Tue Tjur

Nonlinear regression, quasi likelihood, and overdispersion in generalized linear models.

Summary. The aim of this article is to reconsider the methods for handeling of overdispersion in generalized linear models proposed by McCullagh and Nelder. Our starting point will be a nonlinear regression model with normal errors, specified by a mean function, a variance function and a matrix of covariates. Structurally, this model is very similar to a generalized linear model, except that a common dispersion (or squared scale) parameter is a natural ingredient. In this context, we discuss the estimation method known as IRLS (iteratively reweighted least squares) or quasi likelihood. For generalized linear models, this method coincides with maximum–likelihood. We discuss the proposals made by McCullagh and Nelder for situations where such models fail due to overdispersion. For many such models (in particular for discrete responses), the idea of an overdispersion parameter does not make much sense at first sight. Our approach is based on approximation by a non–linear regression model. In particular, we are interested in the validity of approximate F–tests for removal of model terms and approximate T–distribution–based confidence intervals.

American Mathematical Society 1980 subject classification.
Primary 62J12; secondary 62J02.

Key words and phrases.
Nonlinear regression, quasi likelihood, generalized linear model, iteratively reweighted least squares, overdispersion.

0. Introduction and mathematical ingredients.

The methods for handling of overdispersion in generalized linear models, as described in McCullagh and Nelder (1984, 1989), have created much research activity and a lot of discussion, perhaps not so much about the applicability of these methods, but rather about how they can be justified theoretically. Efron (1986) and Jørgensen (1987) discuss parametric extensions of a generalized linear model by a dispersion parameter. At the opposite extreme we have Wedderburn's (1974) distribution–free quasi–likelihood approach, which relies only on first and second–order properties.

The point of view behind the present paper is that parametric extensions of e.g. Poisson and binomial models to account for overdispersion are unnecessarily complicated, when regarded merely as justifications of the relatively simple methods suggested by McCullagh and Nelder, and that the distributions coming out of this may be difficult to swallow in applied contexts. On the other hand, the distribution–free approach is too weak if one really wants to say something about approximate distributions of estimates and test statistics.

We shall argue that a sufficient and much simpler justification of these methods can be based on approximation by non–linear models with normal responses. It should be emphasized that this idea does not introduce conditions that are much more restrictive than usual in statistical practice. For example, the approximate normality of the counts in a multiplicative Poisson model is the necessary subsumed condition for approximate χ^2 distribution of Pearson's classical goodness–of–fit statistic, or the deviance that it approximates.

All models discussed in this paper will be based on the following four ingredients.

(1) An increasing or decreasing function m, called the *mean function*, which to any real number η in some interval (usually the whole real line) assigns a real number $\mu = m(\eta)$.

(2) A function v , called the *variance function*, which to any $\mu = m(\eta)$ assigns a positive real number $v(\mu)$.

(3) A vector $(w_1, \ldots, w_n)$ of positive real numbers w_i, called the *weights*.

(4) An $n \times p$–matrix $\mathbf{X} = ((x_{ij}))$ of rank p. The i'th row of this contains the covariate values associated with the i'th observation.

Whenever convenient, sufficient smoothness of the two functions will be assumed. As a minimum, we need that m is continuously differentiable and that v is continuous.

1. A class of nonlinear regression models.

Let the observations $y_1, \ldots, y_n$ be independent and normal, with $\mathrm{E}y_i = \mu_i = m(\eta_i)$ and $\mathrm{var}\,(y_i) = \lambda v\,(\mu_i)\,/w_i$ where, in turn, the "linear parameter" η_i is specified as a linear combination $\eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ of the covariate values associated with the observation. The unknown parameters of this model are the coefficients $\beta_1, \ldots, \beta_p$ and the squared scale (or dispersion) parameter λ. The model differs from what is standard in the context of non–linear regression in the following respects. The mean structure is more restrictive than usually assumed, since the dependence of the parameters $\beta_1, \ldots, \beta_p$ is assumed to go via the linear combinations η_i, and the mean function is the same for all units. This condition is not essential for the discussion to follow, it is just made here to force the model into a frame shared by the generalized linear models. On the other hand, the variance of an observation is allowed to depend functionally on its mean, which is usually regarded as a speciality in the non–linear regression context.

The log likelihood function becomes

$$\log\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\lambda v\,(\mu_i)\,/w_i}}\exp\left(-\frac{w_i\,(y_i-\mu_i)^2}{2\lambda v\,(\mu_i)}\right)\right)$$

$$= \mathrm{const} - \frac{1}{2}\left(n\log\lambda + \sum_{i=1}^{n}\log v\,(\mu_i) + \frac{1}{\lambda}\sum_{i=1}^{n}\frac{w_i\,(y_i-\mu_i)^2}{v\,(\mu_i)}\right).$$

However, maximum likelihood estimation in this model can not be recommended in general, for the following reason. In many applications, the interesting part of the model has to do with the mean structure, i.e. the function m and the selection of covariates, whereas the specification of a non–constant variance function v is often just a complication imposed by observed heteroscedasticity. Hence, the functional form of v may very well be more or less ad hoc. Only in exclusive situations, where the shape of the variance function is well–documented or structurally justified, would we really want the specification of v to influence the estimates of the β's in a decisive way. More often, we would prefer the variance function to play a more passive role, similar to the role of fixed weights in a standard linear regression situation, where a misspecification of the weights may result in loss of efficiency, but not in bias.

EXAMPLE. Suppose that y_{ig}, $i = 1, \ldots, n$, $g = 1, 2$, are normal with means $\mu_{ig} = \mu_g$ and $\mathrm{var}(y_{ig}) = \lambda\mu_{ig}^2$. Hence, the covariate structure is that of a one–way analysis of variance model with two groups, the mean function m can be taken to be the identity $\mu = \eta$, and the variance function is $v(\mu) = \mu^2$ (constant coefficient of variation). Since this is

just a special variant of a two–sample setup with variance heterogeneity, it is tempting to estimate the two means by $\hat{\mu}_g = \bar{y}_g$ and λ by some weighted average of the two quantities $s_g^2/\bar{y}_g^2$, where s_g^2 denotes the empirical variance in group g. This is what comes out of the IRLS method discussed below, but it is not what comes out of maximum likelihood. Without going into arithmetic details, it suffices to notice that s_1^2/s_2^2 has an $F(n-1, n-1)$ distribution with scale parameter μ_1^2/μ_2^2, and since s_1^2/s_2^2 is independent of $(\bar{y}_1, \bar{y}_2)$, this will obviously influence the estimation of μ_1 and μ_2; which is what one should expect from a method that naively assumes the specified model to be correct. But in most applications, one would probably prefer to disregard the information about mean structure which is directly inferred from the assumed form of the variance function. In the present example, the obvious solution is to estimate the two group means by the group averages, as we would do in a model with two freely varying variances. But this method is only realistic in general when data can be divided into relatively large covariate groups.

More generally, we can say that the IRLS method will only take the variance function into account by letting its inverse values play the role of externally given weights, whereas it will not attempt to improve the fit of the squared residuals to the assumed shape of the variance function.

IRLS estimation. In the context of a more general nonlinear regression model, Seber and Wild (1989, page 46) suggest the following method. For some reason they present it as a method which is robust against distributional assumptions, without mentioning robustness against misspecification of the variance function. Anyway, the method, which coincides with Wedderburn's (1974) method of quasi likelihood in a similar distribution–free setup, goes as follows. Take as estimates of $\beta_1, \ldots, \beta_p$ those that minimize the square sum

$$\sum_{i=1}^{n} \frac{w_i \left(y_i - \mu_i\right)^2}{v_i} \tag{1}$$

when the constants v_i take the values $v_i = v\left(\mu_i\right)$. Notice that this is *not* equivalent to minimization of the expression when the varying quantities $v\left(\mu_i\right)$ are substituted for the v_i's. What we mean is that the estimates of the linear parameters $\eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ should satisfy the equations

$$\sum_{i=1}^{n} \frac{w_i \left(y_i - m\left(\eta_i\right)\right)}{v\left(m\left(\eta_i\right)\right)} m'\left(\eta_i\right) x_{ij} = 0 \tag{2}$$

for $j = 1, \ldots, p$, which are obtained by formal differentiation of (1) with respect to $\beta_1, \ldots, \beta_p$ when the v_i's are regarded as fixed, followed by substitution of $v\left(\mu_i\right)$ for v_i.

The nature of this definition immediately suggests a method for computation of these estimates, which justifies the name "iteratively reweighted least squares". Algebraically, this method is equivalent to the method proposed by Nelder and Wedderburn (1972), which enabled them to solve the likelihood equations of any generalized linear model by an iterative procedure, where each iteration was formally equivalent to the solution of a weighted regression problem. Let $\mu_i^{(k)} = m(\eta_i^{(k)}) = m(\beta_1^{(k)} x_{i1} + \cdots + \beta_p^{(k)} x_{ip})$ and $v_i^{(k)} = v(m(\eta_i^{(k)}))$ denote fitted values and estimated variances after the k'th iteration. To obtain the next guess, we proceed as follows. Linearize the mean function by Taylor expansion to first order around the present values $\eta_i^{(k)}$, i.e. $m(\eta_i) \approx \mu_i^{(k)} + m'(\eta_i^{(k)})(\eta_i - \eta_i^{(k)})$. Substituting the linear approximations for $m(\eta_i)$ in (1), while at the same time replacing $v(m(\eta_i))$ with $v(m(\eta_i^{(k)}))$ (since the variances can be assumed to vary slowly when convergence is approached), we obtain the weighted square sum

$$(3) \qquad \sum_{i=1}^{n} \frac{w_i \left(y_i - \mu_i^{(k)} - m'\left(\eta_i^{(k)}\right)\left(\eta_i - \eta_i^{(k)}\right)\right)^2}{v_i^{(k)}}$$

$$= \sum_{i=1}^{n} \frac{w_i m'\left(\eta_i^{(k)}\right)^2}{v_i^{(k)}} \left(\eta_i^{(k)} + \frac{y_i - \mu_i^{(k)}}{m'\left(\eta_i^{(k)}\right)} - (\beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \right)^2$$

The next set of estimates $\beta_1^{(k+1)}, \ldots, \beta_p^{(k+1)}$ is obtained by minimization of this approximation to the square sum (1). The last expression for it shows that this minimization is computationally equivalent to the solution of a weighted linear regression with observations $\eta_i^{(k)} + (y_i - \mu_i^{(k)})/m'(\eta_i^{(k)})$, weights $w_i m'(\eta_i^{(k)})^2 / v_i^{(k)}$ and covariates as in the original model.

Once the linear parameters are estimated, we estimate the dispersion parameter λ by

$$(4) \qquad \hat{\lambda} = \frac{1}{n-p} \sum \frac{w_i (y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}.$$

which is merely the usual estimate for the proportionality factor λ in a model where the estimates $\hat{\mu}_i$ are regarded as the true means, with the usual bias correction to compensate this assumption (division by $n-p$ rather than n).

Asymptotics. The asymptotic theory for nonlinear regression models can conveniently be based on the assumption $\lambda \to 0$. Clearly — by reduction to group averages — this also covers an asymptotic setup where

data are sampled in finitely many covariate groups of sizes that increase proportionally towards infinity. Intuitively, it is also rather obvious that similar things will happen when the number of units tends to infinity under suitable less restrictive assumptions about the behaviour of the covariates.

We can summarize the asymptotic theory as follows.

Approximate inference in the non–linear model, including confidence limits for contrasts or regression coefficients based on the T–distribution, F–tests for removal of terms from the model etc., can be based on the usual interpretation of estimates, analysis–of–variance table etc. in the analysis of the "linearized observations"

$$y_i^* = \hat{\eta}_i + \frac{(y_i - \hat{\mu}_i)}{m'(\hat{\eta}_i)}$$

by a linear regression model with mean structure given by $\mathbf{X}$ *and weights*

$$w_i^* = w_i \frac{m'(\hat{\eta}_i)^2}{v(\hat{\mu}_i)}.$$

The proof of this is based on the geometric idea that, as λ tends to zero, the multivariate distribution of the data vector becomes more and concentrated around the true mean vector. This means, that in the region where things happen, the approximation of the manifold of possible mean vectors by its tangent plane (or the first order Taylor expansion of the mean function) becomes more and more accurate, and in the limit we simply have a linear model. Which, not surprisingly, is equivalent to the weighted linear regression solved by the iterative procedure in the last iteration.

Notice also that the usual variance estimate $\hat{\lambda}$ in the linear analysis of the y_i^* coincides with (4). By a similar asymptotic argument, this estimate is approximately χ^2_{n-p} with scale parameter $\lambda/(n-p)$. Confidence bounds for λ can be based on this approximation, and so can the test for the hypothesis $\lambda = 1$ (if it is of any interest, which it is usually not in a proper non–linear regression context).

Just to support intuition, we make the following observation. A first order Taylor expansion of m around $\hat{\eta}_i$ gives

$$m(y_i^*) \approx m(\hat{\eta}_i) + m'(\hat{\eta}_i)\frac{y_i - m(\hat{\eta}_i)}{m'(\hat{\eta}_i)} = y_i$$

or

$$y_i^* \approx m^{-1}(y_i)$$

which means that the final analysis is approximately equivalent to an analysis by the same linear model of the "link–function transformed" responses $m^{-1}(y_i)$ instead of the more complicated "linearized observations" y_i^*. This justifies well–known approximate methods, like estimation in logistic regression models by weighted linear regression of the logit–transformed relative frequencies. However, in the discrete case there are wellknown problems with this method, for example that logit transformed frequencies have to be defined in some more or less arbitrary way for frequencies that are 0 or 1. This method can not be recommended in general, but in some cases it is useful for computation of suitable starting values for the IRLS method.

2. Generalized linear models.

A slightly modified definition of a generalized linear model, as introduced by Nelder and Wedderburn (1972, see also McCullagh and Nelder 1989), goes as follows. As for the non–linear regression model considered in section 1, a matrix $\mathbf{X}$ of covariates and a mean function m (whose inverse is called the *link function* in this context) must be given, and the observations $y_1, \ldots, y_n$ should be independent with means $\mu_i = m(\eta_i) = m(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})$. But the normality assumption is replaced with the assumption that the distribution of the i'th observation has a density (or probability function, in the discrete case) of the form

$$\exp\left(w_i\left(y_i\theta_i - b(\theta_i)\right) + c_i(y_i)\right)$$

where $\theta_1, \ldots, \theta_n$ are one–dimensional parameters. The "weigths" w_i are known constants, and the functions b and c_i are, of course, also assumed to be known. Usually, the expression for c_i will involve the weight w_i, but this possibility is subsumed by our notation. The functions c_i can be regarded as logarithmized densities modifying the underlying measure. The essential condition (which implicitely imposes strong conditions on the functions c_i) is that the logarithmized normalizing function b is independent of i.

For each observation, the family of possible distributions constitutes an exponential family of order 1. By standard arguments (see e.g. McCullagh and Nelder 1989, pp. 28–29), the first two derivatives of the function b are related to the moments of the distribution by the relations

$$\mu_i = b'(\theta_i) \quad \text{and} \quad \operatorname{var}(y_i) = \frac{b''(\theta_i)}{w_i}.$$

Consequently, b' is strictly increasing, and if we define $v = b'' \circ b'^{-1}$ we have the relation $\operatorname{var}(y_i) = v(\mu_i)/w_i$ between mean and variance. Hence, as far as means and variances are concerned, we have a structure similar to the structure of the non–linear regression model studied in section 1.

An important difference is that the variance function is determined by the distributional assumptions, whereas in the non–linear regression context we are free to choose any positive continuous function v.

Another important difference is that the dispersion parameter λ is no longer there (or it has been set to 1, if you like), and there is no obvious way of reintroducing it. This is the topic of section 3.

Wedderburn (1974) proved the following main result, which establishes the link between generalized linear models and the non–linear regression models discussed in section 1.

The maximum likelihood estimates of $\beta_1, \ldots, \beta_p$ in the generalized linear model coincide with the IRLS estimates in the non–linear regression model (same observations y_i, same $\mathbf{X}$, m and v). Moreover, the IRLS method coincides, as an algorithm, with the scoring method (modified Newton–Raphson) for maximization of the log likelihood for the generalized linear model.

This result (which follows easily from the fact that differentiation of the log–likelihood results in the left hand side of (2)) may appear a bit surprising, because one would hardly expect multiplicative Poisson or logistic regression models to be exactly the kind of models where robustness against misspecification of the variance function is a decisive matter. However, the robustness argument was for the normal case, and there is really no paradox here, only a pleasant surprise.

3. Generalized linear models with overdispersion.

This has to do with situations where a generalized linear model is somehow the canonical choice, but the model does not fit due to overdispersion.

The first point to be realized here is that there exists no natural definition of such things as "a binomial distribution with overdispersion" or "a Poisson distribution with overdispersion". If Pearson's χ^2 for goodness of fit turns out to be too large in a supposed Poisson model, the unavoidable conclusion is that we must reject that model, unless we are able to find the additional covariate or the modification of the link function which removes the extra–Poisson variation. Attempts have been made to extend the families of Poisson or binomial distributions by a dispersion parameter, but this turns out to be difficult, and the distributions coming out of this have little intuitive appeal as candidates for distributions of counts; see Efron (1986) and Jørgensen (1987).

Extension by a dispersion parameter is possible for the two most important classes of generalized linear models for continuous responses. For normal models, a dispersion parameter can always be introduced directly as a squared scale parameter (and usually it is there from the

beginning). For (supposed) exponentially distributed responses, extension to the class of Γ–distributions will formally extend the model by a dispersion parameter, and often this is a sensible solution in case of non–negative continuous responses, when variation in scale is the natural tool for description of covariate dependence.

From a theoretical point of view, a generally applicable solution is to allow for random (typically normal) effects in the expression for the linear parameters η_i. Random effects in generalized linear models have been studied recently by several authors, see e.g. Kuk (1995), McGilchrist (1994), Longford (1993). However, random variation on the "unit–to–unit level" (which is what overdispersion could be called in this framework) is a speciality in this context, the usual interpretation of random effects has to do with positively correlated measurements on the same individual, block etc.

Moreover, these methods are far more complicated than required when the generalized linear model considered gives a satisfactory description of the mean structure. The remaining problem is that ignorance of the overdispersion will result in incorrect inference (too narrow confidence bands, too many highly significant rejections etc.) when overdispersion is actually present.

The much simpler methods suggested by McCullagh and Nelder (1989) for this situation can roughly be summarized as follows. Recognizing that our generalized linear model does not hold, whereas all sorts of diagnostic checks (residual plots etc.) still support our belief in the estimated mean structure, confidence limits and tests for model reductions are corrected by an estimate of the dispersion parameter, which in the Poisson and binomial cases can be taken to be Pearson's χ^2 for goodness of fit divided by its degrees of freedom. Deviance differences (or log likelihood ratios, or the weighted square sums approximating them) should be rescaled by this estimate, and approximate standard deviations for linear parameters should similarly be multiplied by the square root of this. In order to correct for the random variation of the estimate of the overdispersion parameter, one may use approximate F–tests and T–distribution–based confidence intervals, rather than approximate χ^2–tests and normality–based confidence intervals.

Obviously, these methods can be justified as consequences of a simple approximation of the generalized linear model by the corresponding non–linear regression model. The formal conditions for validity of this approximation can be stated as follows.

(1) The means are as in the generalized linear model.

(2) The variances are proportional to those of the generalized linear model, i.e. $\mathrm{var}\,(y_i) = \lambda v\,(\mu_i)\,/w_i$.

(3) The observations are approximately normal.

The conditions (1) and (2) are roughly those assumed by Wedderburn (1974) in his quasi likelihood approach. Wedderburn outlined proofs of consistency of the quasi likelihood estimates based on these conditions only. This line of arguing can be continued to provide asymptotic normality of estimates under central limit theorem assumptions which are weaker than assumption (3) above, see McCullagh (1983). However, for approximate χ^2 distribution of the estimate of the dispersion parameter — and thus for the methods based on F– and T–distributions — condition (3) is formally required.

Nevertheless, also in situations where (3) is questionable, common sense suggests that it is better to perform this correction for randomness of $\hat{\lambda}/\lambda$ — implicitely making the (more or less incorrect) assumption that the distribution of $\hat{\lambda}$ is approximated well enough by a χ^2–distribution with $n - p$ degrees of freedom — than not to perform any correction at all — implicitely making the (certainly incorrect) assumption that λ is known and equal to $\hat{\lambda}$. This suggestion is supported by simulation studies of the behaviour of T– versus normal test statistics in case of strongly non–normal responses, which will not be reported here (see Tjur 1995).

4. Example: a log–linear Poisson model with overdispersion.

During the four summers 1992–95, The National Environmental Research Institute, Ministry of Environment and Energy, Denmark, performed a study (Odderskær et al 1997) of the impact of pesticides on the reproductivity of skylarks in spring barley fields. Four fields were treated by either conventional spraying or (essentially) no spraying in a nice cross–over arrangement, see table 1. Many things were registered, and the following statistical analysis clearly represents an oversimplification of this large and complex data set. We shall focus on the summary measure for reproductive success constituted by the total number of fledglings produced on the four fields each of the four years, as given in table 1.

Table 1

Field	Ke	Kr	Ku	Rd
Year				
1992	31	*27	*12	27
1993	*34	60	38	*26
1994	33	*17	*27	26
1995	*24	35	56	*23

(* means 'Sprayed')

A model with multiplicative effects of year, field and treatment appears reasonable here. A multiplicative structure compensates in a natural

way for the fact that the fields are not exactly equally sized, and also a proportional effect of the treatment factor seems natural. Thus, it is tempting to try with a multiplicative or log–linear Poisson model, stating that the number y_{af} of fledglings produced year a on field f is Poisson distributed with parameter

$$\mu_{af} = \exp(\alpha_a + \beta_f + \gamma_t)$$

where $t = t(a, f)$ ($= 1$ for sprayed, 2 for unsprayed) denotes the treatment. After estimation in this model, Pearson's χ^2 for goodness of fit becomes

$$\sum_{af} \frac{(y_{af} - \hat{\mu}_{af})^2}{\hat{\mu}_{af}} = 18.72$$

corresponding to a tail probability in the $\chi^2(8)$–distribution of 0.016, whereas the likelihood–ratio test against the full model results in

$$-2\log(\text{likelihood ratio}) = 18.98$$

corresponding to a tail probability in the $\chi^2(8)$–distribution of 0.015. Thus, there is a clear, though not extremely significant, indication of overdispersion here. Except for this, there is no indication of model failure. The plot of normed residuals against fitted values looks fine, and in fact all the normed residuals are between –2 and 2. If we ignore the warning and continue with the likelihood ratio test for "no treatment effect" ($\gamma_1 = \gamma_2$) we get

$$-2\log(\text{likelihood ratio}) = 24.17$$

corresponding to a tail probability in the $\chi^2(1)$–distribution of 0.000001. An extremely convincing conlusion, it seems. The point estimate of $\gamma_1 - \gamma_2$, with 99% confidence limits based on the usual normal approximation, transforms to the following statement concerning the multiplicative scale: "The proportion between skylark reproduction on sprayed and unsprayed fields is estimated to 0.63, with two–sided 99% confidence limits 0.50 and 0.81".

If, however, we take the indication of overdispersion into account and follow the recommendations of the present paper, we end up with an F–test for treatment effect which becomes

$$\mathrm{F}(1, 8) = \frac{23.62}{18.72/8} = 10.09$$

corresponding to a tail probability of 0.013. Still a significant treatment effect, but far from as convincing as it seemed to be in the multiplicative Poisson model. In this case, one would probably tend to report the

estimate of the relevant treatment parameter with 95% confidence limits, but if we stick to 99%, confidence limits for $\gamma_1 - \gamma_2$ based on the T–distribution on 8 degrees of freedom transform to the somewhat more moderate statement that "The proportion between skylark reproduction on sprayed and unsprayed fields is estimated to 0.63, with two–sided 99% confidence limits 0.39 and 1.03".

The method of correction by the estimated overdispersion parameter, disregarding the randomness of this estimate, results in a "corrected deviance"

$$\frac{-2}{\hat{\lambda}} \log(\text{likelihood ratio}) = \frac{24.17}{18.72/8} = 10.33$$

corresponding to a tail probability in the $\chi^2(1)$–distribution of 0.0013. Notice that this is ten times smaller than the P–value 0.013 produced by the F–test for the same hypothesis.

5. Conclusion.

The Nelder–McCullagh approach to overdispersion is useful and generally applicable to situations where the mean structure coincides with the mean structure of some "underlying generalized linear model", provided that the variance can be assumed proportional to the variance in that model. If the observations are approximately normal (and probably also if they are not) F–tests for model reductions and T–distribution–based confidence intervals — which are only vaguely promoted by Nelder and McCullagh — should in general be preferred to the approximations assuming $\lambda = \hat{\lambda}$, when the number of degrees of freedom for the residual is small. Just as for linear regression and analysis–of–variance models, validity of the model for the mean structure (linearity/additivity assumptions and the link function) is the crucial condition for applicability of the model. Correctness of the implied specification of the variance as a function of the mean is also important, but the possible damages caused by misspecification of the variance function are comparable to the damages caused by misspecification of the weights in a linear regression (or by ignorance of heteroscedasticity in an ordinary unweighted regression). Approximate normality of the responses is the least critical condition. To this we can add, in the spirit of Wedderburn's quasi likelihood approach, that an underlying generalized linear model needs not exist at all, it suffices to know the mean function and the variance function.

Most important of all: Don't ever ignore overdispersion!

A practical remark. My public–domain statistical package ISU for DOS/Windows has a procedure FITNONLINEAR for nonlinear regression in the spirit of the present paper. ISU can be downloaded from my homepage http://www.mes.cbs.dk/~sttt/.

6. Acknowledgements.

Thanks are due to Peter Odderskær et al, The National Environmental Research Institute, Department of Landscape Ecology, for kindly allowing me to extract and make use of the data of Table 1 from their data on skylarks and pesticides.

References.

Efron, B. (1986).
Doubly exponential families and their use in generalized linear regression.
J.A.S.A. **81**, 709–721.

Jørgensen, B. (1987).
Exponential Dispersion Models.
J.R.Stat.Soc. B **49**, 127–162.

Kuk, A. Y. C. (1995).
Asymptotically Unbiased Estimation in Generalized Linear Models with Random Effects.
J.R.Stat.Soc. B **57**, 395–407.

Longford, N. T. (1993).
Random Coefficient Models.
Clarendon Press, Oxford.

McCullagh, P. (1983, 1st ed. 1984).
Quasi–likelihood functions.
Ann.Stat. **11**, 59–67.

McCullagh, P. and Nelder, J. A. (1989).
Generalized Linear Models.
Chapman and Hall.

McGilchrist, C. A. (1994).
Estimation in generalized mixed models.
J.R.Stat.Soc. B **56**, 61–69.

Nelder, J. A. and Wedderburn, R. W. M. (1972)
Generalized Linear Models
J.R.Stat.Soc. A **135**, 370–384.

Odderskær, P., Prang, A., Elmegaard, N. and Andersen, P.N. (1997)
Skylark Reproduction in Pesticide Treated Fields (Comparative Studies of Alauda arvensis Breeding Performance in Sprayed and Unsprayed Spring Barley Fields)
Bekæmpelsesmiddelforskning fra Miljøstyrelsen nr. 32. National Environmental Research Institute. Ministry of the Environment and Energy, Denmark. Danish Environmental Protection Agency.

Seber, G. A. F. and Wild, C. J. (1989).
Nonlinear Regression.
Wiley 1989.

Tjur, T. (1995)
Non–linear regression, quasi likelihood and over–dispersion in generalized linear models.
Preprint 1995–2.
Institute of Mathematical Statistics, University of Copenhagen.

Wedderburn, R. W. M. (1974).
Quasi–likelihood functions, generalized linear models, and the Gauss–Newton method.
Biometrika **61**, 439–447.

Tue Tjur
Copenhagen Business School
Department of Management Science and Statistics
Julius Thomsens Plads 10
DK-1925 Frederiksberg C
Denmark
email tuetjur@cbs.dk