STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:2409–2419 (DOI: 10.1002/sim.1047)

A solution to the problem of separation in logistic regression

Georg Heinze*,† and Michael Schemper

Section of Clinical Biometrics, Department of Medical Computer Sciences, University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

SUMMARY

The phenomenon of separation or monotone likelihood is observed in the fitting process of a logistic model if the likelihood converges while at least one parameter estimate diverges to ± infinity. Separation primarily occurs in small samples with several unbalanced and highly predictive risk factors. A procedure by Firth originally developed to reduce the bias of maximum likelihood estimates is shown to provide an ideal solution to separation. It produces finite parameter estimates by means of penalized maximum likelihood estimation. Corresponding Wald tests and confidence intervals are available but it is shown that penalized likelihood ratio tests and profile penalized likelihood confidence intervals are often preferable. The clear advantage of the procedure over previous options of analysis is impressively demonstrated by the statistical analysis of two cancer studies. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: bias reduction; case-control studies; infinite estimates; monotone likelihood; penalized likelihood; profile likelihood

1. INTRODUCTION

In logistic regression it has been recognized that with small to medium-sized data sets situations may arise where, although the likelihood converges, at least one parameter estimate is infinite [1–7]. These situations occur if the responses and non-responses can be perfectly separated by a single risk factor or by a non-trivial linear combination of risk factors. Therefore Albert and Anderson [1] denoted such situations by ‘separation’. The phenomenon is also known as ‘monotone likelihood’ [8]. In general, one does not assume infinite parameter values in underlying populations. The problem of separation is rather one of non-existence [9] of the maximum likelihood estimate under special conditions in a sample. An infinite estimate can also be regarded as extremely inaccurate, the inaccuracy resulting in Wald confidence intervals of infinite width [3, 10, 11].

The problem of separation is by no means negligible and may occur even if the underlying model parameters are low in absolute value. In Table I we show how the probability of separation depends on sample size, on the number of dichotomous risk factors, the magnitude of the odds ratios associated with them and on the degree of balance in their distribution.

*Correspondence to: Georg Heinze, Department of Medical Computer Sciences, University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria
†E-mail: georg.heinze@akh-wien.ac.at

Received July 2000
Accepted April 2001

Table I. The probability of separation (per cent) in logistic regression with dichotomous risk factors. Each entry is based on 1000 samples. The expected marginal balance of responses and non-responses is fixed at 1:1.

Sample size	Number of risk factors	$B_X^* = 1:1$ Odds ratio				$B_X^* = 1:4$ Odds ratio			
		1	2	4	16	1	2	4	16
30	3	0	3	10	53	17	25	43	74
	5	2	7	24	75	30	41	58	85
	10	12	38	78	98	56	71	86	98
50	3	0	0	1	18	2	5	15	46
	5	0	0	2	32	6	9	22	53
	10	0	1	20	78	10	19	36	74

*Degree of balance of dichotomous risk factors.

Sophisticated but computationally demanding procedures exist to detect separation [2, 5]. However, in practice, monitoring the variance in the iteration process is sufficient, for example, by declaring separation if for the related analysis with standardized risk factors at least one of the parameters' variances exceeds 5000 (see SAS reference [12], p. 1945).

Separation was detected in the statistical analysis of an unpublished endometrial cancer study which partly motivated this work. We thank Dr E. Asseryanis from the Vienna University Medical School for providing the data set. Purpose of this study of 79 primarily diagnosed cases of endometrial cancer was to explain the histology of the endometrium by putative risk factors. Histology (HG) was classified as either 0 (grading 0–II) or 1 (grading III–IV) for 30 and 49 patients, respectively. Three risk factors were considered: neovasculization (NV) is coded as 1 (present) for 13 patients and 0 (absent) for 66 patients, and two continuous factors, pulsatility index of arteria uterina (PI) and endometrium height (EH) range from 0 to 49 and from 0.27 to 3.61, respectively, with medians of 16 and 1.64.

As there is no observation for which $\mathrm{NV} = 1$ and $\mathrm{HG} = 0$, quasicomplete separation (see Albert and Anderson [1]) occurs, leading to an infinite ML estimate of the effect of NV.

Currently, if separation caused by a risk factor (NV in our example) is detected, leading to an infinite estimate for parameter β_{NV}, one of the following options will be considered:

1. Omission of NV from the model.
2. Changing to a different type of model.
3. Use of an *ad hoc* adjustment (data manipulation).
4. Exact logistic regression [13].
5. Standard analysis with $\hat{\beta}_{\mathrm{NV}}$ set to a 'high' value (for example, the value of $\hat{\beta}_{\mathrm{NV}}$ of that iteration at which the log-likelihood changed by less than 10^{-6}).

Omission of NV (option 1) provides no information about the effect of this unusually strong and therefore important risk factor and furthermore does not allow adjusting effects of the other risk factors for the effect of NV. Therefore, this option is totally inappropriate.

Expressing effects of risk factors in terms of (log) odds ratios is not only common in the analysis of binary responses but also useful. Models whose parameters have different interpretations that are not risk-related (option 2) may be less appealing.

Ad hoc adjustments (data manipulation) preceding a standard analysis may produce finite estimates (option 3). While simple adjustments of cell frequencies can have undesirable properties (Agresti and Yang, reference [14], p. 20), Clogg *et al.* [15] pursued a more elaborate approach: let $\bar{p} = \sum_{i=1}^{n} y_i/n$ with $y_i \in \{0, 1\}$ denoting the binary dependent variable and let k symbolize the number of parameters to be estimated. Then their basic idea is to add $\bar{p}k/g$ artificial responses and $(1 - \bar{p})k/g$ artificial non-responses to each of the g groups of distinct risk factor patterns, and then to do a standard analysis on the augmented data set. Not much is known about the performance of this procedure and therefore it will be subjected to a comparative investigation in this paper.

Use of exact logistic regression (option 4) permits replacement of the unsuitable maximum likelihood estimate by a median unbiased estimate [4]: let x_{ir} denote the value of the rth risk factor for individual i $(1 \leqslant i \leqslant n;\ 2 \leqslant r \leqslant k)$ and let $x_{i1} = 1$ for all i. Then the median unbiased estimate of a parameter β_r as well as corresponding inference are based on the exact null distribution of the sufficient statistic $T_r = \sum_{i=1}^{n} y_i x_{ir}$ of β_r, conditional on the observed values of the other sufficient statistics $T_{r'}, r' \neq r$. An efficient algorithm is available to evaluate these conditional distributions [16] which should contain a sufficient number of elements. This requirement may be violated with a single continuous risk factor but also with multiple dichotomous risk factors. In the endometrial cancer study we cannot apply exact logistic regression because there are two continuous risk factors in the model leading to degenerate distributions of all sufficient statistics.

A high value for $\hat{\beta}_{\mathrm{NV}}$ (option 5) implies extreme inflation of $\mathrm{var}(\hat{\beta}_{\mathrm{NV}})$ and leads [10, 11] to an insignificant Wald test that may not be plausible for a very strong effect. Though likelihood ratio tests and one-sided profile likelihood confidence intervals can be used, the arbitrary choice for $\hat{\beta}_{\mathrm{NV}}$ remains unsatisfactory.

In this paper we review and suggest a procedure which avoids the arbitrary choice for $\hat{\beta}_{\mathrm{NV}}$ and arrives at a finite estimate for β_{NV} by a modification of the score function of logistic regression. This modification was originally developed by Firth [17–19] to reduce the bias of maximum likelihood estimates in generalized linear models. These estimates are biased away from zero and the occurrence of infinite parameter estimates in situations of separation can be interpreted as an extreme consequence of this property. Several authors [20–24] have discussed the bias of maximum likelihood estimates and have suggested corrections which, however, are only applicable to finite estimates. Our paper focuses on the use of Firth's method with logistic regression, in particular under separation. In the following section we first review some principal ideas of Firth [19], then deal with their implementation in logistic regression (FL), and, finally, suggest confidence intervals based on the profile penalized likelihood. In Section 3 the empirical performance of Firth's procedure is explored in comparison with competing methods to cope with separation. Section 4 revisits the endometrial cancer study and presents and discusses results by an FL analysis. Furthermore, we motivate application of Firth's method also by the analysis of a matched case-control study.

2. FIRTH'S MODIFIED SCORE PROCEDURE

Maximum likelihood estimates of regression parameters β_r $(r = 1, \ldots, k)$ are obtained as solutions to the score equations $\partial \log L/\partial \beta_r \equiv U(\beta_r) = 0$ where L is the likelihood function. In

order to reduce the small sample bias of these estimates Firth [19] suggested basing estimation on modified score equations

$$U(\beta_r)^* \equiv U(\beta_r) + 1/2 \operatorname{trace}[I(\beta)^{-1}\{\partial I(\beta)/\partial\beta_r\}] = 0 \quad (r = 1,\ldots,k) \tag{1}$$

where $I(\beta)^{-1}$ is the inverse of the information matrix evaluated at β. The modified score function $U(\beta)^*$ is related to the penalized log-likelihood and likelihood functions, $\log L(\beta)^* = \log L(\beta) + 1/2 \log|I(\beta)|$ and $L(\beta)^* = L(\beta)|I(\beta)|^{1/2}$, respectively. The penalty function $|I(\beta)|^{1/2}$ is known as Jeffreys invariant prior for this problem. Its influence is asymptotically negligible. By using this modification Firth [19] showed that the $O(n^{-1})$ bias of maximum likelihood estimates $\hat{\beta}$ is removed.

Alternative formulations of penalization in logistic regression have been suggested but with purposes other than bias reduction and solution to the separation problem [25, 26].

If Firth's general idea is applied to a logistic model

$$\text{Prob}(y_i = 1 \mid x_i, \beta) = \pi_i = \left\{1 + \exp\left(-\sum_{r=1}^{k} x_{ir}\beta_r\right)\right\}^{-1}$$

then the score equation $U(\beta_r) = \sum_{i=1}^{n}(y_i - \pi_i)x_{ir} = 0$ is replaced by the modified score equation

$$U(\beta_r)^* = \sum_{i=1}^{n}\{y_i - \pi_i + h_i(1/2 - \pi_i)\}x_{ir} = 0 \quad (r = 1,\ldots,k)$$

where the h_i's are the ith diagonal elements of the 'hat' matrix $H = W^{1/2}X(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W^{1/2}$, with $W = \operatorname{diag}\{\pi_i(1 - \pi_i)\}$. Now Firth-type (FL) estimates $\hat{\beta}$ can be obtained iteratively the usual way (see, for example, Collett, reference [27], Section 3.3.3) until convergence is obtained:

$$\beta^{(s+1)} = \beta^{(s)} + I^{-1}(\beta^{(s)})U(\beta^{(s)})^*$$

where the superscript (s) refers to the sth iteration.

Do finite FL parameter estimates always exist? Alternatively to the method presented above, FL estimates can be obtained by splitting each original observation i into two new observations having response values y_i and $1 - y_i$ with iteratively updated weights $1 + h_i/2$ and $h_i/2$, respectively. The contribution of the new observations to the score function is then $\{(y_i - \pi_i)(1 + h_i/2) + (1 - y_i - \pi_i)h_i/2\}x_{ir} = \{y_i - \pi_i + h_i(1/2 - \pi_i)\}x_{ir}$. We see that the splitting of each original observation into a response and a non-response guarantees finite estimates. Hence the FL method completely eliminates the problem of separation. Only those problems of estimation remain which can also occur with the general linear model, for example, problems due to multicollinearity or nearly degenerate risk factor distributions.

Estimation of standard errors can be based on the roots of the diagonal elements of $I(\hat{\beta})^{-1}$, which is a first-order approximation to $\{-\partial^2 \log L^*/(\partial\beta)^2\}^{-1}$ (see Firth, reference [19], p. 36).

As FL parameter estimates typically will be lower in absolute value than maximum likelihood (ML) estimates, their standard errors will be reduced as well.

Independent of whether $\hat{\beta}$ is obtained by ML or FL, its distribution may be distinctly non-normal and then likelihood ratio tests are preferable. In our case the likelihood ratio statistic LR is defined by $\text{LR} = 2\{\log L(\hat{\gamma}, \hat{\delta})^* - \log L(\gamma_0, \hat{\delta}_{\gamma_0})^*\}$, where $(\hat{\gamma}, \hat{\delta})$ is the joint penalized maximum likelihood estimate of $\beta = (\gamma, \delta)$, the hypothesis of $\gamma = \gamma_0$ being tested, and $\hat{\delta}_{\gamma_0}$ is

the penalized maximum likelihood estimate of δ when $\gamma = \gamma_0$. The values of the profile of the penalized log-likelihood function for γ, $\log L(\gamma, \hat{\delta}_\gamma)^*$, are obtained by fixing γ at predefined values around $\hat{\gamma}$, $\hat{\delta}_\gamma$ denoting penalized maximum likelihood estimates of δ for γ fixed at the predefined values. A profile likelihood $(1-\alpha)$ 100 per cent confidence interval for a scalar parameter γ is the continuous set of values γ_0 for which LR does not exceed the $(1-\alpha)100$th percentile of the χ^2_1-distribution. In Section 4 the profile of the penalized likelihood will be used to judge the adequacy of Wald tests.

Finally, we note that FL and the procedure by Clogg *et al.* [15] (option 3 of Section 1) formally agree for a saturated model where $g = k$, g being the number of distinct risk factor patterns, and with $\bar{p} = \sum_{i=1}^{n} y_i/n = 1/2$. In that case $h_i = 1$ $(i = 1, \ldots, g)$ because $0 \leqslant h_i \leqslant 1$ and $\sum_i h_i = k$ (see, for example, Belsley *et al.* [28], pp. 66–67), and therefore both adjustments, $h_i/2$ and $\bar{p}k/g$, assume a value of $1/2$.

3. AN EMPIRICAL STUDY

A simulation study was conducted to explore and compare the empirical performance of standard maximum likelihood fitting (ML), Firth-type fitting (FL) and fitting with augmented data according to Clogg *et al.* [15] (CL). We also include results from exact logistic regression (XL) for simulated scenarios where it might be an option.

Whereas FL and CL have been defined in previous sections, fitting by ML follows the default implementation of procedure LOGISTIC of SAS/STAT [12]. XL models were obtained by first employing the efficient algorithm described by Hirji *et al.* [29] to compute the exact null distribution of a sufficient statistic and then computing exact maximum likelihood or median unbiased estimates for non-separated and separated data sets, respectively.

The effect of the following factors on the bias of parameter estimates and on the coverage probability of one-sided lower (extending to $-\infty$) and upper (extending to $+\infty$) 97.5 per cent confidence intervals was investigated in a factorial design, generating 1000 samples for each cell: sample size n (30, 50, 100), number of binary risk factors k–1 (3, 5, 10), identical odds ratio $\exp(\beta)$ associated with each risk factor (1, 2, 4, 16) and identical degree of balance B_X of each covariate (1 : 1, 1 : 4). For each cell an intercept parameter β_1 was determined to obtain a balance of responses and non-responses B_Y of 1 : 1 and of 1 : 4.

Values of the binary risk factors x_{ir} $(1 \leqslant i \leqslant n;\ 2 \leqslant r \leqslant k)$ were sampled using the random number generator RANBIN of SAS [30] and x_{i1} was fixed to 1. Outcomes y_i were sampled from a binomial distribution with parameter $\pi_i = \{1 + \exp(-\sum_{r=1}^{k} x_{ir}\beta)\}^{-1}$ again using RANBIN.

Whereas the complete numerical results of the Monte Carlo study are contained in a technical report [31], the typical performance of the investigated procedures can already be understood by means of results selected for Table II. We learn that the bias of $\hat{\beta}_2$ with FL is small, that it is larger with both CL and XL, and that it is largest with ML. Theoretically the bias of the ML method is ∞ if separation occurs with non-zero probability.

For $n = 30$ and $k - 1 = 10$ there is a large proportion of simulated data sets (5–72 per cent) where XL parameter estimates are unavailable due to degenerate conditional distributions (in agreement with Hirji *et al.* [4]). Therefore, respective results are not contained in Table II, as well as those for $n = 100$ where memory and computing time requirements are excessive.

Table II. Average bias × 100 of parameter estimates in logistic regression. Each entry is based on 1000 samples. The expected marginal balance of responses and non-responses is fixed at 1:1.

Sample size	Number of risk factors	Method	$B_X^* = 1:1$ OR†				$B_X^* = 1:4$ OR†			
			1	2	4	16	1	2	4	16
			100β‡				100β‡			
			0	69	139	277	0	69	139	277
30	3	ML	−4	32	102	566	−7	88	186	424
		FL	−3	1	1	−6	−2	−1	−5	−19
		CL	−2	−1	−8	−48	−2	−7	−21	−63
		XL	−3	4	6	−35	−2	−2	−10	−42
	10	ML	−27	574	1118	1168	−8	326	897	1292
		FL	0	3	−23	−130	2	8	−6	−89
		CL	−2	−15	−56	−172	2	−3	−34	−140
100	3	ML	0	4	10	34	1	5	9	58
		FL	0	1	2	2	1	−2	−3	−1
		CL	0	0	0	−11	1	−6	−13	−30
	10	ML	1	11	34	429	1	15	32	233
		FL	1	0	2	8	1	3	4	5
		CL	1	−3	−19	−97	1	1	−10	−71

*Degree of balance of dichotomous risk factors.
†Odds ratio.
‡Parameter value (log-odds ratio).

The empirical coverage under FL by one-sided 97.5 per cent confidence intervals of Wald type and by those based on the profile penalized likelihood was equally satisfactory and close to nominal for odds ratios ranging from 1 to 4 and balanced risk factors. For situations where separation occurs the profile of the penalized likelihood function becomes highly unsymmetric, as will be illustrated by Figure 1 in Section 4. Therefore Wald tests and confidence intervals become unsuitable. This is reflected in one-sided coverage probabilities approaching 100 per cent in situations with a high probability of separation. In these cases coverage by profile penalized likelihood confidence intervals is much more satisfactory, usually still being close to nominal. Detailed tables of coverage rates are given in a technical report [31].

Summarizing, the study confirmed the safe use of FL in general and its clear superiority over ML particularly in situations of high parameter values and/or unbalanced risk factors. Especially in these situations inference should be based on penalized likelihood ratio tests and profile penalized likelihood ratio confidence intervals rather than on Wald type methods.

4. EXAMPLES

4.1. Endometrial cancer study

We return to the endometrial cancer study introduced in Section 1. In the analysis of this data set, quasicomplete separation led to an infinite maximum likelihood (ML) estimate of the effect

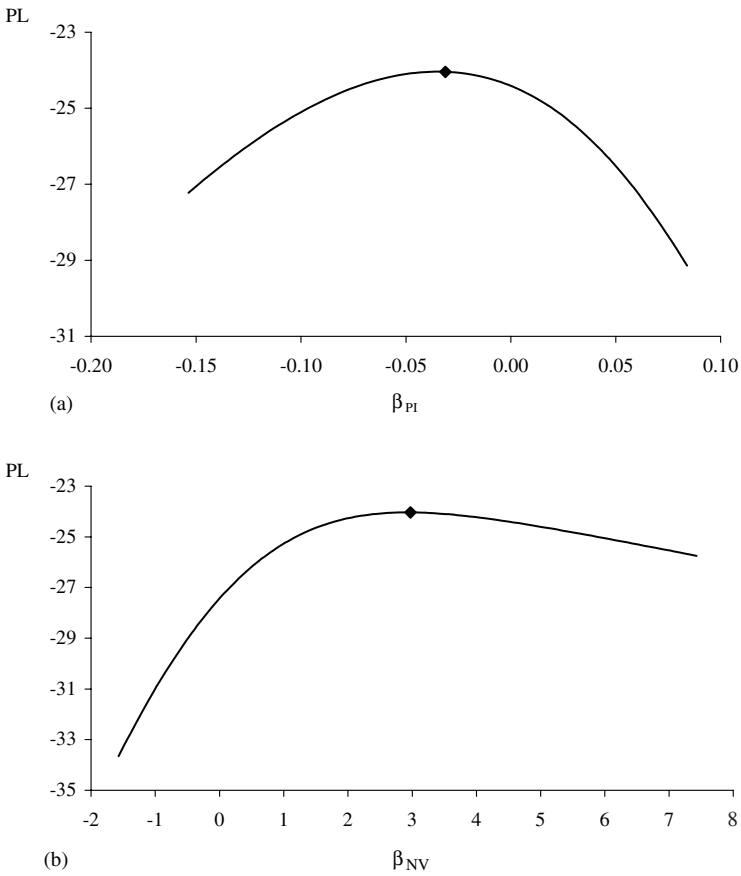

Figure 1. Profile penalized log likelihood function (PL) for factors (a) PI and (b) NV. The functions were obtained by fixing the investigated parameters, β_{PI} and β_{NV}, at 100 predefined values evenly spread within ± 3 standard errors ($\hat{\sigma}(\beta_{\mathrm{PI}}) = 0.04$, $\hat{\sigma}(\beta_{\mathrm{NV}}) = 1.55$) of the point estimates ($\hat{\beta}_{\mathrm{PI}} = -0.03$, $\hat{\beta}_{\mathrm{NV}} = 2.93$) denoted by '$\diamond$'.

of NV. Because the other two risk factors are continuous, exact logistic regression analysis is not available. As the medical investigators were interested in the effect of neovasculization, omission of this risk factor (option 1 of Section 1) cannot be considered here. Odds ratio estimates and corresponding 95 per cent confidence intervals using methods FL and ML are shown in Table III. Point estimates for PI and EH by an FL analysis are slightly smaller than by ML, as expected from the bias reducing property of FL. The estimated odds ratio of 18.7 for NV is a plausible and well communicable result.

ML Wald confidence limits provide no information about the odds ratio produced by NV and therefore are useless. Furthermore, for NV, only a one-sided profile likelihood confidence interval is available by ML. Exploration of the profile penalized log-likelihood function reveals approximately normal shapes for PI and EH but not for NV (see Figure 1). This explains why the relatively strong effect of NV is not declared significant by the FL Wald confidence

Table III. Odds ratio estimates ($\widehat{\mathrm{OR}}$) and two-sided 95 per cent confidence limits for the endometrial cancer study.

Method	Risk factor	$\widehat{\mathrm{OR}}$	95 per cent confidence limits			
			Profile likelihood		Wald	
ML[†]	NV	4.23×10^{6}	3.6	*	3.2×10^{-171}	5.5×10^{181}
	PI[‡]	0.66	0.25	1.47	0.28	1.56
	EH	0.06	0.01	0.24	0.01	0.29
FL	NV	18.71	1.84	2577.46	0.90	391.02
	PI[‡]	0.71	0.29	1.50	0.33	1.54
	EH	0.07	0.01	0.29	0.02	0.34

*Value not available
[†]ML estimation by PROC LOGISTIC of SAS/STAT [12], true $\widehat{\mathrm{OR}}$ point and Wald interval estimates for risk factor NV being $+\infty$ and $[0, +\infty]$, respectively.
[‡]Odds ratio based on changes of 10 units of PI.

interval. With increasing parameter values, distributions of parameter estimates tend to become asymmetric and then profile likelihood confidence intervals are preferable. As can be seen from Table III, two-sided 95 per cent confidence intervals according to Wald and according to the profile penalized likelihood are closest for PI, slightly different for EH, and substantially different for NV.

Thus application of FL and use of profile penalized likelihood confidence intervals have provided better information about the risk factors of this study than the standard ML procedure.

4.2. Lung cancer study

Our second example is a case-control study to assess the risk of developing lung cancer following radiation treatment for carcinoma of the breast. The data set, in which 52 controls were matched to 18 cases by year of birth, year of death, and year of diagnosis of breast cancer, is provided by Mehta and Patel [32]. Two questions are to be answered: whether smoking or radiation treatment are independently associated with lung cancer, and whether there is an interaction of smoking and radiation treatment with respect to developing lung cancer. Thus two different models are estimated: one containing the main effects 'smoking' ($1 =$ yes, $0 =$ no, balance $15:55$) and 'radiation treatment' ($1 =$ yes, $0 =$ no, balance $55:15$) only, and one containing main effects and the interaction term ($1 =$ exposed to both smoking and radiation, $0 =$ otherwise, balance $11:59$).

Several options of analysis can be compared: Though an unconditional unmatched analysis of matched case-control data ('UML') will usually not be considered because parameter estimates are biased towards zero (see Breslow and Day, reference [33], p. 271) we include results from such an analysis for comparative purposes.

Currently, conditional logistic regression for matched sets ('CML') is the standard method of analysis. For matched case-control studies of small sample size or sparse data structure use of an exact analogue is preferred over CML [34]. Results from this approach are abbreviated by 'CXL'.

Table IV. Odds ratio estimates (p-values) for the lung cancer study according to variants of logistic regression (LogR).

Factor	Method					
	UML	CML	CXL	UCML	UCFL	UCCL
Smoking	11.94 (<0.01)	20.70 (<0.01)	19.49 (<0.01)	123.97 (<0.01)	17.46 (<0.01)	15.64 (<0.01)
Radiation	1.17 (0.84)	1.20 (0.86)	1.20 (1.00)	1.28 (0.83)	1.12 (0.91)	1.15 (0.87)
Interaction	3.74 (0.38)	1.67×10^7 (0.10)	2.53 (0.57)	7.15×10^9 (0.03)	9.97 (0.19)	8.25 (0.21)

Methods: UML, unconditional unmatched LogR; CML, conditional matched LogR; CXL, exact CML; UCML, unconditional matched LogR; UCFL, Firth-type UCML; UCCL, Clogg-type UCML.
p-values: refer to likelihood ratio tests except for CXL (exact tests) and UCFL (penalized likelihood ratio tests).
Note that the odds ratio estimates for the interaction given by CML and UCML were obtained by SAS/PROC LOGISTIC [12], the true estimates being ∞.

In unconditional matched analyses of matched case-control studies ('UCML') estimated parameters tend to be biased away from zero (see Breslow and Day, reference [33], Section 7.1). This property is due to the required additional estimation of a large number of nuisance parameters (one for each matched set) which with standard likelihood methods can result in severe bias (Cox and Hinkley, reference [35], p. 292). Firth's modified score procedure substantially alleviates the unsuitable performance of UCML. Results from this approach are abbreviated by 'UCFL'. Finally, we supply results of an unconditional matched analysis by Clogg's procedure ('UCCL'), which by the empirical results of the previous section would be rated between UCML and UCFL. Results obtained by the various methods are summarized in Table IV.

For the analysis of main effects, the CXL, CML, UCFL and UCCL methods arrive at very similar results. Compared to these, UML and UCML produce odds ratio estimates of the smoking effect that are, as expected for matched data, closer to and further away from one, respectively. All methods agree on an almost non-existing effect of radiation. Results are more heterogenous for the analysis of the interaction: in each matched set where at least one control was exposed to both smoking and radiation, also the case was exposed to both risk factors. Therefore, the corresponding CML and UCML parameter estimates are infinite and SAS/PROC LOGISTIC [12] stopped iterations at the implausible parameter values (log-odds ratios) of 16.63 and 22.69, respectively. The significance of the UCML estimate on the 5 per cent level parallels the overestimation of effects by UCML. Surprisingly, the CXL odds ratio estimate is even lower than the UML estimate of 3.74, which itself is assumed to be biased towards one. The median unbiased estimate employed by CXL is based on a conditional distribution of the sufficient statistic of the interaction parameter that consists of two feasible values only. If the other of the two values had been observed, then the CML odds ratio estimate was 0, but the CXL estimate was still 2.53. Therefore we do not consider median unbiased estimates suitable for such situations. However, UCFL provides a plausible estimate

of the interaction odds ratio (9.97), assuming that the lower UML estimate is biased towards one. UCCL results come very close to those by UCFL.

The similarity of the CXL and UCFL main effects estimates and the availability of a plausible estimate of the interaction effect confirm Firth's modified score procedure to be a suitable means of analysis for the lung cancer study.

5. DISCUSSION

Though David Firth's modification of the score function was presented seven years ago, its substantial practical relevance may not have been fully recognized. We have shown that separation is a non-negligible problem for logistic regression and that the modification originally derived to reduce the bias of maximum likelihood estimates provides an ideal solution to this problem. This has been confirmed by an extensive empirical study as well as by the analyses of two different clinical data sets for which the quality of presented results had been substantially improved.

Furthermore, by means of the second example we could show that application of Firth's procedure now also permits the unconditional analysis of matched case-control studies.

Finally, we have demonstrated that in particular for large parameter estimates, where Firth's modification appears most useful, inference based on the profile penalized likelihood is preferable to Wald-type tests and confidence intervals.

The procedure suggested in this paper permits Monte Carlo studies of small sample settings which previously resulted in frequent cases of separation and thus led to uninterpretable results. Similarly, when the bootstrap was applied, separation occurred in some resamples even when no separation was observed in the analysis of the original sample. Thus also the applicability of the bootstrap to logistic regression with small samples is increased.

Application of the Firth-type estimation and inference procedures is facilitated by a program, *FL*, available on request.

ACKNOWLEDGEMENTS

The authors are grateful to David Firth and an anonymous referee for suggestions that considerably improved this paper.

REFERENCES

1. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1984; **71**:1–10.
2. Santner TJ, Duffy DE. A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1986; **73**:755–758.
3. Lesaffre E, Albert A. Partial separation in logistic discrimination. *Journal of the Royal Statistical Society, Series B* 1989; **51**:109–116.
4. Hirji KF, Tsiatis AA, Mehta CR. Median unbiased estimation for binary data. *American Statistician* 1989; **43**:7–11.
5. Clarkson DB, Jennrich RI. Computing extended maximum likelihood estimates for linear parameter models. *Journal of the Royal Statistical Society, Series B* 1991; **53**:417–426.
6. Lesaffre E, Marx BD. Collinearity in generalized linear regression. *Communications in Statistics—Theory and Methods* 1993; **22**:1933–1952.
7. Kolassa JE. Infinite parameter estimates in logistic regression, with application to approximate conditional inference. *Scandinavian Journal of Statistics* 1997; **24**:523–530.

8. Bryson MC, Johnson ME. The incidence of monotone likelihood in the Cox model. *Technometrics* 1981; **23**:381–383.
9. Jacobsen M. Existence and unicity of MLEs in discrete exponential family distributions. *Scandinavian Journal of Statistics* 1989; **16**:335–349.
10. Hauck WW, Donner A. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association* 1977; **72**:851–853.
11. Væth M. On the use of Wald's test in exponential families. *International Statistical Review* 1985; **53**:199–214.
12. SAS Institute Inc. *SAS/STAT User's Guide, Version 8*. SAS Institute Inc.: Cary, NC, 1999.
13. Mehta CR, Patel NR. Exact logistic regression: theory and examples. *Statistics in Medicine* 1995; **14**:2143–2160.
14. Agresti A, Yang M-C. An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics & Data Analysis* 1987; **5**:9–21.
15. Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association* 1991; **86**:68–78.
16. Cytel Software Corporation. *LogXact 2.0*. Cytel Software Corporation: Cambridge, MA, 1996.
17. Firth D. Bias reduction, the Jeffreys prior and GLIM. In *Advances in GLIM and Statistical Modelling*, Fahrmeir L, Francis B, Gilchrist R, Tutz G (eds). Springer-Verlag: New York, 1992; 91–100.
18. Firth D. Generalized linear models and Jeffreys priors: an iterative weighted least-squares approach. In *Computational Statistics, Vol. 1*, Dodge Y, Whittaker J (eds). Physica-Verlag: Heidelberg, 1992; 553–557.
19. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**:27–38.
20. Schaefer RL. Bias correction in maximum likelihood logistic regression. *Statistics in Medicine* 1983; **2**:71–78.
21. Cordeiro GM, McCullagh P. Bias correction in generalized linear models. *Journal of the Royal Statistical Society, Series B* 1991; **53**:629–643.
22. Bull SB, Greenwood CMT, Hauck WW. Jackknife bias reduction for polychotomous logistic regression. *Statistics in Medicine* 1997; **16**:545–560.
23. Cordeiro GM, Cribari-Neto F. On bias reduction in exponential and non-exponential family regression models. *Communications in Statistics—Simulation and Computation* 1998; **27**:485–500.
24. Leung DH-Y, Wang YG. Bias reduction using stochastic approximation. *Australian & New Zealand Journal of Statistics* 1998; **40**:43–52.
25. Anderson JA, Blair V. Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika* 1982; **69**:123–136.
26. le Cessie S, van Houwelingen JC. Ridge estimators in logistic regression. *Applied Statistics* 1992; **41**:191–201.
27. Collett D. *Modelling Survival Data in Medical Research*. Chapman and Hall: London, 1994.
28. Belsley DA, Kuh E, Welsch RW. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley: New York, 1980.
29. Hirji KF, Mehta CR, Patel NR. Computing distributions for exact logistic regression. *Journal of the American Statistical Association* 1987; **82**:1110–1117.
30. SAS Institute Inc. *SAS Language Reference, Version 8*. SAS Institute Inc.: Cary, NC, 1999.
31. Heinze G. Technical Report 10: The application of Firth's procedure to Cox and logistic regression. Department of Medical Computer Sciences, Section of Clinical Biometrics, Vienna University, Vienna, 1999.
32. Mehta CR, Patel NR. *LogXact-Turbo User Manual*. Cytel Software Corporation: Cambridge, MA, 1993.
33. Breslow NE, Day NE. *Statistical Methods in Cancer Research. Volume 1—The Analysis of Case-control Studies*. IARC Scientific Publications: Lyon, 1980.
34. Hirji KF, Mehta CR, Patel NR. Exact inference for matched case-control studies. *Biometrics* 1988; **44**:803–814.
35. Cox DR, Hinkley DV. *Theoretical Statistics*. Chapman and Hall: London, 1974.

