Smooth tests of goodness of fit for the distributional assumption of regression models

DOI: 10.1111/anzs.12361
Handle: http://hdl.handle.net/1942/37334

Aust. N. Z. J. Stat. 2020 doi: 10.1111/j.1467-842X.XXX

Smooth tests of goodness of fit for the distributional assumption of regression models

J. C. W. Rayner[1,3], Paul Rippon[2], Thomas Suesse[3] and Olivier Thas[3,4,5*]

University of Newcastle, University of Wollongong, Hasselt University and Ghent University

Summary

We focus on regression models that consist of (1) a model for the conditional mean of the outcome and (2) a distributional assumption about the distribution of the outcome, both conditional on the regressors. Generalised linear models (GLM) form a well known example. The choice of the outcome distribution is often motivated by prior or background knowledge of the researcher, or it is simply chosen for convenience. We propose smooth goodness of fit tests for testing the distributional assumption in regression models. The tests arise from embedding the regression model in a smooth family of alternatives, and constructing appropriate score tests that correctly account for nuisance parameter estimation. The tests are customised, focussed and comprehensive. We present several examples to illustrate the wide applicability of our method. A small simulation study demonstrates that our tests have power to detect important deviations from the hypothesised model.

Key words: GLM; model diagnostics; Poisson regression; score test; ZIP regression

1. Introduction

Consider regression models for a univariate outcome variable Y and p regressors x_i ($i = 1, \ldots, p$), stacked into the vector $\boldsymbol{x}^\top = (x_1, \ldots, x_p)$. For notational comfort, the first regressor $x_1 = 1$ if an intercept is required in the model, and, similarly, regressors x_i may also represent interaction effects or other transformations of regressors. We will focus on models that are specified as

$$\mathrm{E}(Y \mid \boldsymbol{x}) = g^{-1}(\boldsymbol{x}^\top \boldsymbol{\beta}) = \mu(\boldsymbol{x}^\top \boldsymbol{\beta}) \quad (1)$$

$$Y \mid \boldsymbol{x} \sim f(\cdot; \mu(\boldsymbol{x}^\top \boldsymbol{\beta}), \boldsymbol{\gamma}). \quad (2)$$

*Author to whom correspondence should be addressed.
[1] Centre for Computer-Assisted Research Mathematics and its Applications, University of Newcastle, Australia
[2] School of Mathematical and Physical Sciences, University of Newcastle, Australia
[3] National Institute of Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia
[4] I-BioStat, Data Science Institute, Hasselt University, Belgium
[5] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
Email: olivier.thas@uhasselt.be

Equation (1) specifies the conditional mean of Y as a function of the linear predictor $\boldsymbol{x}^\top\boldsymbol{\beta}$ through the link function g. We will often use the notation $\mu(\boldsymbol{x}^\top\boldsymbol{\beta})$ to denote the conditional mean. Equation (2) states that the density function of the conditional distribution of Y given $\boldsymbol{x}$ is given by f, which depends on the conditional mean μ and on an additional t-dimensional nuisance parameter $\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_t)$. The γ-parameters may be related to the conditional variance of the outcome. For example, for linear regression $\text{var}\,(Y|\boldsymbol{x}) = \gamma$. Note that the model also allows the conditional variance $\text{var}\,(Y|\boldsymbol{x})$ to depend on the mean $\mu(\boldsymbol{x}^\top\boldsymbol{\beta})$.

Generalised linear models (GLM), which were first introduced by Nelder & Wedderburn (1972), form a special class. They arise if f belongs to a one-dimensional (dispersion) exponential family.

Equations (1) and (2) completely specify the conditional outcome distribution, and hence the maximum likelihood (ML) framework can be used for inference on the target parameter $\boldsymbol{\beta}$. The distributional component is often motivated by prior or background knowledge on the probabilistic mechanism that generated the data. For example, count outcomes are known to be often well described by a Poisson distribution (f is Poisson), and binary outcomes often behave like a Bernoulli distribution (f is binomial/Bernoulli). These two examples result in a GLM because Poisson and binomial distributions belong to the exponential family. In modern genomics applications, RNASeq and microbiome 16S RNA sequencing experiments are believed to give count outcomes that can be described by negative binomial (NB) distributions; see e.g., Love, Huber & Anders (2014) and McMurdie & Holmes (2014), which are overdispersed Poisson distributions that contain an overdispersion nuisance parameter. The overdispersion is explained by the biological variability on top of the technical variability that is described by the Poisson distribution. For single cell RNASeq experiments, several papers suggest that the count outcomes should be modelled with a zero-inflated negative binomial (ZINB) distribution; see e.g. Risso et al. (2017). Also in other biological applications, counts often show more zeroes than expected under a Poisson distribution and a zero-inflated Poisson (ZIP) distribution has been suggested to be more appropriate than a Poisson (Thas & Rayner 2005). The zero-inflation is often explained by a second data-generating mechanism that causes the zero counts. The ZIP, NB and ZINB distributions do not belong to the exponential family, but they are still regression models of the form (1) and (2), and hence they fall within the scope of this paper.

Valid asymptotic statistical inference on $\boldsymbol{\beta}$ requires a correct specification of the conditional mean model, and hence several papers have proposed diagnostic methods for detecting violations to the mean model (1); see e.g. Stute & Zhu (2002); Khmaladze et al. (2004); Hart (2013). However, a correct specification of the variance, or the mean-variance relationship is also required. The Quasi-Likelihood approach (Wedderburn 1974) builds upon a semiparametric model with only mean and variance(-mean) specifications. When

the latter is misspecified, the variance of the mean model parameter estimators can be estimated with a robust sandwich estimator, but this shows poor small sample behaviour (Kauermann & Carroll 2001). For GLMs, Huang & Rathouz (2017) demonstrated that the mean model parameters show orthogonality to the outcome distribution, opening the door to first nonparametrically estimate the outcome distribution, and subsequently use this estimate in an empirical likelihood for the estimation of the mean model parameters. Also see Huang (2014). Upon using the orthogonality, the authors showed that asymptotically no efficiency is lost as compared to ML in the correctly specified GLM. In small samples, however, the ML estimator still shows better performance.

Given the arguments and examples provided in the previous paragraphs, we conclude that assessing the distributional assumption, contained in (2), is also of scientific importance. Relatively few methods have been proposed for testing this distributional assumption; see, for example, Dean & Lawless (1989) and Peña & Slate (2006). In practice, formal hypothesis testing can be complemented with graphical inspection of the model fit. Residual plots may be used for assessing the mean model, but in general QQ-plots of residuals cannot be used for gauging the distributional assumption, unless the distribution belongs to a location-shift class (e.g. the normal distribution). QQ-plots and other visualisations based on residuals were strongly promoted by John Tukey in many of his works. We cannot agree more with him that this is indeed of primordial importance for all data analyses. He also developed formal statistical tests for assessing the normality assumption in linear models, but these methods are restricted to additive two-way analysis of variance (Anscombe & Tukey 1963).

In this paper we propose smooth tests of goodness of fit for testing the distributional assumption contained in (2). Smooth tests are well established for testing goodness of fit in the one-sample problem and they can properly account for nuisance parameter estimation. We refer to Rayner, Thas & Best (2009) for a comprehensive overview of the general theory and for several examples, including the normal, Poisson, NB and ZIP distributions. Here we extend the smooth testing method to the regression context as described above. We formulate the theory for regression models of the form (1) and (2), and we show how simplifications arise for special cases, including the class of GLMs (Section 2). In Section 3 we give the Poisson, normal, and ZIP distributions as examples. Small simulation studies for Poisson and ZIP models are presented in Section 4, and conclusions are formulated in Section 5.

2. Smooth tests

The general construction of the test is given in Section 2.1, and special cases resulting in simplifications are provided in Section 2.2.

2.1. The smooth test for general regression models

The development of the theory is very similar to the development detailed in Rayner, Thas & Best (2009, chapters 6 and 8). We therefore limit the exposition here to the overall procedure, with a focus on the details for the regression setting, and to the most important results. Proofs are deferred to Appendix A.1 in Supporting and Supplementary Material.

Derivation of the smooth test

The construction starts with nesting the density function of the regression model (the conditional distribution of the outcome Y, given the regressor $\boldsymbol{x}$) in a family of distributions indexed by the parameter $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_k)$ and then deriving the score test for testing $\boldsymbol{\theta} = \mathbf{0}$ against $\boldsymbol{\theta} \neq \mathbf{0}$. The number k refers to the order of the alternative. The order k embedding density of $Y \mid \boldsymbol{x}$ is given by

$$f_k(y; \mu(\boldsymbol{x}^\top\boldsymbol{\beta}), \boldsymbol{\gamma}, \boldsymbol{\theta}) = C(\mu(\boldsymbol{x}^\top\boldsymbol{\beta}), \boldsymbol{\gamma}, \boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} \theta_i h_i(y; \mu(\boldsymbol{x}^\top\boldsymbol{\beta}), \boldsymbol{\gamma})\right) f(y; \mu(\boldsymbol{x}^\top\boldsymbol{\beta}), \boldsymbol{\gamma}), \tag{3}$$

where C is a normalisation constant and $\{h_i\}$ is a set of functions that are orthonormal to the regression model with density function f. Since both C and $\{h_i\}$ differ from their counterparts of the one-sample smooth tests in the sense that they depend on the regressor $\boldsymbol{x}$, we provide some more details, but first the log-likelihood function is given for a sample of n independently sampled observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$:

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) &= \sum_{j=1}^{n} \log f_k(y_j; \mu(\boldsymbol{x}_j^\top\boldsymbol{\beta}), \boldsymbol{\gamma}, \boldsymbol{\theta}) \\ &= \sum_{j=1}^{n} \log C(\mu(\boldsymbol{x}_j^\top\boldsymbol{\beta}), \boldsymbol{\gamma}, \boldsymbol{\theta}) + \sum_{i=1}^{k} \theta_i \sum_{j=1}^{n} h_i(y_j; \mu(\boldsymbol{x}_j^\top\boldsymbol{\beta}), \boldsymbol{\gamma}) \\ &\quad + \sum_{j=1}^{n} \log f(y_j; \mu(\boldsymbol{x}_j^\top\boldsymbol{\beta}), \boldsymbol{\gamma}). \end{aligned} \tag{4}$$

Note that the last term equals the log-likelihood of the regression model.

For all $\boldsymbol{x} = \boldsymbol{x}_j$, $j = 1, \ldots, n$, the normalisation constants $C(\mu(\boldsymbol{x}^\top\boldsymbol{\beta}), \boldsymbol{\gamma}, \boldsymbol{\theta})$ must guarantee that the area under density (3) equals one. Thus for a given set of parameter values, and for sample of n observations, n normalisation constants are required. For some outcome distributions the normalisation constant may not exist, but the type of score tests that we will derive still do exist (Mardia & Kent 1991; Kallenberg, Ledwina & Rafajlowicz 1997). For the orthonormal functions $\{h_i\}$, again for a given set of parameter values, the set of functions

must be calculated for each of the n regressors $\boldsymbol{x}_j$. In particular, the orthonormality condition reads as

$$\int_{-\infty}^{+\infty} h_u(y;\mu(\boldsymbol{x}^\top\boldsymbol{\beta}),\boldsymbol{\gamma})h_v(y;\mu(\boldsymbol{x}^\top\boldsymbol{\beta}),\boldsymbol{\gamma})f(y;\mu(\boldsymbol{x}^\top\boldsymbol{\beta}),\boldsymbol{\gamma})dy = \delta_{uv}, \tag{5}$$

with $\delta_{uv} = 1$ if $u = v$ and $\delta_{uv} = 0$ otherwise.

In what follows the notation will often be simplified by omitting the dependence of C, f, f_k and $\{h_i\}$ on all parameters and regressors. For example, (5) may be written as $\mathrm{E}_0\left(h_u(Y;\mu,\boldsymbol{\gamma})h_v(Y;\mu,\boldsymbol{\gamma}) \mid \boldsymbol{x}\right) = \delta_{uv}$, in which $\mathrm{E}_0\left(\cdot \mid \boldsymbol{x}\right)$ denotes the conditional expectation under the null hypothesis $\boldsymbol{\theta} = \mathbf{0}$. We will also use μ_j as shorthand notation for $\mu(\boldsymbol{x}_j^\top\boldsymbol{\beta})$.

The smooth test requires the score test statistic for testing $\boldsymbol{\theta} = \mathbf{0}$, and hence the score statistic for $\boldsymbol{\theta}$ is required, as well as the information matrix based on the log-likelihood (4) with all expectations evaluated under the null hypothesis and conditional on the regressors. In this setting, the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are both considered as nuisance parameters and therefore we stack them into a single vector, $\boldsymbol{\eta}^\top = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)$. The nuisance parameter only needs to be estimated under the null hypothesis: its maximum likelihood estimator, which is denoted by $\hat{\boldsymbol{\eta}}$, arises from the hypothesised regression model. The following theorem gives the score statistics and the required information matrices. We will use $\mathcal{X}$ to denote the set of of n regressor vectors $\boldsymbol{x}_j$ in the sample.

Theorem 1. Score statistics and information matrices. *The score statistic for θ_i in model (4) is given by $U_i = U_i(\boldsymbol{\eta}) = (\partial/\partial\theta_i)l(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta}) = \sum_{j=1}^n h_i(y_j;\mu_j,\boldsymbol{\gamma})$. Let $g'(\mu) = (d/d\mu)g(\mu)$. We use the notation $E_k\left(\cdot\right)$ and $E_0\left(\cdot\right)$ to refer to the expectation w.r.t density functions $f_k(\cdot;\mu,\boldsymbol{\gamma},\boldsymbol{\theta})$ and $f(\cdot;\mu,\boldsymbol{\gamma})$, respectively. The elements of the information matrix are given by:*

$$\begin{aligned}
(\boldsymbol{I}_{\theta\theta})_{uv} &= -\,E_k\left(\frac{\partial^2}{\partial\theta_u\partial\theta_v}l \mid \mathcal{X}\right)\Bigg|_{\boldsymbol{\theta}=\mathbf{0}} = n\delta_{uv}\\
(\boldsymbol{I}_{\theta\gamma})_{uv} &= -\,E_k\left(\frac{\partial^2}{\partial\theta_u\partial\gamma_v}l \mid \mathcal{X}\right)\Bigg|_{\boldsymbol{\theta}=\mathbf{0}} = \sum_{j=1}^n E_0\left(h_u(Y_j;\mu_j)\frac{\partial}{\partial\gamma_v}\log f(Y_j;\mu_j) \mid \boldsymbol{x}_j\right)\\
(\boldsymbol{I}_{\theta\beta})_{uv} &= -\,E_k\left(\frac{\partial^2}{\partial\theta_u\partial\beta_v}l \mid \mathcal{X}\right)\Bigg|_{\boldsymbol{\theta}=\mathbf{0}} = \sum_{j=1}^n E_0\left(h_u(Y_j;\mu_j)\frac{\partial}{\partial\mu_j}\log f(Y_j;\mu_j) \mid \boldsymbol{x}_j\right)\frac{x_{jv}}{g'(\mu_j)}.
\end{aligned}$$

The theorem does not give expressions for $\boldsymbol{I}_{\beta\beta}$, $\boldsymbol{I}_{\gamma\gamma}$ and $\boldsymbol{I}_{\gamma\beta}$, because they are not affected by our embedding density (3), i.e. they are the information matrices under the hypothesised regression model.

For the calculation of the smooth test statistic, the expectations in the expressions of the information matrices of Theorem 1 are evaluated; this depends on the exact form of

the regression model, and the nuisance parameters need to be replaced by their maximum likelihood estimates $\hat{\boldsymbol{\eta}}$, after which these matrices are denoted by $\hat{\boldsymbol{I}}$.

Several examples of regression models will follow in Section 3. The next lemma gives the smooth test statistic and its limiting null distribution. The proof is a straightforward application of maximum likelihood theory and is omitted here (see e.g. Boos & Stefanski (2013) for a good exposition to maximum likelihood theory).

Lemma 1. Smooth test statistic and its asymptotic null distribution. *Let $\boldsymbol{V}$ denote the vector $(1/\sqrt{n})(U_1(\boldsymbol{\eta}), \ldots, U_k(\boldsymbol{\eta}))^\top$, and let $\hat{\boldsymbol{V}}$ denote the same vector but with the nuisance parameter $\boldsymbol{\eta}$ replaced with its maximum likelihood estimator $\hat{\boldsymbol{\eta}}$. The smooth test statistic for testing $\boldsymbol{\theta} = \boldsymbol{0}$ against $\boldsymbol{\theta} \neq \boldsymbol{0}$ in model (4) is given by $\hat{S}_k = n\hat{\boldsymbol{V}}^\top \left(\hat{\boldsymbol{I}}_{\theta\theta} - \hat{\boldsymbol{I}}_{\theta\eta}\hat{\boldsymbol{I}}_{\eta\eta}^{-1}\hat{\boldsymbol{I}}_{\theta\eta}^\top\right)^- \hat{\boldsymbol{V}}$ in which $(\cdot)^-$ denotes a generalised inverse. Let r be the rank of $\hat{\boldsymbol{I}}_{\theta\theta} - \hat{\boldsymbol{I}}_{\theta\eta}\hat{\boldsymbol{I}}_{\eta\eta}^{-1}\hat{\boldsymbol{I}}_{\theta\eta}^\top$. Given that $r \geq 1$, under the null hypothesis, as $n \to \infty$, $\hat{S}_k \xrightarrow{d} \chi^2_r$.*

Note that the second term in the matrix $\hat{\boldsymbol{I}}_{\theta\theta} - \hat{\boldsymbol{I}}_{\theta\eta}\hat{\boldsymbol{I}}_{\eta\eta}^{-1}\hat{\boldsymbol{I}}_{\theta\eta}^\top$ corrects the estimated information matrix of the parameter $\boldsymbol{\theta}$ for the nuisance parameter estimation. For many regression models some of the elements of $\hat{\boldsymbol{V}}$ will be exactly zero as a consequence of the estimation of the model parameters; in Section 3 examples will be given. It is then convenient to first remove these components from $\hat{\boldsymbol{V}}$, or, equivalently, remove the corresponding terms $\theta_i h_i(y; \mu(\boldsymbol{x}^\top\boldsymbol{\beta}), \boldsymbol{\gamma})$ from model (3). In this case the estimated covariance matrix $\hat{\boldsymbol{I}}_{\theta\theta} - \hat{\boldsymbol{I}}_{\theta\eta}\hat{\boldsymbol{I}}_{\eta\eta}^{-1}\hat{\boldsymbol{I}}_{\theta\eta}^\top$ will be of full rank r, which is k minus the number of components removed from $\hat{\boldsymbol{V}}$ (or from the model).

A single element from $\hat{\boldsymbol{V}}$, say $\hat{V}_i$, corresponds the parameter θ_i of the density function (3) and to score statistic U_i (Theorem 1), and hence it can serve as the basis of a test statistic for testing $\theta_i = 0$ against $\theta_i \neq 0$. With $\hat{\sigma}_i^2$, the ith diagonal element of $\hat{\boldsymbol{I}}_{\theta\theta} - \hat{\boldsymbol{I}}_{\theta\eta}\hat{\boldsymbol{I}}_{\eta\eta}^{-1}\hat{\boldsymbol{I}}_{\theta\eta}^\top$, the hypotheses can be tested upon using the asymptotic null distribution, i.e. $\hat{V}_i/\hat{\sigma}_i \xrightarrow{d} \mathrm{N}(0,1)$, as $n \to \infty$. The statistic $\hat{V}_i/\hat{\sigma}_i$ is referred to as the ith component of $\hat{S}_k$. Examples will follow later in this paper.

We conclude this section with a note on the convergence of the test statistics to their asymptotic null distributions. Rippon (2012) performed simulation studies for assessing the empirical type I error rates as a function of the sample size. Results for the special case of the Poisson regression can be found at `https://ogma.newcastle.edu.au/vital/access/services/Download/uon:12622/ATTACHMENT02?view=true\#page=93`. These results demonstrate slow convergence of the asymptotic approximations. Sample sizes of at least 100 are needed for good type I error rate control. For normal linear regression models, Peña & Slate (2006) concluded from their simulation study, that the convergence of test statistics similar to our

$\hat{V}_3$ and $\hat{V}_4$ is also very slow. The parametric bootstrap, on the other hand, works well (see simulation studies in Section 4).

2.2. Special cases

Generalised linear models (GLM)

Generalised linear models form a special class of regression models (1) and (2) by restricting the conditional distribution of the outcome variable to the exponential family. Many GLMs belong to a one-parameter exponential family for which the single parameter is related to the conditional mean and hence to the $\boldsymbol{\beta}$-parameter through $\mu(\boldsymbol{x};\boldsymbol{\beta}) = \mathrm{E}\left(Y \mid \boldsymbol{x}\right) = g^{-1}(\boldsymbol{x}^\top\boldsymbol{\beta})$. Examples include logistic regression (binomial distribution) and Poisson regression (Poisson distribution), among others. For this class of models, there is no nuisance parameter γ, and hence some of the information matrices simplify. In particular,

$$(\boldsymbol{I}_{\theta\beta})_{uv} = -\mathrm{E}_0\left(\frac{\partial^2}{\partial\theta_u\partial\beta_v}l \mid \mathcal{X}\right) = \sum_{j=1}^{n}\frac{x_{jv}}{\mathrm{var}_0\left(Y_j \mid \boldsymbol{x}_j\right)g'(\mu_j)}\mathrm{E}_0\left(h_u(Y_j;\mu_j)(Y_j-\mu_j)\right).$$

Furthermore, if the canonical link function is used, $(\boldsymbol{I}_{\theta\beta})_{uv}$ further simplifies to

$$(\boldsymbol{I}_{\theta\beta})_{uv} = -\mathrm{E}_0\left(\frac{\partial^2}{\partial\theta_u\partial\beta_v}l \mid \mathcal{X}\right) = \sum_{j=1}^{n}x_{jv}\mathrm{E}_0\left(h_u(Y_j;\mu_j)(Y_j-\mu_j)\right).$$

The use of the exponential family also allows the use of Iterative Reweighted Least Squares (IRLS) as a general algorithm for $\boldsymbol{\beta}$-parameter estimation.

GLMs may include a dispersion parameter. These models assume that the outcome distribution belongs to the exponential dispersion family. Normal and gamma regression models belong to this class. Although in most GLM literature the dispersion parameter is not estimated by maximum likelihood, but rather by the method of moments, we will further assume that the dispersion parameter is the γ-nuisance parameter. Note that for the normal distribution the maximum likelihood estimator and the method of moment estimator are equivalent.

Score functions are linear combinations of the basis functions

For many distributions the score functions of the nuisance parameters and the mean μ_j can be expressed as a linear combination of the orthonormal basis functions $h_i(y;\mu,\gamma)$,

$i = 1, \ldots, k$. In particular,

$$\frac{\partial}{\partial \gamma_v} \log f(y; \mu_j, \gamma) = \sum_{i=1}^{k} a_{ijv} h_i(y; \mu_j, \gamma) \quad \text{and} \quad \frac{\partial}{\partial \mu_j} \log f(y; \mu_j, \gamma) = \sum_{i=1}^{k} b_{ij} h_i(y; \mu_j, \gamma), \tag{6}$$

for some sets of constants $\{a_{ijv}\}$ and $\{b_{ij}\}$. Often, many of these constants are zero. The normal and exponential distributions are two examples.

Upon using the orthonormality, the elements of the information matrix now become

$$(\boldsymbol{I}_{\theta\gamma})_{uv} = \sum_{j=1}^{n} a_{ujv}, \qquad (\boldsymbol{I}_{\theta\beta})_{uv} = \sum_{j=1}^{n} \frac{x_{jv} b_{uj}}{g'(\mu_j)},$$

$$(\boldsymbol{I}_{\gamma\beta})_{uv} = \sum_{j=1}^{n} \frac{x_{jv} (\sum_{i=1}^{k} a_{iju} b_{ij})}{g'(\mu_j)}, \qquad (\boldsymbol{I}_{\gamma\gamma})_{uv} = \sum_{j=1}^{n} \sum_{i=1}^{k} a_{iju} a_{ijv},$$

$$(\boldsymbol{I}_{\beta\beta})_{uv} = \sum_{j=1}^{n} \frac{x_{ju} x_{jv} \sum_{i=1}^{k} b_{ij}^2}{(g'(\mu_j))^2}.$$

Note that for GLMs from a one-parameter exponential family, the score function of μ_j is a first order polynomial in y, proportional to $y - \mu_j$.

For several common distributions the orthonormal basis consists of orthonormal polynomials in $y - \mu_j$. Rayner, Thas & Best (2009, Appendix C) gives explicit forms for many examples.

3. Examples

In the sections following, smooth tests for several specific regression models will be discussed in more detail. All tests are constructed as earlier described. Poisson regression, which is treated in Section 3.1 is an example of a GLM with no nuisance parameters and with a score function linearly related to the orthonormal polynomials (Sections 2.2 and 2.2 apply). Logistic regression belongs to the same type of regression models; some details are given in Appendix A.2 in the Supporting and Supplementary Material. The normal linear regression model is discussed in Section 3.2. Again Sections 2.2 and 2.2 apply, but now a nuisance parameter is present (the error term variance). In Section 3.3 the smooth test for zero-inflated Poisson (ZIP) regression models are developed. ZIP regression models do not belong to the class of GLMs, but still our general theory applies.

For all these regression models, a numerical example is provided. The p-values are computed by means of a parametric bootstrap procedure with $2,000$ bootstrap runs. In

simulation studies presented in Section 4, we will demonstrate that this bootstrap procedure succeeds in controlling the type I error rate.

3.1. Poisson regression

Test statistic

Poisson regression fits into the GLM framework with a Poisson distribution for the outcome variable and a canonical log-link, i.e. $g(\mu(\boldsymbol{x};\boldsymbol{\beta})) = \log(\mu(\boldsymbol{x};\boldsymbol{\beta})) = \boldsymbol{x}^\top\boldsymbol{\beta}$. No nuisance parameters other than the β-parameters are involved. The Poisson–Charlier polynomials are known to form an orthonormal basis w.r.t. the Poisson distribution. The polynomials of order one and two are given by $h_1(y;\mu) = (y-\mu)/\sqrt{\mu}$ and $h_2(y;\mu) = [(y-\mu)^2 - y]/(\mu\sqrt{2})$. Higher order polynomials can be found in Rayner, Thas & Best (2009, Appendix C). The score function for the mean is given by $(\partial/\partial\mu)\log f(y;\mu) = (y-\mu)/\mu = h_1(y;\mu)/\sqrt{\mu}$. From this expression, we see that the simplifications of Section 2.2 apply with $b_{1j} = 1/\sqrt{\mu_j}$ and $b_{ij} = 0$ for $i = 2,\ldots,k$. Hence, we find

$$(\boldsymbol{I}_{\theta\beta})_{1v} = \sum_{j=1}^{n} x_{jv}\sqrt{\mu_j} \quad (\boldsymbol{I}_{\beta\beta})_{uv} = \sum_{j=1}^{n} x_{ju}x_{jv}\mu_j \quad \text{and } (\boldsymbol{I}_{\theta\beta})_{uv} = 0, \quad \text{for } u = 2,\ldots,k.$$

With these expressions, we find that $\hat{\boldsymbol{I}}_{\theta\theta} - \hat{\boldsymbol{I}}_{\theta\eta}\hat{\boldsymbol{I}}_{\eta\eta}^{-1}\hat{\boldsymbol{I}}_{\theta\eta}^\top$ is a diagonal matrix with first element equal to ω^2 (see further for its definition) and all other diagonal elements equal to one. The diagonal structure results in a decomposition of the order k smooth test statistic: $\hat{S}_k = \hat{V}_1^2/\omega^2 + \hat{V}_2^2 + \cdots + \hat{V}_k^2$, where $\hat{V}_i = \sum_{j=1}^n h_i(y_j;\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}})/\sqrt{n}$ and where ω^2 can be conveniently expressed using matrix notation. Let $\boldsymbol{X}$ denote the usual $n \times p$ design matrix and $\boldsymbol{D}$ a diagonal matrix with elements $\sqrt{\mu_1},\ldots,\sqrt{\mu_n}$. With $\boldsymbol{I}_n$ the $n \times n$ identity matrix and $\mathbf{1}_n$ a column vector with all elements set to one, we write $\omega^2 = \mathbf{1}_n^\top\left(\boldsymbol{I}_n - \boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{D}^2\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}\right)\mathbf{1}_n/n$.

We discuss the first few components in some detail. The numerator of the first component is given by the square of

$$\hat{V}_1 = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} h_1(y_j;\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{y_j - \hat{\mu}_j}{\sqrt{\hat{\mu}_j}} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{y_j - \exp(\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}})}{\sqrt{\exp(\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}})}},$$

in which $y_j - \hat{\mu}_j$ is the residual of the jth observation. The denominator $\sqrt{\hat{\mu}_j}$ is the standard error of the residual if the true $\boldsymbol{\beta}$ would have been used instead of its MLE $\hat{\boldsymbol{\beta}}$. To correct for the estimation of $\boldsymbol{\beta}$, the factor $1/\omega^2$ appears in the first component. Hence, $\hat{V}_1^2/\omega^2$ can be interpreted as a goodness of fit statistic for the specification of the mean model, which includes both the specification of the linear predictor $\eta = \boldsymbol{x}^\top\boldsymbol{\beta}$ and the link function $g(\cdot)$.

The second component can be written as the square of

$$\hat{V}_2 = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h_2(y_j; \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{(y_j - \hat{\mu}_j)^2 - y_j}{\sqrt{2}\hat{\mu}_j} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{(y_j - \exp(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}))^2 - y_j}{\sqrt{2}\exp(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})}.$$

In the numerator we interpret $(y_j - \hat{\mu}_j)^2 - y_j$ as the residual of the jth observation w.r.t. the model-based specification of the variance of the outcome. In particular, the conditional expectation of $(Y_j - \hat{\mu}_j)^2$, given $\boldsymbol{x}_j$, is, by definition, the conditional variance of the outcome Y_j, and the conditional expectation of Y_j given $\boldsymbol{x}_j$ trivially is the conditional mean of the outcome. Hence the conditional expectation of the residual is a contrast between the conditional variance and mean, which, if the Poisson regression model holds true, should be equal. Hence, the second component can be interpreted as a statistic measuring the goodness of fit of the second moment of the Poisson regression model. The higher order components $\hat{V}_3^2, \ldots, \hat{V}_k^2$ are interpreted in a similar fashion. Finally, we note that $\hat{V}_2$ is closely related to a test statistic proposed by Dean & Lawless (1989) for detecting overdispersion in Poisson regression. Their test statistic is given by $\sum_{j=1}^{n}[(y_i - \exp(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}))^2 - y_i] / \sqrt{2n \sum_{j=1}^{n} \left[\exp(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})\right]^2}$.

Numerical example

Spinelli, Lockhart & Stephens (2002) presented an example in which the expected frequency of cases of bladder cancer in male aluminium workers is analysed with a Poisson regression model with age and exposure to coal tar pitch volatiles as regressors. The age is included as a factor variable with 11 levels, referring to age groups, each spanning five years. The exposure is included as a continuous regressor, but it is actually an ordinal variable taking four values. The model also includes an offset, which is set to the logarithm of the total person years at risk. The dataset includes 4213 workers. One of the conclusions from the data analysis is that the exposure has a significant effect ($p = 0.00184$) on the number of bladder cancer cases per person year, correcting for age. The effect is estimated as a factor 2.089 increase in the expected number of bladder cancer cases per person year, when the exposure level increases with one level.

The smooth test of order $k = 4$ has been applied to this example. The p-values were computed using the parametric bootstrap with $2,000$ bootstrap samples. Table 1 shows the results. The overall order $k = 4$ test gives $p = 0.632$, and hence at the 5% level of significance there is no evidence against the Poisson assumption. Neither do any of the component tests suggest any deviation from the Poisson assumption.

Table 1. Results of the order $k = 4$ smooth test applied to the bladder cancer example. The two-sided p-values are computed from 2,000 parametric bootstrap runs (denoted by p_B) and from the asymptotic distribution (p_A). Results are shown for two models: "Age Factor" refers to the model with age included as a factor variable, and "Age Ordinal" refers to the model with age as an ordinal regressor.

	Age Factor			Age Ordinal		
Statistic	value	p_A	p_B	value	p_A	p_B
$\hat{S}_4$	0.631	0.960	0.653	1.306	0.860	0.798
$\hat{V}_1/\omega$	0.123	0.902	0.951	1.025	0.305	0.020
$\hat{V}_2$	-0.733	0.464	0.842	-0.502	0.616	0.390
$\hat{V}_3$	0.280	0.779	0.958	0.013	0.999	0.694
$\hat{V}_4$	0.003	0.998	0.998	0.049	0.961	0.962

Table 1 also shows the results of the smooth test applied to a Poisson regression model in which age is not included as a factor variable, but as a continuous regressor which takes the values 1 up to 11, referring to the 11 age classes (age is here thus an ordinal variable). The first component test gives a two-sided p-value of 0.020, and hence it suggests that the mean model is not correctly formulated. Although the results are not presented here, we generally also advise looking at the conventional residual plots for assessing the correctness of the mean model.

Finally, Table 1 also shows the p-values calculated from the asymptotic null distributions. Although the conclusions at the 5% level of significance are almost the same, these results illustrate the discrepancy between the bootstrap and the asymptotic approximation for the sample size of this example ($n = 44$).

3.2. Linear regression with normal error terms

Test statistic

For the normal linear regression model, the link function is the identity function, and the residual variance, say σ^2, is the only nuisance parameter in the normal distribution; thus $\gamma = \sigma^2$. Hence, the conditional mean is written as $\mu(\boldsymbol{x}^\top\boldsymbol{\beta}) = \boldsymbol{x}^\top\boldsymbol{\beta}$. The system of orthonormal polynomials is given by the Hermite polynomials (Rayner, Thas & Best 2009, Appendix C), of which the first few are given by $h_0(y; \boldsymbol{x}^\top\boldsymbol{\beta}, \sigma^2) = 1$, $h_1(y; \boldsymbol{x}^\top\boldsymbol{\beta}, \sigma^2) = (y - \boldsymbol{x}^\top\boldsymbol{\beta})/\sigma$ and $h_2(y; \boldsymbol{x}^\top\boldsymbol{\beta}, \sigma^2) = [(y - \boldsymbol{x}^\top\boldsymbol{\beta})^2 - \sigma^2]/(\sigma^2\sqrt{2})$.

For the normal regression model, the score functions for $\boldsymbol{\beta}$ and σ^2 can be written as ($u = 1, \ldots, p$)

$$\frac{\partial}{\partial\beta_u}\log f(y; \mu_j, \sigma^2) = x_{ju}(y_j - \boldsymbol{x}_j^\top\boldsymbol{\beta}) \quad \frac{\partial}{\partial\sigma^2}\log f(y; \mu_j, \sigma^2) = \left[(y_j - \boldsymbol{x}_j^\top\boldsymbol{\beta})^2 - \sigma^2\right].$$

Note that the right hand sides of the equations involve the Hermite polynomials h_1 and h_2 and that the score functions take the form of (6) and hence the simplifications of Section 2.2 apply. Also note that the maximum likelihood estimator (MLE) of σ^2 is the solution of $\sum_{j=1}^{n}(\partial/\partial\sigma^2)\log f(y_j;\mu_j,\sigma^2) = \sigma^2\sum_{j=1}^{n} h_2(y_j;\mu_j;\sigma^2) = 0$. This implies that the second component statistic $\hat{V}_2 = S_2(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)/\sqrt{n} = 0$. The interpretation is that the smooth test cannot detect a wrongly specified variance, because the variance is estimated by matching the variance parameter σ^2 to the empirical variance (up to the asymptotically negligible factor $n/(n-1)$). We therefore continue with the construction of the smooth test, with the second orthonormal Hermite polynomial removed from model (3). Upon applying the methods described in this paper, including the simplifications of Section 2.2, we find again, as for Poisson regression, that $\hat{\boldsymbol{I}}_{\theta\theta} - \hat{\boldsymbol{I}}_{\theta\eta}\hat{\boldsymbol{I}}_{\eta\eta}^{-1}\hat{\boldsymbol{I}}_{\theta\eta}^{\top}$ is the identity matrix with the first element replaced by $\omega^2 = \frac{1}{n}\mathbf{1}_n(\boldsymbol{I}_n - \boldsymbol{H})\mathbf{1}_n^{\top}$, in which $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}$ is the hat matrix of the linear regression model. Thanks to the diagonal structure of the matrix, the order k smooth test statistic becomes $\hat{S}_k = \hat{V}_1^2/\omega^2 + \hat{V}_3^2 + \cdots + \hat{V}_k^2$, with $\hat{V}_i = \sum_{j=1}^{n} h_i(y_j;\boldsymbol{x}_j^{\top}\hat{\boldsymbol{\beta}},\hat{\sigma}^2)/\sqrt{n}$ $(i = 1, 3, 4, \ldots, k)$. For $i = 1$, we write

$$\hat{V}_1 = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} h_1(y_j;\boldsymbol{x}_j^{\top}\hat{\boldsymbol{\beta}},\hat{\sigma}^2) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{y_j - \boldsymbol{x}_i^{\top}\hat{\boldsymbol{\beta}}}{\hat{\sigma}} = \frac{1}{\hat{\sigma}\sqrt{n}}\sum_{j=1}^{n}\left(y_j - \boldsymbol{x}_i^{\top}\hat{\boldsymbol{\beta}}\right) = 0,$$

in which the equality to zero is a consequence of the residuals summing to zero when ML or least-squares was used for parameter estimation in a linear regression model that includes an intercept. Hence, for such regression models the first order Hermite polynomial is also removed from model (3). The final test statistic is then given by $\hat{S}_k = \hat{V}_3^2 + \cdots + \hat{V}_k^2$.

The test statistic thus shows a natural decomposition into $k-2$ components. Because of the polynomial nature of the orthonormal functions, the components can be roughly interpreted in terms moments; see Henze & Klar (1996); Henze (1997); Klar (2000); Thas (2010) for detailed discussions on this issue. For example, a large $\hat{V}_3^2$ is an indication that the skewness of the true outcome distribution does not agree with the skewness of the hypothesised normal outcome distribution. Since the latter is zero (symmetric distribution), a large $\hat{V}_3^2$ suggests that the true outcome distribution is skewed. Similarly, a large $\hat{V}_4^2$ suggests that the kurtosises of the true distribution and the hypothesised normal distribution do not agree. The two squared components ($\hat{V}_3^2$ and $\hat{V}_4^2$) are equivalent to the components $\hat{S}_3^2$ and $\hat{S}_4^2$ of Peña & Slate (2006).

Numerical example

Davison (2003, example 8.25) reports data on an experiment in which 48 animals were randomly allocated to 12 groups of four animals. Each group was given one of three poisons

Table 2. Results of the order $k = 4$ smooth test applied to the poison data. Smooth test results for the normal distributional assumption are shown. The two-sided p-values are computed from the asymptotic distribution (p_A) and from 2,000 parametric bootstrap runs (denoted by p_B). Results for the ANOVA models with the survival time and with the reciprocal survival time are shown.

	survival time			reciprocal survival time		
Statistic	value	p_A	p_B	value	p_A	p_B
$\hat{S}_4$	19.4546	0.001	0.008	1.1467	0.887	0.874
$\hat{V}_3$	3.2628	0.001	<0.001	0.9664	0.334	0.370
$\hat{V}_4$	2.9679	0.003	0.012	-0.4612	0.645	0.760

and one of four treatments, resulting in a balanced design. The outcome is the survival time in 10-hour units. An analysis based on an additive two-factor analysis of variance (ANOVA) model (i.e. regression model with dummies coding for the two factors) revealed that both poison and treatment have significant effects on the mean outcome. This dataset will be referred to as the poison data.

Table 2 shows the results from the smooth test of order $k = 4$, and its component tests. At the 5% level of significance we can conclude that the normality assumption is not satisfied. Both the third and fourth order component tests give highly significant results, suggesting that perhaps within each or some poison/treatment groups the outcome shows a skewed distribution with too heavy or too light tails. Given the rather small samples sizes in each group (4 animals), our analysis gives a warning that the ANOVA p-values may not be trusted. Davison (2003) suggested applying a Box–Cox transformation to the outcome to resolve the issue.

We have also applied the smooth tests to the same model, but with the reciprocal survival time as outcome. The results are also presented in Table 2. Now no significant goodness of fit is observed.

Finally, Table 2 also shows the p-values calculated from the asymptotic null distributions. Once more the results illustrate some disagreement between the bootstrap and the asymptotic approximation for the sample size of this example ($n = 48$), but the differences are not as large as in the previous example.

3.3. Zero inflated Poisson regression

Test statistic

The Zero Inflated Poisson (ZIP) distribution is a mixture distribution of a Poisson distribution and a point probability at zero. It thus allows for an excess of zeroes as compared to what is expected under a Poisson distribution. The probability of an excess zero is

quantified through an additional parameter, which is considered here as a nuisance parameter. The ZIP distribution does not belong to the exponential family, and hence ZIP regression models are not within the class of GLMs. Neither does the simplification of Section 2.2 apply.

The mean of the Poisson component is related to the linear predictor $\boldsymbol{x}^\top\boldsymbol{\beta}$ through the log-link, i.e. $g(\mu(\boldsymbol{x};\boldsymbol{\beta})) = \log(\mu(\boldsymbol{x};\boldsymbol{\beta})) = \boldsymbol{x}^\top\boldsymbol{\beta}$. The probability of an excess zero is denoted by the parameter γ, which acts as a nuisance parameter. The score functions for γ and μ are given by

$$\begin{aligned}\frac{\partial}{\partial\mu}\log f(y;\mu,\gamma) &= \frac{y-(1-\gamma)\mu}{\mu} - \gamma\left(1-\frac{\delta_0(y)}{\gamma+(1-\gamma)\exp(-\mu)}\right)\\ \frac{\partial}{\partial\gamma}\log f(y;\mu,\gamma) &= \frac{1}{1-\gamma}\left(\frac{\delta_0(y)}{\gamma+(1-\gamma)\exp(-\mu)}-1\right),\end{aligned}$$

where $\delta_0(y) = 1$ if $y = 0$ and $\delta_0(y) = 0$ otherwise. With these score functions, the MLEs of $\boldsymbol{\beta}$ and γ can be obtained (e.g. iterative estimation scheme).

The polynomial of order one is given by $h_1(y;\mu,\gamma) = [y-(1-\gamma)\mu]/\sqrt{(1-\gamma)\mu+\gamma(1-\gamma)\mu^2}$. Higher order orthonormal polynomials can be computed using the Emerson recursion relation (see e.g. Rayner, Thas & De Boeck (2008)). Polynomials up to order 4 are explicitly given in Appendix C of Rayner, Thas & Best (2009). With these polynomials the score functions cannot be written in the form of (6), and hence the simplification of Section 2.2 does not apply here. Therefore we need all the elements of the information matrix as given in Lemma 1. These elements require the following expectations

$$\begin{aligned}&\mathrm{E}_0\left(h_u(Y_j;\mu_j)\frac{\partial}{\partial\gamma}\log f(Y_j;\mu_j,\gamma)\mid\boldsymbol{x}_j\right) = \frac{h_u(0;\mu_j,\gamma)}{1-\gamma}\\ &\mathrm{E}_0\left(h_u(Y_j;\mu_j)\frac{\partial}{\partial\mu_j}\log f(Y_j;\mu_j,\gamma)\mid\boldsymbol{x}_j\right) = \delta_1(u)\sqrt{(1-\gamma)/\mu_j+\gamma(1-\gamma)}+h_u(0;\mu_j,\gamma)\gamma\\ &\mathrm{E}_0\left(\frac{\partial}{\partial\gamma}\log f(Y_j;\mu_j,\gamma)\frac{\partial}{\partial\mu_j}\log f(Y_j;\mu_j)\mid\boldsymbol{x}_j\right) = \frac{1}{1-\gamma}\left(\frac{\gamma}{\gamma+(1-\gamma)\exp(-\mu_j)}-1\right).\end{aligned}$$

With these expressions, and with the parameters replaced with their MLEs, the matrix $\hat{\boldsymbol{I}}_{\theta\theta} - \hat{\boldsymbol{I}}_{\theta\eta}\hat{\boldsymbol{I}}_{\eta\eta}^{-1}\hat{\boldsymbol{I}}_{\theta\eta}^\top$ can be calculated. In contrast to the two previous examples, this matrix is not diagonal, and thus the smooth test statistic $\hat{S}_k$ cannot be written as the sum of its components.

Numerical example

Kostic et al. (2015) investigated the gut microbiome of 33 infants who were genetically predisposed to develop type I diabetes (T1D). The infants were followed during 3 to 4

years in a longitudinal study. At regular visits stool samples were taken for microbiome analysis through 16S rRNA sequencing, resulting in abundance data for 2,239 OTUs (OTU = operational taxonomic unit, which is a proxy for a microorganism species identification). Here we consider only the data of the last visit of each of the 33 infants; we also know the age of the child at this last visit (in days) and whether the child was diagnosed with T1D or not. One of the original research questions in this study was to test for differential abundance of microbial species between T1D cases and the healthy infants, while correcting for age. This should be tested for each OTU separately. The microbiome data come as counts from the sequencing technology. For each biological sample, the sum of the counts of all OTUs is known as the library size, which varies substantially between the samples and which is considered to be an irrelevant technical artefact. A typical data analysis starts with assuming a count distribution, and modelling the log-transformed mean parameter of this distribution as $\log(\text{lib. size}) + \beta_0 + \beta_1 \text{T1D} + \beta_2 \text{AGE}$, with $\log(\text{lib. size})$ an offset, T1D the 0/1 disease indicator, and AGE the age of the child. Several count distributions have been proposed for OTU count data: ZIP, negative binomial and zero inflated negative binomial (see e.g. Xu et al. (2015)). Here we test the ZIP distributional assumption in the regression model. We only present the results for two OTUs. The data are shown in Table 1 in Supporting and Supplementary Material.

As for the Poisson regression example, we use a smooth test of order $k = 4$, and p-values were computed based on 2,000 parametric bootstrap runs. Results are shown in Table 3. The table also shows the results for testing the Poisson distribution with the smooth test of Section 3.1.

For the OTU 195929 data the Poisson model is problematic because the V_3 component is significant at the 5% level. However for the ZIP model all components and S_4 have p-values greater than 0.05 and we can conclude this is an acceptable model. For the OTU 1954177 data both the Poisson and ZIP models have two components significant at the 0.05 level and neither model is acceptable.

It is worth noting that while all components have the same asymptotic null distributions in all models, in small samples their null distributions no longer coincide. A similar comment applies to S_4. Thus for OTU 1954177, the S_4 p-values for the ZIP and the Poisson models of 0.006 and 0.088, respectively, do not necessarily indicate that the Poisson is a more acceptable model than the ZIP. The log likelihood at the MLEs for the Poisson is -118.37 with three degrees of freedom while that for the ZIP is -99.32 with four degrees of freedom. Since the class of ZIP models includes the class of Poisson models the favoured ZIP model will always be at least as good a model as the favoured Poisson model.

As for the two previous examples, the p-values based on the asymptotic null distributions are also reported in Table 3. We observe a strong deviation between the asymptotic and

Table 3. Results of the order $k = 4$ smooth test applied to two OTUs of the infant gut microbiome example. Smooth tests for the ZIP and the Poisson distributional assumption are shown. The two-sided p-values are computed from 2,000 parametric bootstrap runs (denoted by p_B) and from the asymptotic distribution (p_A). For the ZIP, the parameter ω equals 1.

	OTU 194177					
	Poisson			ZIP		
Statistic	value	p_A	p_B	value	p_A	p_B
$\hat{S}_4$	2782.015	<0.001	0.088	101.941	<0.001	0.006
$\hat{V}_1/\omega$	3.924	<0.001	<0.001	2.256	0.024	0.470
$\hat{V}_2$	17.894	<0.001	<0.001	9.185	<0.001	<0.001
$\hat{V}_3$	16.748	<0.001	0.240	6.285	<0.001	0.001
$\hat{V}_4$	46.538	<0.001	0.212	1.333	0.183	0.118
	OTU 195929					
	Poisson			ZIP		
Statistic	value	p_A	p_B	value	p_A	p_B
$\hat{S}_4$	63.776	<0.001	0.489	13.301	0.010	0.544
$\hat{V}_1/\omega$	-0.434	0.664	0.257	1.996	0.046	0.541
$\hat{V}_2$	7.583	<0.001	0.753	3.246	0.001	0.384
$\hat{V}_3$	-2.063	0.039	0.014	0.426	0.670	0.544
$\hat{V}_4$	1.348	0.178	0.311	-1.066	0.286	0.173

bootstrap p-values, which often results in opposite conclusions at the 5% level of significance. Recall that the asymptotic approximation for the data analysis presented in Table 2 was better.

The results of all three data examples make us conclude that the discrepancies will vary with both the model and the sample size. In the next section we will empirically demonstrate that the bootstrap succeeds in controlling the type I error rate and is hence to be recommended.

4. Simulation study

The type I error rate and power of the order $k = 4$ smooth test and its component tests are evaluated in a simulation study. By no means do we intend to present a comprehensive simulation study that includes smooth tests for many different distributional assumptions. Instead we only show the results for Poisson and ZIP regression models for illustrative purposes.

All results are based on 2,000 Monte Carlo runs, and p-values are computed from 200 parametric bootstrap runs. All tests are performed at the 5% level of significance.

4.1. Poisson regression

In each Monte Carlo simulation run, we simulated $n = 15$ observations from a negative binomial (NB) regression model with $\log \mu(x) = 2.6 + 2x$ with x taking values 0, 0.5 and 1, each for $n/3$ of the simulated observations. In each simulation run, a Poisson model with mean model $\log \mu(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x$ is fitted to the data (i.e. no mean-model misspecification). The variance of a negative binomial distribution with mean μ is given by $\mu + \tau\mu^2$, with τ the overdispersion parameter. With $\tau = 0$, the NB collapses to a Poisson distribution. For a range of values for τ, the results are shown in the top panel of Figure 1. The graph shows that all bootstrap tests control the type I error rate. As expected, the overdispersion is best detected with the second order component test (V_2), but the power of the order 4 smooth test (S_4) is not much less.

In a second set of simulations, $n = 25$ observations are simulated with a Poisson distribution with $\log \mu(x) = 1 + 3x_1 + \zeta x_2$ with (x_1, x_2) taken values in the 5×5 grid pattern generated with $x_1 \in \{-1, -0.5, 0.5, 1\}$ and $x_2 \in \{-1.2, -0.7, -0.2, 0.3, 0.8\}$, with $n/25$ observations in each point, and with $\zeta \in [0, 0.5]$. The generated data, however, are analysed with a misspecified Poisson model with only one regressor (x_1). The results are shown in the bottom panel of Figure 1. The bootstrap tests control the type I error rate. The order 4 smooth test has good power. However, despite the first moment of the model being misspecified, it is the second order component test (V_2) that gives the largest power. The missing regressor in the model causes the data to appear overdispersed. Hilbe (2011) called this kind of situation 'apparent overdispersion' to distinguish it from cases where the distributional assumption really is violated. This suggests that one should apply goodness of fit tests always in combination with other diagnostic tools for assessing the correctness of the mean model (e.g. Pearson or deviance residual plots).

4.2. ZIP regression

Here we test the null hypothesis of a ZIP outcome distribution with $\log \mu(x) = 1 + \beta x$. Two simulation scenarios are considered.

In the first case we sampled from a zero inflated negative binomial distribution with a probability of 0.2 for zero-inflation, and $\log \mu(x) = 1 + x$ in which x is standard normally distributed. The conditional variance of the outcome is given by $\mu + \tau\mu^2$, with τ the overdispersion parameter of the NB. The parameter τ takes the values $\tau = 0.0, 0.2, \ldots, 2.0$.

In the second case we simulated data from a ZIP-regression model with $\log \mu(\boldsymbol{x}) = 1 + x_1 + \zeta x_2$ in which x_1 and x_2 are standard normally distributed and $\zeta = 0.0, 0.1, 0.2, \ldots, 1.2$. For $\zeta \neq 0$ the hypothesised ZIP-regression model has a misspecified mean model.

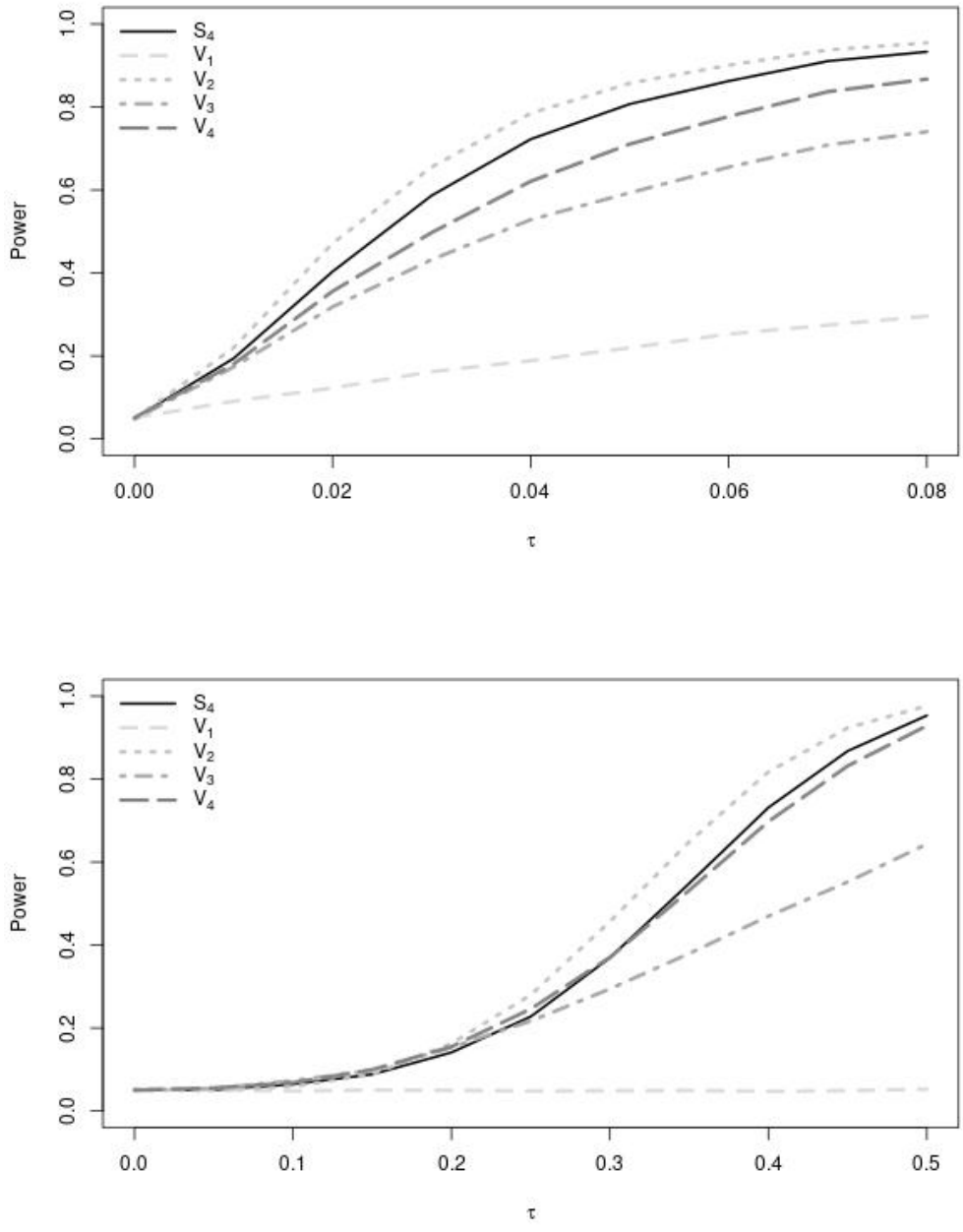

Figure 1. Powers of the smooth test and its component tests for the Poisson regression model. All results are based on 2,000 Monte Carlo simulation runs and 200 parametric bootstrap runs. Top: data generated with negative binomial regression model with overdispersion parameter τ. Bottom: data generated with Poisson regression model with a misspecified mean model that includes an additional term ζx_2.

From the results (Figure 2) it can be seen that V_1 is not very useful for detecting the considered alternatives, but the other four tests detect the alternatives well. The order 4 smooth test (S_4) appears most powerful and can be generally recommended. All tests control the type I error rate.

5. Conclusion

Many regression models contain a distributional assumption for the outcome, conditional on the regressors, allowing for maximum likelihood estimation of the regression parameters. An important class of such models is formed by Generalised Linear Models (GLM). We have developed smooth goodness of fit tests for testing this distributional assumption. The construction starts from sets of polynomials that are orthonormal to the conditional outcome distribution, and thus for each observed regressor another set of orthonormal functions must be computed. The smooth test statistic is a score test statistic developed in a larger class of distributions that embeds the hypothesised regression model. Our methods correctly account for the estimation of nuisance parameters (i.e. regression parameters and other parameters of the conditional outcome distribution). The test statistic asymptotically has a chi-squared null distribution, but simulation studies have shown a slow convergence to the limiting distribution. Therefore it is suggested the parametric bootstrap should be used for p-value calculation. The test statistic is build up from components each of which can also be used as a test statistic. These tests can be helpful in understanding in what sense the data deviate from the hypothesised distribution in terms of e.g. skewness and kurtosis. Even when the data analyst does not want to use formal hypothesis testing for the assessment of the distributional assumption, these components are informative and may be helpful in exploring potential deviations from the hypothesised distribution.

The first component is based on a contrast between the observed outcomes and the fitted mean model. From this perspective it may seem like a statistic for assessing the correct specification of the mean model, but when higher order moments are misspecified by the distributional assumption, its diagnostic property may be lost; see e.g. Klar (2000) for a similar issue with the one-sample smooth tests. Also, a misspecified mean model may cause the higher order components to give small p-values, even when the distributional assumption holds true. The 'apparent overdispersion' (Hilbe 2011) in the Poisson regression example is an illustration of this issue. As a consequence, we always advise not blindly applying the smooth tests proposed in this paper, but to always complement them with other diagnostic tools for assessing the correctness of the mean model (e.g. Pearson or deviance residual plots).

When the regression model is a GLM or when the score functions of its nuisance parameters are linear combinations of the orthonormal basis functions, the construction of the

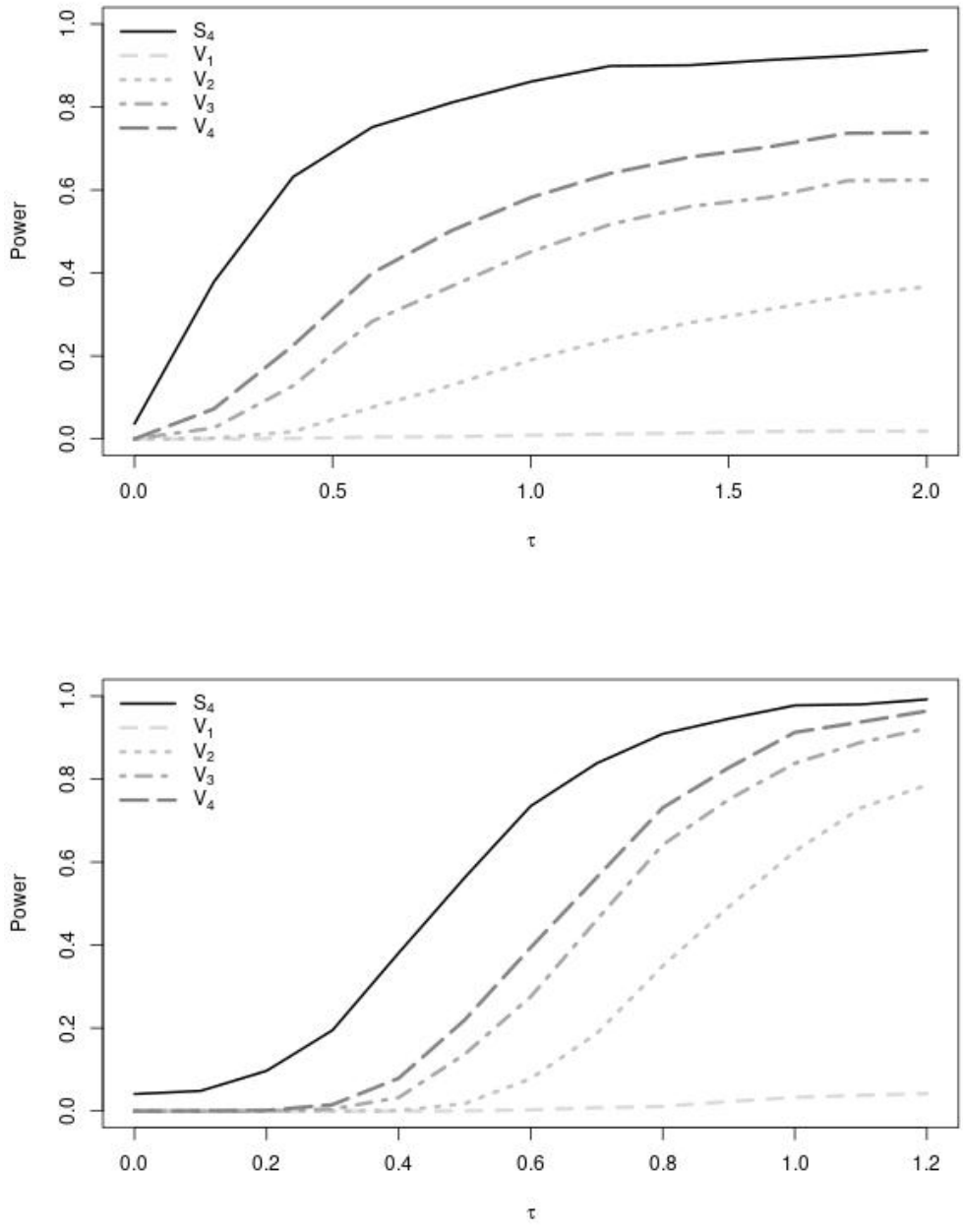

Figure 2. Powers of the smooth test and its component tests for the ZIP regression model. All results are based on 2000 Monte Carlo simulation runs and 200 parametric bootstrap runs. Top: data generated with zero inflated negative binomial regression model with overdispersion parameter τ. Bottom: data generated with ZIP regression model with a misspecified mean model that includes an additional term ζx_2.

smooth test permits simpler expressions. We have given explicit forms of the test statistics for Poisson regression, normal linear regression, zero-inflated Poisson (ZIP) regression and logistic regression. The simulation study, which was restricted to Poisson and ZIP regression, demonstrates that the new tests control the type I error rate when the parametric bootstrap is used. In the PhD thesis of Rippon (2012) smooth tests for more regression models are presented and evaluated in simulation studies. All tests have power for interesting alternative hypotheses. In conclusion, the new tests are customised, focussed and comprehensive.

References

ANSCOMBE, F.J. & TUKEY, J.W. (1963). The examination and analysis of residuals. *Technometrics* **5**, 141–160.

BOOS, D.D. & STEFANSKI, L.A. (2013). *Essential Statistical Inference: Theory and Methods*, vol. 120. New-York: Springer Science & Business Media.

DAVISON, A. (2003). *Statistical Models*. Cambridge: Cambridge University Press.

DEAN, C. & LAWLESS, J.F. (1989). Tests for detecting overdispersion in poisson regression models. *Journal of the American Statistical Association* **84**, 467–472.

HART, J. (2013). *Nonparametric smoothing and lack-of-fit tests*. Springer Science & Business Media.

HENZE, N. (1997). Do components of smooth tests of fit have diagnostic properties? *Metrika* **45**, 121–130.

HENZE, N. & KLAR, B. (1996). Properly rescaled components of smooth tests of fit are diagnostic. *Australian Journal of Statistics* **38**, 61–74.

HILBE, J.M. (2011). *Negative Binomial Regression*. Cambridge: Cambridge University Press.

HUANG, A. (2014). Joint estimation of the mean and error distribution in generalized linear models. *Journal of the American Statistical Association* **109**, 186–196.

HUANG, A. & RATHOUZ, P.J. (2017). Orthogonality of the mean and error distribution in generalized linear models. *Communications in Statistics-Theory and Methods* **46**, 3290–3296.

KALLENBERG, W., LEDWINA, T. & RAFAJLOWICZ, E. (1997). Testing bivariate independence and normality. *Sankhya, Series A* **59**, 42–59.

KAUERMANN, G. & CARROLL, R.J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96**, 1387–1396.

KHMALADZE, E.V., KOUL, H.L. et al. (2004). Martingale transforms goodness-of-fit tests in regression models. *The Annals of Statistics* **32**, 995–1034.

KLAR, B. (2000). Diagnostic smooth tests of fit. *Metrika* **52**, 237–252.

KOSTIC, A.D., GEVERS, D., SILJANDER, H., VATANEN, T., HYÖTYLÄINEN, T., HÄMÄLÄINEN, A.M., PEET, A., TILLMANN, V., PÖHÖ, P., MATTILA, I. et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell host & microbe* **17**, 260–273.

LOVE, M.I., HUBER, W. & ANDERS, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 550.

MARDIA, K. & KENT, J. (1991). Rao score tests for goodness-of-fit and independence. *Biometrika* **78**, 355–363.

MCMURDIE, P.J. & HOLMES, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology* **10**, e1003531.

NELDER, J. & WEDDERBURN, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

PEÑA, E.A. & SLATE, E.H. (2006). Global validation of linear model assumptions. *Journal of the American Statistical Association* **101**, 341–354.

RAYNER, J., THAS, O. & DE BOECK, B. (2008). A generalised Emerson recurrence relation. *Australian and New Zealand Journal of Statistics* **50**, 235240.

RAYNER, J.C.W., THAS, O. & BEST, D.J. (2009). *Smooth Tests of Goodness of Fit: Using R*. Singapore: Wiley.

RIPPON, P. (2012). Application of smooth tests of goodness of fit to generalized linear models. Ph.D. thesis, University of Newcastle, Newcastle, Australia.

RISSO, D., PERRAUDEAU, F., GRIBKOVA, S., DUDOIT, S. & VERT, J.P. (2017). Zinb-wave: A general and flexible method for signal extraction from single-cell rna-seq data. *bioRxiv* , 125112.

SPINELLI, J.J., LOCKHART, R.A. & STEPHENS, M.A. (2002). Tests for the response distribution in a poisson regression model. *Journal of statistical planning and inference* **108**, 137–154.

STUTE, W. & ZHU, L.X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics* **29**, 535–545.

THAS, O. (2010). *Comparing Distributions*. New York, USA: Springer.

THAS, O. & RAYNER, J. (2005). Smooth tests for the zero-inflated Poisson distribution. *Biometrics* **61**, 808–815.

WEDDERBURN, R.W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss?newton method. *Biometrika* **61**, 439–447.

XU, L., PATERSON, A.D., TURPIN, W. & XU, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PloS one* **10**, e0129606.