Biometrika (2012), **99**, 1, *pp.* 223–229
doi: 10.1093/biomet/asr075
Advance Access publication 13 January 2012

Miscellanea

Proportional likelihood ratio models for mean regression

By ALAN HUANG

Department of Statistics, University of Chicago, Chicago, Illinois 60637, U.S.A.
alanh@uchicago.edu

and PAUL J. RATHOUZ

Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, Wisconsin 53792, U.S.A.
rathouz@biostat.wisc.edu

Summary

The proportional likelihood ratio model introduced in Luo & Tsai (2012) is adapted to explicitly model the means of observations. This is useful for the estimation of and inference on treatment effects, particularly in designed experiments and allows the data analyst greater control over model specification and parameter interpretation.

Some key words: Empirical likelihood; Exponential tilting; Generalized linear model; Multi-way layout; Proportional likelihood ratio model; Quasilikelihood; Semiparametric model.

1. Introduction

Luo & Tsai (2012) introduced proportional likelihood ratio models as an extension of density ratio (e.g., Fokianos et al., 2001) and biased sampling models (e.g., Vardi, 1985). While these models have proved to be useful in many applied statistical settings, for regression problems, especially those pertaining to designed experiments, treatment effects as measured by contrasts in the mean response have a more immediate interpretation. In this note, we adapt the proportional likelihood ratio model to explicitly model the observation means. We focus particularly on the multi-way layout, noting that data from designed experiments in agriculture, engineering and biometry often come in this form.

The multi-way layout has $K < \infty$ groups or subpopulations, where each group k is defined by levels of a q-vector X_k of covariates. Within group k, a sample of n_k independent and identically distributed observations $\{Y_{k1}, \ldots, Y_{kn_k}\}$ is drawn from a distribution F_k that depends only on X_k. Assume each F_k has a density dF_k with respect to some common dominating measure. A proportional likelihood ratio model parameterized via mean regression in the multi-way layout is then defined by two components. First, there is an explicit mean model,

$$E(Y_{ki} \mid X_k) = \eta(X_k^{\mathrm{T}}\beta), \tag{1}$$

where η is a user-specified inverse link function and β is a q-vector of unknown parameters. Secondly, there is an assumption that the distributions F_k are related via an exponential tilting of some reference distribution F. More precisely, for $k = 1, \ldots, K$,

$$dF_k(y) = \exp(b_k + \theta_k y)\, dF(y), \tag{2}$$

where

$$b_k = -\log \int \exp(\theta_k y)\, dF(y) \tag{3}$$

are normalizing constants and, in order to satisfy (1), each tilt $\theta_k \equiv \theta_k(\beta, F)$ is implicitly defined as the solution to the mean constraint

$$\eta(X_k^{\mathrm{T}}\beta) = \int y \exp(b_k + \theta_k y)\, dF(y). \tag{4}$$

In other words, each density dF_k is an exponential tilt of some reference density dF, with the amount of tilting indirectly determined by the mean $\eta(X_k^{\mathrm{T}}\beta)$ of group k.

The main difference between models (1)–(4) and the proportional likelihood ratio model of Luo & Tsai (2012) is that the linear predictor $X_k^{\mathrm{T}}\beta$ appears explicitly in the mean model (1) and only indirectly in the tilts θ_k in the distributional model (2). It is this mean specification that distinguishes our model from proportional likelihood ratio models, density ratio models and biased sampling models.

For our model to be well-defined, we require the Laplace transform of F to exist in a neighbourhood of the origin. For identifiability, we may use the convention $b_K = \theta_K = 0$, so that F coincides with F_K. Other properties of this model can be found in Rathouz & Gao (2009). In that paper, some favourable results were obtained for the special case of polytomous responses. In contrast, this note deals with the multi-way layout where covariates take on a finite number of levels, quantitative or categorical, but the response space is arbitrary.

When the reference distribution $F \equiv F_K$ is specified, the model reduces to a standard generalized linear model and the loglikelihood function based on (1)–(4) can be used to derive maximum likelihood estimators $\hat{\beta}_{\mathrm{MLE}}$ and to carry out likelihood-based inferences on β in the usual way. For example, if dF_K is a Gaussian kernel, then normal linear regression is recovered; if dF_K is a Poisson kernel, then Poisson regression is obtained. Of course, such estimators will typically be inefficient and likelihood-based inferences will be biased unless the specified F_K coincides with the true underlying error distribution. The main advantage of adopting the exponential tilt perspective of (1)–(4) is that it lends itself immediately to a semiparametric extension that obviates the need to specify, and possibly misspecify, an error distribution. The idea is to leave the reference distribution F_K unspecified, instead treating it as an infinite-dimensional parameter in the loglikelihood function

$$l(\beta, F_K) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\{\log dF_K(Y_{ki}) + b_k(\beta, F_K) + \theta_k(\beta, F_K)Y_{ki}\},$$

where $b_K = \theta_K = 0$ and for $k = 1, \ldots, K-1$, $b_k(\beta, F_K)$ and $\theta_k(\beta, F_K)$ are the joint solutions to equations (3) and (4). Our approach is thus a semiparametric extension of generalized linear models.

2. PROFILE EMPIRICAL LIKELIHOOD

A variety of parametric and nonparametric model specifications for F_K are possible; here we construct an empirical likelihood (Owen, 2001) by replacing the density dF_K in the loglikelihood by nonnegative probability masses $\{p_{ki} : i = 1, \ldots, n_k, k = 1, \ldots, K\}$ on the observed support $\{Y_{ki}\}$. The resulting empirical loglikelihood function is

$$l(\beta, p) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(\log p_{ki} + b_k + \theta_k Y_{ki}),$$

where, for $k = 1, \ldots, K$, each pair (b_k, θ_k) satisfies jointly the normalization and mean constraints

$$1 = \sum_{j=1}^{K}\sum_{i=1}^{n_j} p_{ji} \exp(b_k + \theta_k Y_{ji}), \tag{5}$$

$$\eta(X_k^{\mathrm{T}}\beta) = \sum_{j=1}^{K}\sum_{i=1}^{n_j} Y_{ji}\, p_{ji} \exp(b_k + \theta_k Y_{ji}). \tag{6}$$

Equations (5) and (6) are empirical analogues of the normalization and mean constraints (3) and (4). By setting $b_K = \theta_K = 0$ for identifiability, the probability masses $\{p_{ki}\}$ have the interpretation of a multinomial distribution with mean $\eta(X_K^{\mathrm{T}}\beta)$ on the observed support. For the mean equation (6) to be solvable, β must satisfy $Y_{\min} \leqslant \eta(X_k^{\mathrm{T}}\beta) \leqslant Y_{\max}$ for all k, where $Y_{\min} = \min\{Y_{ki} : i = 1, \ldots, n_k, k = 1, \ldots, K\}$ and $Y_{\max} = \max\{Y_{ki} : i = 1, \ldots, n_k, k = 1, \ldots, K\}$ are the minimum and maximum observations, respectively. This is commonly referred to as the convex hull condition in the empirical likelihood literature. For β outside this range, the convention is to set the empirical loglikelihood to $-\infty$; see Owen (2001, pp. 209–10) for more discussion.

A profile empirical loglikelihood for β can then be defined by

$$l_{\mathrm{p}}(\beta) = \sup_{p} l(\beta, p),$$

where the supremum is taken over all multinomial distributions on the observed support with mean $\eta(X_K^{\mathrm{T}}\beta)$. The maximum empirical likelihood estimator for β is then defined as $\hat{\beta} = \mathrm{argmax}_{\beta} l_{\mathrm{p}}(\beta)$.

If $\hat{\beta}$ satisfies the convex hull condition strictly, that is, if $Y_{\min} < \eta(X_k^{\mathrm{T}}\hat{\beta}) < Y_{\max}$ for all k, and the tilts $\theta_1, \ldots, \theta_{K-1}$ remain finite, then the corresponding maximum empirical likelihood estimate of p, denoted $\hat{p} = \hat{p}(\hat{\beta})$, exists and is unique, by an argument similar to that in Vardi (1985). In this case, we can define a maximum empirical likelihood estimator of the reference distribution $F_K(y)$ by

$$\hat{F}_K(y) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{p}_{ki} 1(Y_{ki} \leqslant y). \tag{7}$$

3. Main results

Let $n = n_1 + \cdots + n_K$ denote the total sample size. In the following asymptotic considerations, let the sample size ratios $\rho_k = n_k/n \to \rho_k^*$ as $n \to \infty$, where $0 < \rho_k^* < 1$ for all $k = 1, \ldots, K$.

In Proposition 1, we establish the consistency and asymptotic normality of the maximum empirical likelihood estimator for β, and show that its asymptotic variance has a negative inverse profile Hessian representation. Proposition 1 also justifies Wald tests and confidence regions for β, provided an accurate estimate of the profile Hessian can be obtained.

Proposition 1. *The equation* $0 = dl_{\mathrm{p}}(\beta)/d\beta$ *admits a solution* $\hat{\beta}$ *that is consistent for the true value* β^*. *Furthermore,* $H_P^{1/2}(\hat{\beta} - \beta^*) \to N(0, I)$ *in distribution as* $n \to \infty$, *where* $H_P = -E\{d^2 l_{\mathrm{p}}(\beta^*)/d\beta d\beta^{\mathrm{T}}\}$.

Likelihood-based inferences are often preferred to Wald-type methods because of asymptotic optimality properties. Another, more practical, advantage is that likelihood-based methods do not require an estimate of the asymptotic variance of $\hat{\beta}$. In Proposition 2, we show that the profile empirical loglikelihood behaves asymptotically like a true loglikelihood for the testing of hypotheses and construction of confidence regions for β. Recall that β is a $q \times 1$ vector.

Proposition 2. *Under* $H_0 : \beta = \beta^*$, $2\{l_{\mathrm{p}}(\hat{\beta}) - l_{\mathrm{p}}(\beta^*)\} \to \chi_q^2$ *in distribution as* $n \to \infty$.

A finite-sample adjustment can be obtained by comparing $2\{l_{\mathrm{p}}(\hat{\beta}) - l_{\mathrm{p}}(\beta^*)\}$ to a $q\mathcal{F}_{q,n-q}$ distribution instead. This is justified because $q\mathcal{F}_{q,n-q} = \chi_q^2 + o_P(1)$ in distribution.

General composite hypotheses $H_0 : M\beta^* = \gamma$ for a given matrix M and vector γ can be reparameterized into the following form. Write $\beta = (\beta^{(1)}, \ldots, \beta^{(q)})^{\mathrm{T}}$ componentwise and let B_0 be a subspace of $\mathbb{R}^q$ given by $B_0 = \{\beta : \beta^{(1)} = \cdots = \beta^{(r)} = 0\}$ for some $r \leqslant q$. Let $\hat{\beta}_0 = \mathrm{argmax}_{\beta \in B_0} l_{\mathrm{p}}(\beta)$ be the maximum empirical likelihood estimator over B_0 and $\hat{\beta}$ be the unconstrained maximum empirical likelihood estimator over $\mathbb{R}^q$. An analogous result for composite hypotheses is then given by the following corollary.

Corollary 1. *Under* $H_0 : \beta \in B_0$, $2\{l_{\mathrm{p}}(\hat{\beta}) - l_{\mathrm{p}}(\hat{\beta}_0)\} \to \chi_r^2$ *in distribution as* $n \to \infty$.

Again, a finite-sample adjustment can be obtained by comparing the profile empirical loglikelihood ratio to an $r\mathcal{F}_{r,n-q}$ distribution instead.

A fundamental issue in generalized linear models is the selection of an appropriate error distribution for the data. To this end, consistent estimation of the underlying error distribution is justified by Proposition 3.

PROPOSITION 3. *As* $n \to \infty$, $\hat{F}_K$ *exists with probability tending to* 1 *and* $n^{1/2}\{\hat{F}_K(y) - F_K^*(y)\} \to N\{0, W(y)\}$ *in distribution, where* F_K^* *is the true reference distribution and* $W(y) = W(y; \beta^*, b^*, \theta^*, F_K^*)$ *is some covariance function.*

Proofs of Propositions 1 and 2 are outlined in the Supplementary Material. The proof of Corollary 1 is straightforward and is omitted. Proposition 3 is analogous to Theorem 1 of Qin & Lawless (1994) and its proof is omitted.

4. SIMULATION STUDY

To examine the performance of the proposed method in practice, we conducted a limited simulation study using small to medium sample sizes, looking at Type I errors under the null hypothesis. Given the level of generality and flexibility of our proposed method, the only truly comparable approach from the existing literature appears to be quasilikelihood coupled with a sandwich estimator of variance. Neither method requires a correctly specified variance function and both are guaranteed to be asymptotically correct, although in smaller samples, a correctly specified working variance model should work better than an incorrect one. As a benchmark, we also looked at the standard Wald test from a correctly specified parametric model.

First, a series of simulations was carried out using continuous data simulated from a 2×2 additive mean model with normal errors,

$$Y_{jk,i} \sim N(\mu_0 + \alpha_1 I_{(j=2)} + \alpha_2 I_{(k=2)}, \sigma^2) \quad (j, k = 1, 2; i = 1, \ldots, n_{\text{rep}}),$$

with intercept $\mu_0 = 2{\cdot}304$, main effects $\alpha_1 = -0{\cdot}012$, $\alpha_2 = 0{\cdot}750$, error standard deviation $\sigma = 1$ and number of replicates $n_{\text{rep}} = 2$, 4 and 8. The parameter values were chosen such that the simulated values and covariate effects were similar to that of the chemical dataset from Myers et al. (2010, p. 74). For each configuration, 10 000 replicate datasets were simulated. The average R^2 of the simulated datasets was around 50%.

A second series of simulations was performed using nonnegative continuous data simulated from a 2×2 log-additive mean model with gamma errors,

$$Y_{jk,i} \sim \text{Ga}\{\mu_{jk} = \exp(\mu_0 + \alpha_1 I_{(j=2)} + \alpha_2 I_{(k=2)}), \nu\} \quad (j, k = 1, 2; i = 1, \ldots, n_{\text{rep}}),$$

with mean parameters $\mu_0 = 5{\cdot}414$, $\alpha_1 = 0{\cdot}0617$, $\alpha_2 = -0{\cdot}15$, shape parameter $\nu = 100$ and number of replicates $n_{\text{rep}} = 2$, 4 and 8. The parameter values were chosen such that the simulated values and covariate effects were similar to that of the resistivity dataset from Myers et al. (2010, p. 221). Again, for each configuration, 10 000 replicate datasets were simulated.

For each simulated dataset, the empirical likelihood ratio test of Corollary 1, referred to as an $\mathcal{F}_{1,n-4}$ distribution, was used to test for an interaction term. The Type I errors were compared with those obtained from using the Wald test from a quasilikelihood plus sandwich estimation approach, referred to as a t_{n-4} distribution. The quasilikelihood plus sandwich estimation approach correctly assumes a constant working variance function $V(\mu) = \sigma^2$ for normal data and a quadratic working variance function $V(\mu) = \phi\mu^2$ for gamma data, but allows for possible misspecification through a sandwich estimator of variance. Computations were carried out using the MINOS optimizer in AMPL (Fourer et al., 2003).

The simulation results are displayed in Table 1. We see that in the smallest scenario of only two replicates, both our proposed method and quasilikelihood with sandwich estimation performed quite poorly compared with a correctly specified Wald test. This should not be surprising, however, since the mean model under the alternative has four parameters, leaving us with only four degrees of freedom to estimate the reference distribution or variance function. This is a formidable task. However, by four replicates, the

Table 1. *Type* I *errors in a* 2×2 *layout with normal additive and gamma log-additive data*

	Normal			Gamma		
	Nominal Type I error (%)			Nominal Type I error (%)		
	1	5	10	1	5	10
Two replicates						
ELRT ($\mathcal{F}_{1,4}$)	0·6	7·1	17·8	0·4	4·9	16·3
QL + SW (t_4)	3·2	13·2	21·7	3·5	12·5	20·2
Wald test (t_4)	0·9	4·8	10·2	1·3	5·4	10·5
Four replicates						
ELRT ($\mathcal{F}_{1,12}$)	0·9	6·6	13·9	1·4	7·3	13·4
QL + SW (t_{12})	1·8	8·5	15·2	2·3	8·8	14·4
Wald test (t_{12})	1·1	5·3	10·0	1·2	5·7	10·5
Eight replicates						
ELRT ($\mathcal{F}_{1,28}$)	0·9	5·9	11·1	0·9	5·6	10·9
QL + SW (t_{28})	1·4	6·8	11·7	1·4	6·4	11·8
Wald test (t_{28})	0·9	5·0	9·9	0·9	4·9	9·3

ELRT, empirical likelihood ratio test; QL + SW, quasilikelihood with sandwich estimator of variance.

empirical likelihood ratio test was reasonably well calibrated by the $\mathcal{F}$ distribution for both normal and gamma data and proves to be quite useable in practice. Furthermore, with the possible exception of two replicates, the empirical likelihood ratio test outperformed the Wald test based on quasilikelihood with a sandwich estimator of variance. The simulation standard errors for 1, 5 and 10% Type I errors are 0·1, 0·2 and 0·3%, respectively.

Empirical likelihood methods typically have larger Type I errors than their nominal values and the convergence is usually from above (e.g., DiCiccio et al., 1991). The observed over-conservativeness at lower significance levels for small sample sizes here is therefore unexpected, but may be due to numerical inaccuracies in the algorithm used in fitting the model, especially for iterations near the boundary of the convex hull. We tried to deal with this by using an additional stability parameter that penalizes iterations that are too close to the convex hull boundary. As with all empirical likelihood methods, the convex hull issue diminishes as sample size increases.

5. Worsted yarn experiment data analysis

An experiment investigating the effects of three factors, x_1, length, x_2, amplitude and x_3, load, on the cycles-to-failure, y, of worsted yarn is described in Box & Cox (1964), also in Myers et al. (2010, p. 234). Each factor took on three quantitative values in a 3^3 factorial design, with the values of each factor normalized and coded as -1, 0 and 1.

Myers et al. (2010) noted that the cycles-to-failure times in this experiment were nonnegative discrete random variables expected to have an asymmetric distribution with a long right tail. Such data are frequently modelled by exponential, Weibull, lognormal or gamma distributions. Ultimately, the data were analysed using a log-linear mean model with a gamma distribution, although no initial justification was provided as to why this error distribution is appropriate. The fitted mean model assuming gamma errors is $\hat{\mu} = \exp(6{\cdot}349 + 0{\cdot}843x_1 - 0{\cdot}631x_2 - 0{\cdot}385x_3)$, with estimated scale parameter $\hat{r} = 31{\cdot}621$.

In contrast, data analysis using our proposed method does not require any specification of an underlying error distribution; instead, we estimate the error distribution and the mean model parameters simultaneously from the data. Our fitted mean model is $\hat{\mu} = \exp(6{\cdot}399 + 0{\cdot}798x_1 - 0{\cdot}599x_2 - 0{\cdot}402x_3)$, which is very similar to that of using a gamma model.

In Table 2, we compare 95% confidence intervals for the coefficients based on the empirical likelihood ratio test of Corollary 1 to those obtained from standard Wald tests for gamma regression. Confidence intervals based on our approach are not necessarily symmetric around their point estimates, unlike

Table 2. *Worsted yarn experiment: estimated coefficients and confidence intervals*

	Gamma model (Wald-based 95% CI)	Proposed method (ELRT-based 95% CI)
x_1 (length)	0·843 (0·756, 0·929)	0·798 (0·734, 0·898)
x_2 (amplitude)	−0·631 (−0·718, −0·545)	−0·599 (−0·695, −0·528)
x_3 (load)	−0·385 (−0·472, −0·298)	−0·402 (−0·471, −0·268)

CI, confidence interval; ELRT, empirical likelihood ratio test.

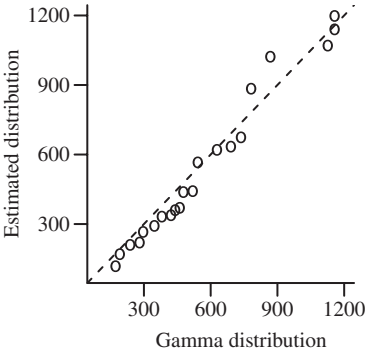

Fig. 1. Worsted yarn experiment: quantile-to-quantile plot of the maximum empirical likelihood estimate $\hat{F}$ against the gamma distribution.

Wald intervals. Also, neither method produces uniformly narrower confidence intervals. Overall, however, our proposed method provided a very similar fit to the data in terms of estimation and inferences on the mean model.

In Fig. 1, we plot the quantiles of the estimated error distribution $\hat{F}$ given by (7) against the quantiles of the gamma distribution function with scale = 31·621 and mean 861·3, the same as that of $\hat{F}$. The Kolmogorov–Smirnov statistic for comparing the two distributions is approximately 0·25 on 23 degrees of freedom, giving a scaled statistic of 1·2. Both the plot and the Kolmogorov–Smirnov statistic suggest that the gamma distribution is an acceptable approximation to the data. Our approach provides a diagnostic and justification for the gamma log-linear model fitted by Myers et al. (2010), without relying on the correctness of the gamma response model for inferences on regression coefficients.

6. DISCUSSION

Empirical likelihood approaches for generalized linear and quasilikelihood models have been investigated before in the framework of general estimating equations (e.g., Kolaczyk, 1994). These methods typically construct likelihood functions that do not correspond to any explicit probability model for the data. In contrast, the proportional likelihood ratio approach in this note assumes an explicit probability model for the data, up to an unspecified infinite-dimensional distribution parameter, which is then estimated via empirical likelihood. The theoretical and simulation results suggest that the proposed method performs quite favourably compared with existing methods. Our new class of semiparametric models can also be used for model selection and diagnostics within the classical generalized linear model framework.

ACKNOWLEDGEMENT

We thank the editor and a referee for pointing out the similarities between this paper and Luo & Tsai (2012), and for helpful suggestions that streamlined this paper. We thank Mihai Anitescu for providing help with the MINOS solver in AMPL.

Supplementary material

Supplementary material available at *Biometrika* online includes an outline of the proofs of Propositions 1 and 2.

References

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc.* B **26**, 211–52.

DiCiccio, T. J., Hall, P. & Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19**, 1053–61.

Fokianos, K., Kedem, B., Qin, J. & Short, D. A. (2001). A semiparametric approach to the one-way layout. *Technometrics* **43**, 56–65.

Fourer, R., Gay, D. M. & Kernighan, B. W. (2003) *AMPL: A Modeling Language for Mathematical Programming*, 2nd ed. Pacific Grove: Cengage Learning.

Kolaczyk, E. D. (1994). Empirical likelihood and generalized linear models. *Statist. Sinica* **4**, 199–218.

Luo, X. & Tsai, W. Y. (2012). Proportional likelihood ratio model. *Biometrika* **99**, 211–22.

Myers, R. H., Montgomery, D. C., Vining, G. G. & Robinson, T. J. (2010). *Generalized Linear Models: With Applications in Engineering and the Sciences*, 2nd ed. New York: Wiley.

Owen, A. B. (2001). *Empirical Likelihood*. Boca Raton: Chapman & Hall.

Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300–25.

Rathouz, P. J. & Gao, L. P. (2009). Generalized linear models with unspecified reference distribution. *Biostatistics* **10**, 205–18.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178–203.

[*Received January* 2011. *Revised October* 2011]