SPECIFICATION TESTS FOR THE RESPONSE DISTRIBUTION IN GENERALIZED LINEAR MODELS

Bernhard Klar, Simos G. Meintanis

Department of Mathematics, Karlsruhe Institute of Technology (KIT), Kaiserstraße 89, 76133 Karlsruhe, Germany

and

Department of Economics, National and Kapodistrian University of Athens, 8 Pesmazoglou Street, 105 59 Athens, Greece

Abstract. Goodness-of-fit tests are proposed for the case of independent observations coming from the same family of distributions but with different parameters. The most popular related context is that of generalized linear models (GLMs) where the mean of the distribution varies with regressors. In the proposed procedures, and based on suitable estimators of the parameters involved, the data are transformed to normality. Then any test for normality for i.i.d. data may be applied. The method suggested is in full generality as it may be applied to arbitrary laws with continuous or discrete distribution functions, provided that an efficient method of estimation exists for the parameters. We investigate by Monte Carlo the relative performance of classical tests based on the empirical distribution function, in comparison to a corresponding test which instead of the empirical distribution function, utilizes the empirical characteristic function. Standard measures of goodness-of-fit often used in the context of GLM are also included in the comparison. The paper concludes with several real-data examples.

Keywords. Empirical characteristic function, Empirical distribution function, Goodness-of-fit test.

1 Introduction

Let $\{Y_j\}_{j=1}^n$ be independent observations, each following an arbitrary distribution function (DF) $\{F_j\}_{j=1}^n$, respectively. Denote by $\mathcal{F}_\vartheta$ a specific family of distributions indexed by a parameter ϑ. We wish to test the null hypothesis

$$(1.1) \qquad H_0 : F_j \equiv \mathcal{F}_{\vartheta_j}, \text{for some } \vartheta_j \in \Theta,\ j = 1, ..., n,$$

with $\Theta \subseteq \mathbb{R}^q,\ q \geq 1$, i.e. that all F_j belong to a specific parametric family of distributions, and that their only stochastic difference is indexed by the varying parameter ϑ_j. Although goodness-of-fit (GOF) testing is one of the classical problems of inference, most of the standard procedures are restricted to the i.i.d. case. The scenario described above however occupies a central position in a variety of modeling situations, and most importantly in the context of generalized linear models (GLMs). In GLMs, the varying parameter ϑ_j is related to the mean of the underlying distribution. In turn, this mean is linearly related (through a function termed 'the link') to a vector of regressors with value $\mathbf{x}_j$ via a parameter vector β. If $\mathbf{x}_j = (x_{0j}, x_{1j}, ..., x_{pj})^T$ with $x_0 \equiv 1$, and $\beta = (b_0, ..., b_p)^T$, the following cases are popular among practitioners:

- Poisson regression: $F_j = \text{Poisson}(\mu_j)$, with $\mathbf{E}(Y_j) = \mu_j$ and $\log \mu_j = \mathbf{x}_j^T \beta$.
- Logistic regression: $F_j = \text{Bernoulli}(\mu_j)$, with $\mathbf{E}(Y_j) = \mu_j$ and $\log[\mu_j/(1-\mu_j)] = \mathbf{x}_j^T \beta$.
- Negative Binomial regression: $F_j = \text{Poisson}(\mu_j)$, with $\log \mu_j = \mathbf{x}_j^T \beta + u_j$, where $Z_j := e^{u_j} \overset{i.i.d.}{\sim} \text{Gamma}(1, \nu)$, with density $\Gamma(\nu)^{-1} z^{\nu-1} e^{-z}$, and $\mathbf{E}(Z_j) = \nu$.
- Gamma regression: $F_j = \text{Gamma}(\mu_j, \nu)$, with density $\Gamma(\nu)^{-1}(\nu/\mu_j)^\nu y^{\nu-1} e^{-\nu y/\mu_j}$, and $\mathbf{E}(Y_j) = \mu_j$ with $\log \mu_j = \mathbf{x}_j^T \beta$.
- Inverse Gaussian regression: $F_j = \text{Inverse Gaussian}(\mu_j, \lambda)$, with density $\sqrt{\lambda/(2\pi y^3)}$ $\exp\left(-\lambda(y-\mu_j)^2/(2\mu_j^2 y)\right)$, and $\mathbf{E}(Y_j) = \mu_j$ with $\log \mu_j = \mathbf{x}_j^T \beta$.

For GLMs, standard measures of GOF are the so-called deviance statistic and the generalized Pearson χ^2 statistic (McCullagh and Nelder, 1989); modifications of these statistics may be found in Paul and Deng (2002) and Wood (2002). Note also that apart from various attempts to apply the definition of the R^2 measure of linear fit to GLMs (refer to Mittlböck and Heinzl, 2002, and Hu and Shao, 2008), different notions of residuals in GLMs (Pearson, deviance, and Anscombe, residuals) lead to alternative measures of fit based on these residuals; see for instance Shayib and Young (1989). Despite all these attempts there is limited work

towards applying the standard omnibus GOF tests to a specific GLM. These tests include the Cramér-von Mises statistic, and the Anderson-Darling statistic, which utilize the empirical DF (EDF). One reason for the lack of research in this direction may be that when estimation of parameters is involved, these statistics depend not only on the distribution being tested but also on the estimation method and on the true values of the parameters. This dependence of course complicates the implementation of EDF-tests in the context of GLMs, as the asymptotic null distribution of the test statistics, being specific to the hypothesized family $\mathcal{F}_\vartheta$ and the method of estimation, would also depend on both the values of the regressors as well as on the values of the regression parameter β.

In this paper we propose a method in order to circumvent the aforementioned difficulties involved in applying classical EDF-GOF tests in the context of GLMs. In doing so, and in order to illustrate its performance, we apply the proposed method to classical and more recent omnibus GOF tests. The rest of the paper unfolds as follows. In Section 2 we present the new method and indicate how to apply it to the GOF statistics in the context of GLM. Section 3 contains a simulation study in which the omnibus procedures are compared to some measures which are routinely employed by practitioners when trying to assess the fit of specific GLM. In Section 4 certain limitations of the proposed procedure are exposed in the case of i.i.d. data. Finally, Sections 5 and 6 contain real data examples and the conclusions of this study, respectively.

2 Test statistics

2.1 Description of the procedure

In the context of GLMs, consider the data $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{Y}$ denotes the vector with elements Y_j, and $\mathbf{X}$ denotes the matrix with j^{th} row equal to $\mathbf{x}_j^T$, $j = 1, ..., n$. In order to reduce testing for H_0 in (1.1) to tests for composite normality for i.i.d. data we adapt the following procedure originally found in Chen & Balakrishnan (1995):

(i) Consistently estimate ϑ_j by $\widehat{\vartheta}_j$, where $\widehat{\vartheta}_j := \widehat{\vartheta}_j(\mathbf{X}, \mathbf{Y})$.

(ii) Compute $U_j = \mathcal{F}_{\widehat{\vartheta}_j}(Y_j), 1 \leq j \leq n$, and the ordered values $U_{(1)} \leq \ldots \leq U_{(n)}$.

(iii) Compute $\Upsilon_j = \Phi^{-1}(U_{(j)})$, and then $\bar{\Upsilon} = n^{-1} \sum_{j=1}^n \Upsilon_j$, and $S_\Upsilon^2 = (n-1)^{-1} \sum_{j=1}^n (\Upsilon_j - \bar{\Upsilon})^2$.

(iv) Apply the tests in (2.1) and (2.2) with U_j replaced by $\Phi(Z_j),\ j = 1, 2, ..., n$, where

$$Z_j = \frac{\Upsilon_j - \bar{\Upsilon}}{S_\Upsilon},\ j = 1, 2, ..., n. \tag{2.1}$$

Since all distributions in GLMs belong to the exponential family, maximum likelihood estimation of the parameter vector β is always possible in step (i) (see, e.g., McCullagh and Nelder (1989), section 2.2.2). Further unknown parameters may also be estimated by maximum likelihood, or by the method of moments (McCullagh and Nelder (1989), section 8.3.6). Step (ii) transforms to uniformity (under the null hypothesis H_0 and the corresponding DF), apart from sampling variability due to the estimation step. In turn step (iii) renders the observations Υ_j approximately normally distributed. This finally brings us to step (iv), and the classical tests for normality incorporating the standardized observations Z_j, and the standard normal DF, $\Phi(\cdot)$. Note that the transformation in step (ii) is not monotone due to the dependence of the parameter on the observation, and therefore, as a slight modification of the Chen & Balakrishnan procedure, the U_j have to be ordered. We note here that the actual (quantitative) level of dependence induced by estimation in the pairs $(U_j, U_k),\ j \neq k$, would be determined by the specific estimation method employed and by the type of family being tested.

Notice that the U_j in step (ii) are just the *crude residuals* of Cox and Snell (1968). Loynes (1980) analyzed the asymptotic behavior of the empirical process based on the U_j; from this, the asymptotic distribution of EDF based tests can be derived. In practice, however, these results are difficult to apply due to the dependence on the distribution and the true values of the parameters. On the other hand, the quantities Υ_j, $j = 1, \ldots, n$, are the *quantile residuals* introduced by Dunn and Smyth (1996). They differ from the U_j only by a deterministic transformation, but are more suitable for a visual inspection since people are most familiar with the normal distribution. From this point of view, the Z_j in (2.1) are a different type of residuals, and could be termed *standardized quantile residuals*. For visualization purposes, for example Q-Q plots, there is usually not much of a difference between the last two types of residuals; refer for instance to Example 2, in §5.1.

The procedure described by (i)–(iv) works well for continuous DF, but should be suitably adjusted when $\mathcal{F}_\vartheta$ is discrete. In this connection, we follow the modification proposed by Dunn and Smyth (1996) which led to the definition of the so-called *randomized quantile residuals*. According to this modification, for discrete DF step (ii) should be replaced by:

(ii) Let $a_j = \lim_{y \uparrow Y_{(j)}} \mathcal{F}_{\widehat{\vartheta}_j}(y)$ and $b_j = \mathcal{F}_{\widehat{\vartheta}_j}(Y_{(j)})$, and generate U_j as a random deviate

following a uniform distribution on the interval $(a_j, b_j]$.

2.2 Tests for normality

The method proposed in the last subsection reduces the problem of testing H_0 in (1.1), to that of testing normality with estimated parameters. To recall some popular normality tests assume for a minute that we have i.i.d. observations $\{X_j\}_{j=1}^n$ in ascending order. Then the Cramér-von Mises and the Anderson-Darling statistics are given by (refer to D'Agostino & Stephens, 1986 or Thode, 2002),

$$\text{(2.2)} \qquad \mathrm{W}^2 = \sum_{j=1}^{n} \left(U_j - \frac{2j-1}{2n} \right)^2 + \frac{1}{12n},$$

$$\text{(2.3)} \qquad \mathrm{A}^2 = -n - \frac{1}{n} \sum_{j=1}^{n} \left[(2j-1) \log U_j + (2n+1-2j) \log(1-U_j) \right],$$

respectively, where $Z_j = (X_j - \bar{X})/S_X$ are the standardized observations transformed to $U_j = \Phi(Z_j)$ by using the standard normal DF $\Phi(\cdot)$, with $\bar{X} = n^{-1} \sum_{j=1}^{n} X_j$ and $S_X^2 = (n-1)^{-1} \sum_{j=1}^{n} (X_j - \bar{X})^2$. Asymptotic percentage points and modifications of the statistics for finite sample size can be found in Table 4.7 in D'Agostino & Stephens (1986).

Turning now to competitive procedures other than the EDF tests, we would also like to consider a recent test for normality which in the GLM-context of (1.1) takes the form

$$\text{(2.4)} \qquad CF = n \int_{-\infty}^{\infty} |\varphi_n(t) - e^{-(1/2)t^2}|^2 \, \frac{1}{\sqrt{2\pi}} \, e^{-(1/2)t^2} dt,$$

where $\varphi_n(t) = n^{-1} \sum_{j=1}^{n} e^{itZ_j}$ is the empirical characteristic function of Z_j, $j = 1, ..., n$, which are computed from (2.1). It is clear from (2.4), that instead of comparing the EDF to the hypothesized DF as in the case of the W^2 and A^2 statistics, the test statistic CF compares the empirical characteristic function to the hypothesized characteristic function. After some straightforward algebra we have from (2.4),

$$CF = \frac{1}{n} \sum_{j,k=1}^{n} e^{-(Z_j - Z_k)^2/2} - \sqrt{2} \sum_{j=1}^{n} e^{-Z_j^2/4} + \frac{n}{\sqrt{3}} \, .$$

Epps & Pulley (1983) proposed this test statistic and showed that CF is very competitive to the classical EDF tests. An approximation to the limit distribution of CF under normality based on Johnson distributions is derived in Henze (1990). Using a simple transformation of CF, the test can also be carried out for finite samples as long as sample size is larger than or equal to 10 (Henze, 1990, p. 17).

2.3 Deviance statistics

An often-used goodness of fit measure for generalized linear models is the scaled deviance

$$D = 2(\log L(y) - \log L(\hat{\mu}))/\varphi,$$

where $L(y)$ is the likelihood of the saturated model, $L(\hat{\mu})$ is the likelihood under the model considered and φ is the dispersion parameter of the pertaining exponential family of distributions which is assumed to be known. For Poisson and logistic regression, φ equals 1. The same holds for the negative binomial model which also belongs to the exponential family if we assume that the parameter ν is known. For gamma and inverse Gaussian regression models, φ equals $1/\nu$ and $1/\lambda$, respectively. In the case of the normal distribution, the scaled deviance has a $\chi^2_{n-(p+1)}$ distribution where $p+1$ is the number of regression parameters. For other distributions in the exponential family, a similar assertion may be approximately correct; however, it is now well recognized that this approximation can be very poor even for large sample sizes (Davison (2003), p. 483).

For the models with continuous response, distribution theory for the deviance is further complicated by the fact that φ is generally unknown. In this case, φ has to be estimated which can be done by maximum likelihood or by the moment method, using the residual chi-squared statistic divided by the residual degrees of freedom.

The deviance is often used with the upper tail as critical region. However, we used a two sided test which seems more appropriate here as an omnibus test. As critical values, we used quantiles of the χ^2_{n-p-1}-distribution. For more details on deviance–type statistics, as well as for some other measures of GOF in the context of GLMs the reader is referred to Zheng (2000).

If there are several measurements with identical covariable values, one can use the scaled deviance taking into account the group structure

$$D_{group} = 2(\log L(\bar{y}) - \log L(\hat{\mu}))/\phi,$$

where $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_m)$ is the vector of means of the different groups. Again, we considered two sided tests based on D_{group} using quantiles of the chi-squared distribution with $m-p-1$ degrees of freedom as critical values. For grouped data, this approximation improves if the number of observations increases in each group.

3 Simulations

3.1 Continuous response

In this and the following section, we report results of several simulation studies. Since we are mainly interested in the performance for small and medium sample sizes, we used simple models with a small number of parameters. For all computations, we used the statistics software R (R Development Core Team, 2011). Estimation of the parameter vector β is done by maximum likelihood, while the dispersion parameter is estimated by the method of moments. Computation of (randomized) quantile residuals for generalized linear models can be done using the function `qresiduals` in the R library `statmod` (Smyth, 2011).

For the hypothesis of a GLM with inverse Gaussian distribution, we considered the following two models.

Model $IG1$. $\mu_j = \exp(2.6 + 2x_{1j})$, for $j = 1, \ldots, n$. The covariate x_{1j} has one third of the values equal to each of $0, 0.5$, and 1. The means μ_j are then $13.46, 36.60$ and 99.48, each for one third of the sample. As sample size, we took $n = 15$, $n = 30$ and $n = 90$. In the null model, the response variable Y_j is inverse Gaussian with dispersion parameter $\phi = 0.5$ corresponding to $\lambda = 2$ in the usual parametrization.

Model $IG2$. $\mu_j = \exp(0 + 3x_{1j})$, for $j = 1, \ldots, n$. The covariate x_{1j} has one third of the values equal to each of $0, 0.5$, and 1. The μ_j are then $1.00, 4.48$ and 20.09, each for one third of the sample. Again, we put $\lambda = 2$.

Model $IG3$, a model with low mean values: $\mu_j = \exp(-2 + 2x_{1j})$, for $j = 1, \ldots, n$ with the same values of x_{1j} as in $IG1$ and $IG2$. The μ_j are then $0.135, 0.368$ and 1, each for one third of the sample. Again, λ is set to 2.

Other models considered include those with gamma distributed response variable, with the same mean vectors as those of $IG1, IG2$ and $IG3$, and dispersion or shape parameter equal to 1; these models are denoted by $\Gamma 1, \Gamma 2$ and $\Gamma 3$.

We also considered cases of the standard Gaussian linear model; in particular, we choose $\mu_j = 1 + 4x_{1j}$ (denoted Normal 1) and $\mu_j = 1 + 8x_{1j}$ (Normal 2) with the same values of x_{1j} as above. The μ_j are then $1, 3, 5$ and $1, 5, 9$, respectively.

Results:

Table 1 shows that the new tests maintain their theoretical level very well. This also holds for the models $IG3$ and $\Gamma 3$ which have low mean values in each of the three groups. There is no marked difference between the tests based on W^2, A^2 and CF.

		W^2		A^2		CF		D		D_{group}	
	n	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
	15	10.5	5.2	11.6	5.6	11.9	6.1	86.2	81.3	23.1	16.4
$IG1$	30	9.9	5.0	10.0	5.1	10.4	5.5	89.9	86.7	19.6	13.0
	90	9.3	4.9	9.3	4.8	9.1	4.7	86.4	82.9	14.8	9.0
	15	9.6	4.9	9.7	4.7	9.7	4.5	45.9	36.9	13.9	8.3
$IG2$	30	9.5	4.9	9.5	4.8	9.2	4.6	48.3	39.4	12.3	7.0
	90	8.7	4.3	8.5	4.1	8.4	3.8	53.5	45.4	10.4	5.3
	15	9.4	4.5	9.5	4.8	9.5	4.5	1.6	0.6	8.2	4.1
$IG3$	30	9.6	4.9	9.6	4.9	9.9	4.9	2.3	0.8	7.8	3.9
	90	10.0	5.1	10.0	5.1	10.2	4.9	3.5	1.3	8.1	3.7
	15	9.8	4.8	10.1	4.9	10.0	4.8	17.3	9.9	10.3	4.9
$\Gamma1$	30	10.1	5.1	9.9	5.1	10.1	5.2	23.2	14.1	10.2	5.0
	90	9.7	4.8	9.8	4.8	9.9	4.6	36.9	26.0	9.8	4.9
	15	9.6	4.8	9.8	4.8	9.8	4.9	17.7	10.5	9.7	4.9
$\Gamma2$	30	9.7	4.9	9.6	4.8	9.9	5.0	23.0	14.2	9.9	4.9
	90	9.1	4.7	9.5	4.9	9.1	4.4	37.1	25.9	10.2	5.2
	15	10.1	5.1	10.4	5.1	10.4	5.4	17.3	10.0	9.6	4.5
$\Gamma3$	30	10.1	5.1	10.1	5.1	10.1	5.3	23.2	13.8	9.9	4.7
	90	9.9	5.0	9.8	5.0	10.0	4.8	37.1	25.7	10.2	4.9
	15	10.6	5.2	10.7	5.1	10.4	5.0	-	-	9.0	4.0
Normal 1	30	10.0	5.1	10.0	5.1	9.9	4.9	-	-	9.7	4.7
	90	9.4	4.6	9.7	4.7	9.6	4.9	-	-	10.7	5.1
	15	9.2	4.7	9.2	4.6	9.2	4.7	-	-	9.0	3.8
Normal 2	30	10.1	4.8	9.9	4.9	9.8	5.1	-	-	9.6	4.8
	90	9.7	5.0	9.6	5.0	9.8	4.9	-	-	10.0	5.1

Table 1: Empirical levels of the tests based on W^2, A^2, CF, D and D_{group}, theoretical level $\alpha = 10\%$ and 5%, based on 10000 replications

			W^2		A^2		CF		D		D_{group}	
Sim	Est	n	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
		15	61.0	52.6	62.9	54.7	65.8	57.9	82.0	76.3	9.5	5.4
$\Gamma 1$	IG	30	86.5	81.3	87.7	83.1	89.6	85.3	95.1	93.1	8.7	4.5
		90	99.8	99.8	99.9	99.8	99.9	99.9	99.9	99.9	7.7	3.8
		15	68.0	60.3	69.7	62.1	72.7	65.1	79.3	74.0	9.2	4.7
$\Gamma 2$	IG	30	91.3	87.8	92.1	88.7	93.3	90.2	93.4	91.1	9.2	4.9
		90	100.0	100.0	100.0	100.0	100.0	100.0	99.8	99.6	9.0	4.5
		15	48.0	35.7	50.8	39.2	54.3	42.4	6.6	1.6	11.4	4.5
$IG1$	Γ	30	83.9	76.8	87.5	81.1	90.3	84.4	33.7	25.0	11.1	4.8
		90	100.0	99.9	100.0	100.0	100.0	100.0	97.6	95.9	10.5	4.9
		15	17.7	11.6	18.6	12.5	20.3	13.9	4.0	1.2	9.6	4.0
$IG2$	Γ	30	28.0	19.1	32.5	22.6	35.6	25.7	18.6	11.1	10.9	5.2
		90	66.8	53.6	81.3	69.7	82.6	72.2	80.0	73.5	11.6	6.1

Table 2: Empirical power of the tests based on W^2, A^2, CF, D and D_{group}, $\alpha = 10\%$ and 5%, based on 10000 replications

The test based on the scaled deviance D, hence ignoring the grouping, does not maintain its level neither for the gamma nor for the inverse Gaussian distribution and should not be used. In particular, the test is extremely liberal for the first two inverse Gaussian models, but conservative for the low mean model $IG3$. For all three models with gamma distributed response, the empirical level is much higher than the theoretical one. For the normal distribution with estimated dispersion parameter, D equals the mean of a χ^2_{n-2}- distribution, i.e. $D = n - 2$, and a test based on D makes no sense.

On the other hand D_{group}, although liberal for small sample size at least under the first and second IG model, shows an observed level which seems to converge to the theoretical one as the sample size increases, even if the dispersion parameter is estimated.

In Table 2, we present the results of a small simulation to study the empirical power of the different tests. In the upper part, data are generated according to models $\Gamma 1$ and $\Gamma 2$, and the hypothetical model is inverse Gaussian. In the lower part, we assume a gamma regression model, but generate data from $IG1$ and $IG2$.

The first three tests show a similar behaviour with power clearly increasing with sample size. The characteristic function test has a slight edge over the test based on A^2, with the based on W^2 being the least powerful, but differences in power are not pronounced.

Turning to the deviance statistics we see that D_{group} has no power at all (we only refer to D_{group} since the D test fails to maintain the theoretical level). At first view, this may be quite surprising. However, for grouped data, deviance tests are specification tests for the mean, but not for the response distribution. Since intercept and the sole regressor are in the model, means are consistently estimated in the saturated as well as the hypothetical model, and deviance becomes small even if it is based on the wrong distributional assumption.

In this and the subsequent section we used designs with replications to be able to compare the results with the test based on the grouped deviance. We also considered random designs, using the same models as above but replacing the values of x_{1j} $(j = 1, \ldots, n)$ by random values from a uniform distribution on $(0, 1)$. The results for the tests based on W^2, A^2, CF and D are nearly unchanged, and, hence, are omitted.

3.2 Discrete response

Poisson regression

As basic models for mean, we used the same models as Spinelli et al. (2002), and a third model with low mean values:

Model 1. $\mu_j = \exp(2.6 + 2x_{1j})$, for $j = 1, \ldots, n$. The covariate x_{1j} has one third of the values equal to each of $0, 0.5$, and 1. The μ_j are then $13.46, 36.60$ and 99.48, each for one third of the sample.

Model 2. $\mu_j = \exp(0 + 3x_{1j})$, for $j = 1, \ldots, n$. The covariate x_{1j} has one third of the values equal to each of $0, 0.5$, and 1. The μ_j are then $1.00, 4.48$ and 20.09, each for one third of the sample.

Model 3. $\mu_j = \exp(-2 + 2x_{1j})$, for $j = 1, \ldots, n$ with the same values of x_{1j} as in Model 1. The μ_j are then $0.135, 0.368$ and 1, each for one third of the sample.

As sample size, we took $n = 15$, $n = 30$ and $n = 90$. In the null model, the response variable Y_j is Poisson with mean μ_j; with means from Model 1 (resp. Model 2/3) above, it is called Poisson 1 (resp. Poisson 2/3). Note that there exist some difficulties in simulating and estimating the Poisson 3 model: if the simulated values are mainly zeros, fitted rates can occur which are numerically 0, and no valid model is provided. These cases (about 9% of all

cases for $n = 15$, 0.7% for $n = 30$ and 0% for $n = 90$) have been discarded in the simulations.

Alternatives are similar to the models given in Spinelli et al. (2002):
(i) The negative binomial distribution is an overdispersed alternative to the Poisson. It has mean μ_j and variance $\mu_j + \mu_j^2/\nu$. The models with means as in Model 1 and Model 2 and $\nu = 5$ are called Nbinom 1 and Nbinom 2, respectively.
(ii) A further common overdispersed alternative is the mixture of two Poisson distributions. Two equally weighted Poisson distributions with means $0.5\mu_j$ and $1.5\mu_j$ have been chosen. The resulting mixtures are called Pmix 1 and Pmix 2.
(iii) The binomial distribution $B(n, p)$ provides an underdispersed alternative. We have chosen $\mu_j = n_j p_j$ and $n_j = \text{floor}(1.2\mu_j)$ (the integer part of $1.2\mu_j$); hence, the probability of success is approximately 0.8.
(iv) The beta-binomial distribution can be used to provide an equally dispersed alternative. This distribution can be defined as binomial distribution with n_j trials and with the probability of success p a random variable from the beta distribution. We have put $n_i = \text{floor}(\mu_j) + 3$ with parameters chosen so that the mean of the beta-binomial distribution is μ_j and the variance equals the mean.

Table 3 shows that the new tests maintain the theoretical level very well. The tests based on W^2, A^2 and CF behave very similarly and power is again clearly increasing with sample size.

The deviance test based on D works very well for the Poisson models 1 and 2. However, for the low mean model Poisson 3, the test is conservative for small sample size but becomes liberal for larger samples. Its power is much higher than that of the other tests for most alternatives; as expected, the equally dispersed beta-binomial alternative can not be detected (this would also be the case for the underdispersed alternatives if we would have chosen a one-tailed test). On the other hand, the power of the test based on D_{group} is poor and has the unpleasant feature of not being increasing with sample size (which suggests an inconsistent test); the reason is the same as in the case of continuous distributions.

Logistic regression

In this section, the hypothetical models are:

Binom 1: $Y_j \sim Bin(10, p_j)$ with $p_j = (1 + \exp(-2 + 4x_{1j}))^{-1}$ for $j = 1, \ldots, n$. The covariate x_{1j} has one third of the values equal to each of $0, 0.5$, and 1. The p_j are then $0.12, 0.5$ and 0.88, each for one third of the sample.

		W^2		A^2		CF		D		D_{group}	
Sim	n	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
	15	9.6	5	10	4.9	9.7	5	10.3	5.1	10.3	5
Poisson 1	30	10	4.9	10.1	5.1	10.2	4.9	10.5	5.2	10	4.9
	90	9.9	5.1	9.8	5.1	9.9	5	10.2	5.3	9.7	4.8
	15	10	5.3	10.2	5.1	10.1	5.1	9	4.5	10	5.3
Poisson 2	30	10.3	5.4	10.5	5.2	10.2	5.1	9.5	4.7	10.3	5.1
	90	10	4.9	10	5	9.9	4.8	11.6	6.2	9.9	5
	15	9.4	4.6	9.4	4.5	9.4	4.6	3.5	1.4	8.7	4.9
Poisson 3	30	9.4	4.7	9.4	4.6	9.2	4.6	7.5	3.8	9.8	5.1
	90	9.4	4.6	9.4	4.5	9.3	4.4	15.5	8	11.3	5.6
	15	21.2	16.5	21.3	16.5	20.8	16.2	99.9	99.8	47.7	40.9
NBinom 1	30	40.5	35.8	40.6	35.9	39.6	35.8	100	100	48.5	41.2
	90	80.7	78.5	80.8	78.9	80.4	78.4	100	100	48.9	41.8
	15	12	6.7	12.6	7	12.2	6.3	77.8	71.1	16.6	10.2
NBinom 2	30	16	9.5	17.4	10.3	17.5	10.3	96	93.8	16.4	9.9
	90	33.7	23.3	37.6	26.6	38.6	27.8	100	100	16.5	10.1
	15	39.4	27	39.2	26.3	34.3	21.7	99.9	99.9	54	47.3
Pmix 1	30	79.5	65.3	80.5	65.5	72.3	52.7	100	100	52.9	46.1
	90	100	99.9	100	100	100	99.9	100	100	52.3	45.4
	15	12.1	6.3	13.2	6.9	13.9	7.5	93.4	90.6	19.3	12.2
Pmix 2	30	15.4	8.7	16.2	9.5	17.8	10.7	99.8	99.6	19.2	11.8
	90	20.6	13	22.7	14.3	25.9	16	100	100	18.2	11
	15	11.4	5.9	11.7	6.2	11.6	6.1	99.9	99.6	12.9	6.7
Binom 1	30	14	7.7	14.9	8.3	16.2	9.2	100	100	12.8	6.6
	90	23.6	14.9	26.6	17.1	30.4	20	100	100	13	6.6
	15	16.4	9.2	17.9	10.1	18.3	10.4	100	99.9	20.9	10
Binom 2	30	25.4	16	29.6	19	30.1	19.7	100	100	23.3	11.8
	90	55.4	41.8	66.7	52.8	64	50.6	100	100	22.1	11.3
	15	92.5	88.4	92.5	88.5	89.5	85.6	60.7	53.7	12.9	7.1
Beta-binom 1	30	99.9	99.9	99.9	99.9	99.8	99.6	61.6	54.7	11.3	6.1
	90	100	100	100	100	100	100	65.9	59.6	9.9	4.8
	15	34.6	23.3	35.4	24.3	33.8	23.4	22.9	16.3	11.1	5.5
Beta-binom 2	30	67.7	56.1	68.3	56.9	65.1	54.3	29.2	21.5	10.4	5.3
	90	99.3	98.4	99.2	98.3	98.4	97	44	35.2	10.2	5.2

Table 3: Empirical level (lines 3-11) and power (lines 12-35) of the tests based on W^2, A^2, CF, D and D_{group} for the Poisson regression model, $\alpha = 10\%$ and 5%, based on 10000 replications

Binom 2: $Y_j \sim Bin(10, p_j)$ with $p_j = (1 + \exp(0 + 2x_{1j}))^{-1}$; x_{1j} as in Binom 1. The p_j are then $0.5, 0.73$ and 0.88, each for one third of the sample.

As alternatives, we have chosen two beta-binomial distributions representing different degrees of overdispersion relative to the binomial distribution. Specifically, we used a binomial distribution with 10 trials and with parameters chosen so that the mean of the beta-binomial distribution is the same as that of Binom 1 and Binom 2, but with variance 3.25 times larger. These models are called called Beta-Binom A1/A2. Similar models but with factor 5.5 between variances are called Beta-Binom B1/B2.

Results for sample size $n = 15, 30$ and 90 are given in Table 4. Again, the newly introduced tests maintain their level very well, and show similar behaviour under alternatives with power clearly increasing with sample size. For the logistic model, the test based on D does not work well: the theoretical level is not maintained. Hence, the results for alternative distributions have little or no meaning. For D_{group}, we observe the same behaviour as in the previous simulations. In particular, power does not increase (or even decreases, slightly) with sample size.

4 The i.i.d. case revisited

The work of Chen and Balakrishnan (1995) shows that their procedure works well for some standard lifetime distributions. This fact was further strengthened by the results in Meintanis (2009). We have done a similar simulation with a larger range of distributions and a higher number of replications to see if the procedure is more widely applicable.

Results of the simulations for sample sizes $n = 10, 20, 40$ based on 50000 replications showed that the procedure does not always work satisfactorily: it works very well for gamma, inverse Gaussian and lognormal distributions; further, it is good for Weibull and acceptable for logistic and t_5 distributions. The results become worse if the degrees of freedom of the t-distribution declines, eventually failing for the Cauchy distribution. This behaviour is not really surprising: for the t-distribution with 2 degrees of freedom and the Cauchy distribution, the second moment does not exist and the standardization in (2.1) is not meaningful. Even if the second moment exists, the speed of convergence of mean and standard deviation to the population counterparts is low for heavy tailed distributions.

Hence, the procedure is not a general purpose procedure, but acceptable for a remarkable

		W^2		A^2		CF		D		D_{group}	
Sim	n	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
	15	9.4	4.8	9.5	4.7	9.4	4.6	10.5	5.1	10.9	6.6
Binom 1	30	9.7	4.8	9.7	4.9	9.8	4.9	12.2	6.3	10.6	5.8
	90	9.7	4.8	10	4.8	10	4.7	21.5	12.6	10.2	5.2
	15	10.4	5.3	10.5	5.3	10.2	5.1	11.5	6	10.3	5.8
Binom 2	30	9.5	4.9	9.6	4.7	9.6	4.8	13	7.1	10.4	5.5
	90	10	5.3	10.1	5.4	10	5	18.8	11.2	9.6	4.8
	15	10.7	5.2	11	5.3	9.8	4.9	91.3	87.5	32.8	25.2
Beta-Binom A1	30	12.4	6.2	12.3	6.1	10.1	4.9	99.6	99.3	31.6	24.3
	90	20.5	11.7	19.6	10.8	10.9	5.5	100	100	30.6	23
	15	18.3	10.7	18.9	10.9	19	11	95.3	92.5	32.3	24.4
Beta-Binom A2	30	33.3	22.4	34	22.8	36.8	24.7	99.9	99.8	32.3	24
	90	79.3	69.3	79.9	69.9	82.2	72.4	100	100	30.6	22.9
	15	12.8	6.5	12.7	6.5	9.7	4.2	97.5	96.4	48	41
Beta-Binom B1	30	19.5	11.1	18.1	10.1	11.1	5.9	99.9	99.9	45.3	37.6
	90	40	26.3	36.9	22.7	12	6.4	100	100	43	35.7
	15	40.8	29.7	41	29.4	40.5	28.7	99.3	99	46	38.8
Beta-Binom B2	30	72	62.2	71.3	61.2	71.2	60.9	100	100	45	37.8
	90	99.2	98.5	99.1	98.2	98.7	97.7	100	100	44.2	36.4

Table 4: Empirical level (lines 3-8) and power (lines 9-20) of the tests based on W^2, A^2, CF, D and D_{group} for the logistic regression model, $\alpha = 10\%$ and 5%, based on 10000 replications

number of hypothetical distributions. We opted not to report the aforementioned simulation results in order to save space. These results however can be obtained from the authors upon request.

5 Real data examples

5.1 Examples with continuous distributions

5.1.1 Example 1

In our first example, we consider a dataset with motor insurance claims in Sweden for the year 1977 analyzed by Hallin and Ingenbleek (1983). The dataset can also be found together with further description and references on the Statistical Science Web under
`http://www.statsci.org/data/general/motorins.html`.
The dataset comprises the following variables.

Response variable `Payment`: Total value of payments
Offset: `Insured`: Number of insured in policy-years
Covariates `Kilometres`: Kilometres travelled per year, ordered factor from 1 to 5, taken as numeric quantity
`Zone`: Geographical zone (Zone 1: major cities)
`Bonus`: No claims bonus. Equal to the number of years, plus one, since last claim.
`Make`: Different car models.
`Claims`: Number of claims.

The total dataset has 1797 observations, among them 295 in Zone 1. In the source given above, one may find the following citation: "The number of claims in each category can be treated as Poisson to a good approximation. The amount of each claim can be treated as gamma. The total payout is therefore compound Poisson." In this example, we want to assess the gamma hypothesis for the individual claims.

For the whole dataset and all examined models, all GOF tests yield a p-value of zero. Hence, we restrict attention to data from Zone 1, as it is done in Faraway (2006), p. 139. For the analysis of the total value of payments, he used a gamma GLM with log-link:
`Payment ~ offset(log(Insured)) + Kilometres + Make + Bonus`.
The results of the GOF tests are given in the second line of Table 5. They range from 0.004

	CF	W^2	A^2
Gamma GLM	0.016	0.004	0.007
log linear model	0.010	0.012	0.011

Table 5: p-values for the models in Example 1

to 0.016. Hence, even for this subset of the original data, the gamma model is questionable. An alternative analysis uses a linear model with `log(Payment)` as response variable. For this model, the p-values of all GOF tests are similar to those of the gamma model (2nd line of Table 5).

5.1.2 Example 2

As a second example, we analyze a data set with the logarithm of the surface temperature and the light intensity of 43 stars in the star cluster CYG OB1. This data set is available as data set `star` in the R library `faraway` (Faraway (2011), the data set `star` gives the log of the light intensity and, in addition, contains data of four giant stars).

We fitted gamma models with the light intensity as response variable, using the canonical link function $1/\mu$, the log link and the identity link. Figure 1 shows the data and the fitted mean functions.

The tests applied to the gamma model with canonical link yield p-values around 0.3; with log link, the p-values are around 0.5; and with identity link, the p-values are around 0.2 (see Table 6). There is no wide difference between the link functions in this example, but the canonical and log link seem to have an edge over the identity link.

Figure 2 shows Q-Q plots of the quantile residuals (υ_j in §2.1) and the standardized quantile residuals (Z_j in §2.1) for the gamma model with log link. One can observe only a slight difference between the plots; both indicate a reasonable fit of the model.

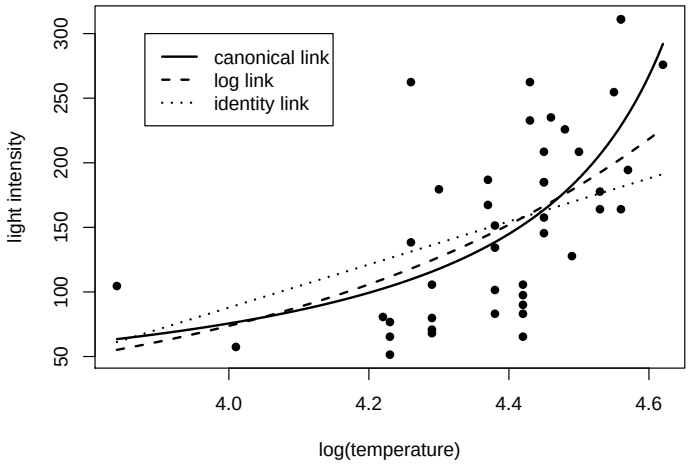

Figure 1: Data points and fitted mean functions in Example 2

	CF	W^2	A^2
canonical link	0.31	0.26	0.30
log link	0.43	0.51	0.53
identity link	0.18	0.23	0.20

Table 6: p-values for the gamma model with different link functions in Example 2

5.2 Examples with discrete distributions

5.2.1 Example 3

Here, we consider again the data set of Example 1, but with the number of claims as response variable. Again, we only consider data of Zone 1.

First we fitted a Poisson model (with canonical link), using the number of insured as offset and Kilometres, Make and Bonus as covariates. The deviance for this model is 782, compared with the null deviance (i.e. the deviance of the model only including the intercept) of 6978. However, since some estimated means are incompatible with the corresponding observations, this model is clearly rejected by all tests.

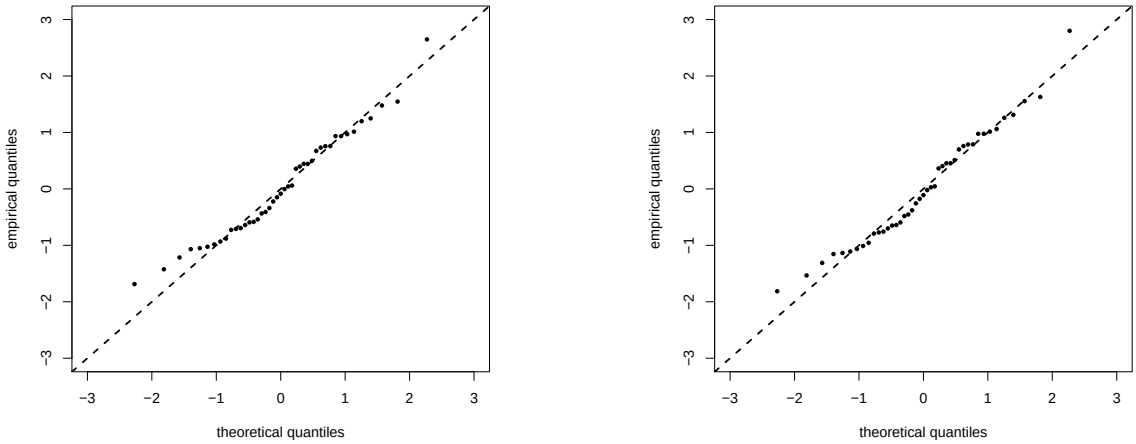

Figure 2: Q-Q plots of the quantile residuals (left) and the standardized quantile residuals (right) for the gamma model with log link in Example 2

As a more flexible alternative, we fitted a negative binomial regression model (again with logarithmic link function). Here, we used the function `glm.nb` in the R library `MASS` (Venables and Ripley (2002)). Deviance and null deviance are 278 and 1157, respectively. Table 7 shows the p-values from three executions of the different (randomized) tests. The last line of Table 7 shows the mean value of 100 replications of the test. Contrary to the Poisson distribution, the negative binomial distribution seems to yield an acceptable model.

	CF	W^2	A^2
1st trial	0.09	0.19	0.12
2nd trial	0.03	0.09	0.05
3rd trial	0.05	0.14	0.09
mean value	0.07	0.18	0.11

Table 7: p-values for the negative binomial model in Example 3

	CF	W^2	A^2
1st trial	0.22	0.43	0.40
2nd trial	0.91	0.74	0.75
3rd trial	0.04	0.15	0.09
mean value	0.54	0.57	0.55

Table 8: p-values for the Poisson model in Example 4

5.2.2 Example 4

As our last example, we analyze data originating from a study of cancer in 4213 male aluminum workers. The data set is used and described in detail in Spinelli et al. (2002). The 4213 workers were divided into 44 groups indexed by weighted years of exposure (treated as numeric predictor) and age (a factor with 11 levels). The observations Y_j of the response variable are the counts of cases of bladder cancer for group j.

A Poisson model (with log link) has a residual deviance of 22.9 on 32 degrees of freedom and null deviance of 70.3 on 43 degrees of freedom. Three replications of the randomized test yield the results in Table 8. Since the fitted means are very small (the median is 0.24), the randomization has a pronounced effect, contrary to the previous example. In such a situation, we recommend also to look at the mean value of several replications of the test; this is given in the last line of Table 8 (for 100 replications).

Fitting a Poisson model without the covariate exposure yields a model with deviance of 33.7 on 33 degrees of freedom. Hence, the covariate has a significant influence on the mean function. This is in agreement with p-values of zero for all goodness-of-fit tests for this sub-model.

Next, we tried negative binomial regression models, again with logarithmic link function. With all predictors, the estimated additional parameter is very large, and the model is not distinguishable from a Poisson model. Without the predictor exposure, the negative binomial model is rejected. From the results, we conclude that the Poisson model with both covariates age and exposure is a very satisfactory model for this dataset.

6 Conclusion

In this paper we propose a method of testing goodness-of-fit for the distribution of observations in the context of GLMs. The main idea is to apply on these observations a variation of a parametric transformation suggested by Chen & Balakrishnan (1995) in the i.i.d. context. As a result of this transformation, the inherent dependence of the null distribution of any goodness-of-fit test statistic on the underlying parameters, regressors, and methods employed in estimating these parameters, is thereby removed and critical points, apart from becoming free of all these specifications, they also remain invariant among different GLMs. The technique reduces the problem to a goodness-of-fit test for normality with estimated parameters, but it is nevertheless suitable for continuous as well as discrete response, by making use of the quantile residuals and the randomized quantile residuals of Dunn & Smyth (1996).

A detailed Monte Carlo study reveals the sampling properties of the proposed methods compared to conventional tests based on the notion of deviance, under some popular GLM situations such as gamma and inverse Gaussian regression, Poisson and logistic regression, as well as the classical normal linear model. In particular, the results show that the deviance statistics may not be appropriate to use under all GLMs as they often lead to distortion of type I error probabilities and consequently power results are hard to assess. On the other hand, the proposed procedures always yield omnibus tests (though in certain cases not as powerful as the deviance), which consistently respect the theoretical level of significance.

Acknowledgement: The authors wish to sincerely thank three anonymous referees for the constructive criticism that resulted to a considerable improvement in the presentation of this work.

References

Chen, G., & Balakrishnan, N. (1995). A general purpose approximate goodness-of-fit test. *J. of Quality Technology*, *27*, 154–161.

D'Agostino, R., & Stephens, M. (1986). *Goodness-of-fit techniques.* Marcel Dekker, Inc., New York.

Davison, A.C. (2003). *Statistical Models.* Cambridge University Press.

Dunn, P.K., & Smyth, G.K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist.*, *5*, 236–244.

Epps, T.W., & Pulley, L.B. (1983). A test for normality based on the empirical characteristic function procedures. *Biometrika*, *70*, 723–726.

Faraway, J.J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Chapman & Hall/CRC.

Faraway, J.J. (2011). *faraway: Functions and Datasets for Books by Julian Faraway,* http://cran.r-project.org/web/packages/faraway/index.html

Hallin, M., & Ingenbleek, J.-F. (1983). The Swedish automobile portfolio in 1977. A statistical study. *Scandinavian Actuarial Journal*, 49–64.

Hardin, J.W., & Hilbe, J.M. (2007). *Generalized Linear Models and Extensions, 2nd Edition.* Stata Press.

Henze, N. (1990). An approximation to the limit distribution of the Epps-Pulley test statistic for normality. *Metrika*, *37*, 7–18.

Hu, B., & Shao, J. (2008). Generalized linear model selection using R^2 . *J. Statist. Plann. Infer.*, *138*, 3705–3712.

Loynes, R.M. (1980). The empirical distribution function of residuals from generalised regression. *Ann. Statist.*, *8*, 285–298.

McCullagh, P., & Nelder, J.A. (1989). *Generalized Linear Models. (2nd ed.)* Chapman and Hall, London.

Meintanis, S.G. (2009). Goodness-of-fit testing by transforming to normality: comparison between classical and characteristic function-based methods. *J. Statist. Comput. Simul.*, *79*, 205–212.

Mittlböck, M., & Heinzl, H. (2002). Measures of explained variation in gamma regression models. *Commun. Statist.–Simul. Comput.*, *31*, 61–73.

Paul, S.R., & Deng, D. (2002). Score tests for goodness of fit of generalized linear models to sparse data. *Sankhya*, *64*, 179–191.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,* http://www.R-project.org.

Shayib, M.A., & Young, D.H. (2002). Modified goodness of fit of tests in gamma regression. *J. Statist. Comput. Simul.*, *33*, 125–133.

Smyth, G. with contributions from Hu, Y., Dunn, P., Phipson, B. (2011). *statmod: Statistical Modeling,* http://CRAN.R-project.org/package=statmod

Spinelli, J.J., Lockhart, R.A., & Stephens, M.A. (2002). Tests for the response distribution in a Poisson regression model. *J. Statist. Plan. Inf.*, *108*, 137–154.

Thode, H.C. (2002). *Testing for Normality.* Marcel Dekker, Inc., New York.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth Edition.* Springer, New York.

Wood, G.R. (2002). Generalized linear accident models and goodness of fit testing. *Accident Analysis and Prevention*, *34*, 417-427.

Zheng, B. (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statist. Med.*, *19*, 1265-1275.