values resulting in wider CIs.

It can be seen from Table 5 that for $n = 200$ and $n = 500$, Wald-type and parametric bootstrap CIs are too short and cannot reach their nominal coverage, with the sandwich CI being the worst. Their performance improves in large samples ($n = 1000$), especially for Wald CIs based on the observed or asymptotic variance (93.6% coverage for both), but they are still slightly below target. Boostrap CI were not of much help with respectively 90 and 92% for the non-parametric and parametric cases. This is in agreement with what has been observed for skewed continuous data (Zhou and Dinh, 2005). The problem was shown to be linked to the difference of skewness between the two groups, which also occurs here.
The sandwich approach is not recommended for such models as the sandwich variance tends to be biased negatively in small samples as also observed here for $n = 1000$. Fortunately, the profile likelihood CI always achieves its nominal target with a coverage around 96%. In this simulation we always added a penalty term even in situations where it was not necessary (i.e. when a lower bound can be found by direct LRT inversion). Our suggestion is to invert the LRT and add the penalty only when it is needed, unless investigators are happy to report only the upper bound (which is the one that matters).

We also investigated the effect of model misspecification on the (penalized) likelihood ratio CI. We generated data using a different dispersion ($\delta_1 \neq 0$) across treatment arms but fitted a model with common dispersion, using a sample size of $n = 200$ (100 in each group). Results are presented in Table 6. As expected, the procedure deteriorates for large values of δ_1 (i.e. much smaller dispersion in the treated arm) but coverage probabilities may be considered as acceptable for values that are not too large ($\delta_1 < 0.40$). In comparison, the procedure maintains accurate coverage probability if the dispersion model is correctly specified. This confirms the need for careful modelling of the dispersion parameter, even when using the orthogonal parametrization.

6 Discussion

PiG regression is an alternative Poisson mixture model to negative binomial regression that can be used to model count data. It has been used successfully to model counts in diverse subject areas such as MRI lesion counts in multiple sclerosis (Sormani et al., 2001); actuarial claims (Willmot, 1987); word frequencies (Sichel, 1974); and species abundance (Ord and Whitmore, 1986). We were interested in modelling the number of falls in PD patients in a clinical trial. A common strategy is to model the log mean response

as a linear combination of the covariates ($x^\top \beta$), and assume a common dispersion across the treatment groups. More generally, the GAMLSS family offers the possibility to model the dispersion, usually on the log scale, as another combination of covariates $w^\top \gamma$, possibly different from those of the mean model. The standard approach of assuming a constant dispersion leads to biased MLE $\hat{\beta}$ and wrong inference when the dispersion model is misspecified (e.g. when treatment actually reduces the dispersion but is ignored at the modelling stage). The problem is caused by the dependence of $\hat{\beta}$ on the dispersion parameter estimate $\hat{\gamma}$. We propose an orthogonal parametrization that makes the regression parameter $\hat{\beta}$ asymptotically independent of the nuisance parameter $\hat{\delta}$. The asymptotic variance of $\hat{\beta}$ where δ is unknown is the same as that where δ is known. This form of orthogonality (Cox and Reid, 1987), also called 'information orthogonality', may not be sufficient to prove consistency mathematically (see Woutersen (2011) where a slightly more stringent definition is proposed). However it performs well in practice as $\hat{\delta}_\beta$, the MLE of δ for specified β, varies only slowly in β in the neighbourhood of $\hat{\beta}$, specifically $\hat{\delta}_\beta - \hat{\delta} = O_p(n^{-1})$ if $\beta - \hat{\beta} = O_p(n^{-1/2})$; see, for instance Young and Smith (2005, p.145).

It is worth noting that regression models under the two parametrizations, while having the same response distribution, have different model parameters and hence are different models in the presence of continuous covariates or an offset term. One may be more appropriate and fit better than the other, a feature unrelated to the properties of the estimators.

Inference in this model is more difficult than anticipated, with standard confidence intervals (Wald, sandwich, bootstrap) all failing to reach their nominal level for sample sizes as large as $n = 1000$, even when the dispersion model is well specified. This lack of accuracy is caused by the asymmetry of the data, more specifically the relative skewness of the two arms. This is in agreement with the findings of Zhou and Dinh (2005) in the continuous case. In contrast, the profile likelihood CI provides reliable inference for n as low as 200. Support in favour of the LRT and corresponding CI was also suggested in Hilbe (2011, p. 106) for the NB model and in Rigby and Stasinopoulos (2014) for asymmetric models for costs data. A possible difficulty arises for the PiG model as the LRT inversion provides only the upper bound of the CI. A practical solution to this consists of adding a small quadratic penalty prior to inverting.

For inference purposes, a proper dispersion model is preferable and we recommend the anticipation of a possible treatment effect on the dispersion when writing the statistical analysis plan of a clinical study. While these findings were presented here in the context of PiG regression, it is quite clear that similar issues will arise for any response distribution where the mean and dispersion parameters are not orthogonal. While the problem that we encountered could be avoided by considering other, nonparametric, inference methods, such approaches are not available when there is a need to control for other covariate and offset effects. More generally and going beyond the analysis of trials, the misspecification of the dispersion model could arise when, for example, the effect of a covariate on the dispersion parameter is incorrectly specified as linear when its effect is nonlinear. This is an area of future research.

The software available as supplementary material is sufficient for basic modelling using the orthogonal parametrization. We are in contact with the authors of the `gamlss` package and they may consider adding the orthogonal version of the PiG model in a future release.

Acknowledgements The trial and post hoc analyses were funded by Lundbeck. SRH has provided consulting services to Lundbeck NA Ltd. The authors thank Lundbeck, and particularly Dr L. Arthur Hewitt, for the use of the data derived from clinical trials and post-hoc analyses published separately. The authors also thank the two referees for their useful comments. Trial Registration: ClinicalTrials.gov NCT01176240.

Conflict of Interest

The authors have declared no conflict of interest.

References

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. John Wiley & Sons Ltd.

Canty, A. and B. D. Ripley (2016). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-18.

Cole, S. R., H. Chu, and S. Greenland (2013). Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *American Journal of Epidemiology 179*(2), 252–260.

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological) 49*, 1–39.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap methods and their application*. Cambridge University Press.

Dean, C., J. Lawless, and G. Willmot (1989). A mixed Poisson-inverse Gaussian regression model. *Canadian Journal of Statistics 17*(2), 171–181.

Greenwood, M. and G. U. Yule (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society 83*, 255–279.

Hauser, R. A., S. Heritier, G. J. Rowse, L. A. Hewitt, and S. H. Isaacson (2016). Droxidopa and reduced falls in a trial of Parkinson disease patients with neurogenic orthostatic hypotension. *Clinical Neuropharmacology 39*(5), 220–226.

Hilbe, J. M. (2011). *Negative Binomial Regression, second edition*. New York: Cambridge University Press.

Hollander, M., D. A. Wolfe, and E. Chicken (2013). *Nonparametric statistical methods*. John Wiley & Sons.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 221–233.

Huzurbazar, V. S. (1950). Probability distributions and orthogonal parameters. In *Proc. Camb. Phil. Soc*, Volume 46, pp. 281–284. Cambridge Univ Press.

IBM Corp (2013). *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY.

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

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling 13*(4), 275–303.

Neuhäuser, M. (2011). *Nonparametric statistical tests: A computational approach*. CRC Press.

Ord, J. K. and G. A. Whitmore (1986). The Poisson-inverse Gaussian distribution as a model for species abundance. *Communications in Statistics-Theory and Methods 15*(3), 853–871.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rigby, R. and D. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(3), 507–554.

Rigby, R., D. Stasinopoulos, and C. Akantziliotou (2008). A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. *Computational Statistics & Data Analysis 53*(2), 381–393.

Rigby, R. A. and D. M. Stasinopoulos (2014). A comparison of GLM and GAMLSS for modeling heteroskedastic and skewed health cost data. Personal communication.

Royston, P. et al. (2007). Profile likelihood for estimation and confidence intervals. *Stata Journal 7*(3), 376–387.

SAS Institute Inc. (2011). *SAS/STAT Software, Version 9.3*. Cary, NC.

Sichel, H. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. In *Proceedings of the Third Symposium on Mathematical Statistics*, pp. 51–97. SACSIR, Pretoria.

Sichel, H. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society. Series A (General)*, 25–34.

Sormani, M., P. Bruzzi, M. Rovaris, F. Barkhof, G. Comi, D. Miller, G. Cutter, and M. Filippi (2001). Modelling new enhancing MRI lesion counts in multiple sclerosis. *Multiple Sclerosis 7*(5), 298–304.

Stasinopoulos, M. D., R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press.

Stein, G. Z. and J. M. Juritz (1988). Linear models with an Inverse Gaussian Poisson error distribution. *Communications in Statistics-Theory and Methods 17*(2), 557–571.

Stein, G. Z., W. Zucchini, and J. M. Juritz (1987). Parameter estimation for the Sichel distribution and its multivariate extension. *Journal of the American Statistical Association 82*(399), 938–944.

Venzon, D. and S. Moolgavkar (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 37*(1), 87–94.

Willmot, G. E. (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal 1987*(3-4), 113–127.

Woutersen, T. (2011). Consistent estimation and orthogonality. In *Missing Data Methods: Cross-sectional Methods and Applications*, pp. 155–178.

Young, G. and R. Smith (2005). *Negative Binomial Regression, second edition*. New York: Cambridge University Press.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software 16*(9), 1–16.

Zhou, X. H. and P. Dinh (2005). Nonparametric confidence intervals for the one- and two-sample problems. *Biostatistics 6*(2), 187–200.

