Eviter les estimations infinies avec la régression logistique – théorie, solutions, exemples

Georg Heinze

Avoiding infinite estimates in logistic regression – theory, solutions, examples

Georg Heinze

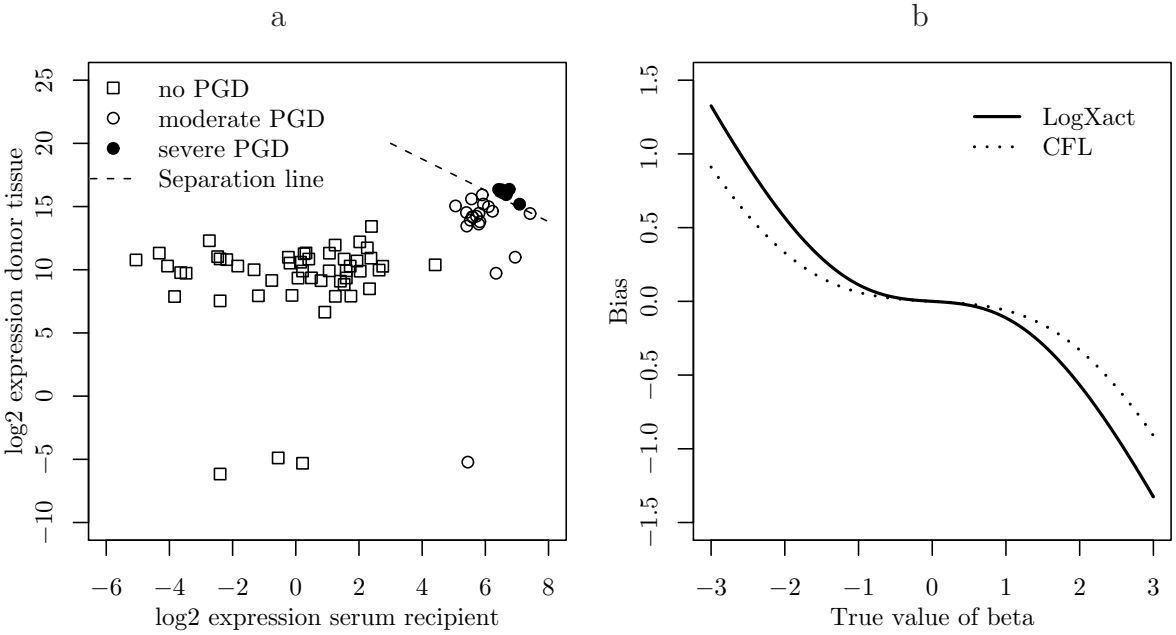
Core Unit for Medical Statistics and Informatics, Medical University of Vienna
Spitalgasse 23, A-1090 Vienna, Austria
e-mail: georg.heinze@meduniwien.ac.at

In logistic regression analyses of small or sparse data sets, results obtained by maximum likelihood methods cannot be generally trusted. In such analyses, although the likelihood meets the convergence criterion, at least one parameter may diverge to plus or minus infinity. This situation has been termed 'separation'. Examples of two studies are given, where the phenomenon of separation occurred: the first one investigated whether primary graft dysfunction of lung transplants is associated with endothelin-1 mRNA expression measured in lung donors and in graft recipients. In the second example, conditional logistic regression was used to analyze a randomized animal experiment in which animals were clustered into sets defined by equal follow-up time. I show that a penalized likelihood approach provides an ideal solution to both examples, and provide comparative analyses including possible alternative approaches. The estimates obtained by the penalized likelihood approach have reduced bias compared to their maximum likelihood counterparts, and inference using penalized profile likelihood is straightforward. Finally, I provide an overview of software that can be used to apply the proposed penalized likelihood approach.

Eviter les estimations infinies avec la regression logistique - theorie, solutions, exemples

Dans les analyses d'échantillons de petite taille ou avec données manquantes, les résultats de la régression logistique obtenus par la méthode du maximum de vraisemblance (MV) ne sont généralement pas fiables. Dans de telles analyses, bien que le critère de convergence puisse être atteint, au moins un paramètre peut diverger vers plus ou moins l'infini. Cette situation a été appelée 'séparation'. Nous donnons des exemples de deux études où ce phénomène s'est produit. Le premier exemple porte sur l'association entre le rejet précoce d'une greffe de poumon et l'expression des ARNm de l'endothéline-1 mesurée chez le donneur et le receveur de greffe. Dans le second exemple, la régression logistique conditionnelle a été utilisée pour analyser un essai randomisé en grappe sur des animaux. Nous montrons que l'approche par la vraisemblance pénalisée (VP) constitue une solution adéquate pour les deux exemples, et nous présentons une analyse comparative avec des approches alternatives. Des résultats de simulations suggèrent que la méthode de la VP conduit des estimations quasi non biaisées, même lorsque la probabilité d'estimations infinies par la MV est non négligeable. Les taux de couverture des intervalles de confiance basés sur la VP sont proches du taux nominal, et les tests sont plus puissants que ceux des autres approches. Enfin, nous donnons un aperçu des logiciels qui peuvent être utilisés pour appliquer l'approche de la VP proposée.

Figure 1: a: ET-1 expression in donor tissue vs. recipient serum and PGD. b: Exact bias evaluation for animal experiment.



1 Theory

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. These situations occur if the responses and non-responses can be perfectly separated by a single independent variable or, as seen in Fig. 1a, by a non-trivial linear combination of independent variables. Therefore, Albert and Anderson (1984) denoted such situations by 'separation'.

The data of Fig. 1 stems from a study recently performed at the Medical University of Vienna, which aimed at investigating the association of endothelin-1 (ET-1) mRNA expression in pulmonary tissue of lung donors and in the serum of lung recipients with short-term outcome of the transplantation. The data set, which consists of 76 lung donors and recipients, was provided by Dr. Mohamed Salama to whom I am indebted. Within the first three days after transplantation, presence of severe primary graft dysfunction (PGD) was assessed according to the standards set by the International Society for Heart and Lung Transplantation. Six out of 76 patients were diagnosed with severe PGD, which is associated with poor long-term outcome. Figure 1a shows the log transformed ET-1 expression values and the PGD grading of the 76 study subjects.

Logistic regression analysis using the standard maximum likelihood estimation tech-

nique does not converge, such that the parameter estimates of both covariates are ∞, with infinite variance.

In general, one does not assume infinite parameter values in underlying populations. The problem of separation is rather one of non-existence of the maximum likelihood estimate under special conditions in a sample. In the following, I show how a method originally developed by Firth (1993) to reduce the bias of maximum likelihood estimates may be used to obtain more plausible parameter estimates. 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, e. g. Cordeiro and McCullagh (1991) or Bull *et al* (1997), have discussed the bias of maximum likelihood estimates and have suggested corrections which, however, are only applicable to finite estimates. This paper focuses on the use of Firth's method with logistic regression, in particular under separation.

2 Solution

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 (1993) suggested to base estimation on modified score equations

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

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 influence of the penalty function $|I(\beta)|^{1/2}$ is asymptotically negligible. By using this modification Firth (1993) showed that the $O(n^{-1})$ bias of maximum likelihood estimates $\hat{\beta}$ is removed. Heinze and Schemper (2002) further showed that Firth-type estimates in logistic regression are always finite. These authors also proposed to use the profile penalized likelihood for the construction of confidence intervals and tests. In our case the likelihood ratio statistic LR is defined by $LR = 2\left\{\log L(\hat{\gamma}, \hat{\delta})^* - \log L(\gamma_0, \hat{\delta}_{\gamma_0})^*\right\}$, 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 $(1-\alpha)100\%$ 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. Simulation studies of small data sets revealed that Firth-type estimates are often nearly unbiased and confi-

Table 1: Comparison of estimation methods in two example studies

Study	Method	Variable	Odds ratio	95% CI	P-value
Lung	ML	Donor tissue	∞	[0, ∞]	1.000
transplants		Recipient serum	∞	[0, ∞]	1.000
	Firth	Donor tissue	7.2	[1.05, 73]	0.039
		Recipient serum	0.6	[0.22, >1000]	0.289
Animal	CML	Heparin (no vs. yes)	∞	[0, ∞]	1.000
experiment	LogXact	Heparin (no vs. yes)	6.18	[0.71, ∞]	0.103
	CFL	Heparin (no vs. yes)	11.04	[1.07, 1491]	0.042

dence intervals based on profile penalized likelihood yield coverage rates that are close to their nominal values (Heinze, 2006; Heinze and Schemper, 2002).

3 Examples

An analysis of the lung transplant data of Section 1 using the Firth-type method proposed above yields finite and plausible odds ratio estimates and 95% confidence intervals for the covariates (Tab. 1). Please note that since covariates are $\log_2$ ET-1 mRNA expression values, the odds ratio estimates refer to a doubling of mRNA expression. We may conclude that an effect of ET-1 mRNA expression on severe PGD can only be confirmed for expression in donor tissue.

My second example is an animal experiment (Bergmeister *et al*, 2008) which was provided to me by Dr. Helga Bergmeister from the Medical University of Vienna, Divison of Biomedical Research. Purpose of the study was to investigate heparin-crosslinked and non-heparinized, xenogeneic vascular substitutes in a rat model. In 38 of the 76 study objects, implants were heparin-crosslinked. Prostheses were implanted into the abdominal aorta of the rats, which were divided into strata followed up for one day, three days, seven days, ten days, one month, three months, and six months. Each stratum comprised the same number of rats from the heparin-crosslinked and the non-heparinized groups.

Unlike in the lung transplant study, now the study subjects are clustered into 7 groups defined by equal follow-up time. Conditional logistic regression (Breslow and Day, 1980) eliminates the 7 cluster-specific parameters from the likelihood by conditioning on their sufficient statistics, which are given by the number of cases in each cluster. In total, only four aneurysms were documented, and they all occurred in rats provided with non-heparinized implants. Therefore, the conditional maximum likelihood estimate of the group effect is not finite and one would conclude infinite higher risk for aneurisms in the non-heparinized group. Table 1 contains the odds ratio estimates and 95 per cent

confidence intervals comparing the risk of aneurysms between non-heparinized ($x = 1$) *versus* heparin-crosslinked implants ($x = 0$). Odds ratios were estimated by standard conditional logistic regression (CML), by transferring the Firth-type penalization to the conditional likelihood (CFL), and by the software package LogXact (Cytel, 2005). The latter approach generates a permutational distribution of the sufficient statistic of the group parameter $T = \sum_i x_i y_i$, and estimates β by maximizing the probability of the observed sufficient statistic. In case of separation, LogXact uses a median unbiased estimate, which is obtained by choosing $\hat{\beta}$ such that $\Pr(T|\beta) = 1/2$.

Both CFL and median unbiased estimates yield finite odds ratio estimates, but CFL shrinks the estimate less than does the median unbiased approach (Table 1). Exact conditional analysis yields the following distribution of T under the null hypothesis that the odds ratio is one (or $\beta = 0$): $T = 0, 1, 2, 3, 4$ with $\Pr(T|\beta = 0) = 0.052, 0.248, 0.400, 0.248, 0.052$, respectively. Given this exact conditional distribution, one can compare the exact bias of CFL and LogXact estimation for various assumptions about β (Fig. 1b). We learn that both estimators of β show a bias towards zero, but the bias of CFL is less severe than that of LogXact. Comparing the confidence intervals (Table 1), we notice that the profile penalized likelihood interval excludes the value of 1, while from the exact and normal approximation intervals one would not conclude a significant group difference. Since the point probabilities of the two most extreme values of T (0 and 4) under $\beta = 0$ both are 0.052, the exact test of the hypothesis $\beta = 0$ will not reject the null hypothesis, even if one of these extremes is observed. Hence, the corresponding exact 95 per cent confidence interval cannot exclude an odds ratio of 1 under any assumption about β. We conclude that in this example, the profile penalized likelihood confidence interval provides a good trade-off between coverage and power.

The Firth-type penalization method for small sample unconditional or conditional logistic regression analysis presented in this contribution improves on existing approaches in various aspects. First, it provides a convenient solution to the problem of separation and prevents the analyst from reporting unplausible infinite (or zero) odds ratio estimates. Second, it considerably reduces the bias of point estimates. Third, it provides powerful inference and avoids the conservatism often encountered in exact conditional analysis.

With increasing sample size, the relative impact of penalization on the estimates will vanish, and differences between Firth-type and maximum likelihood analyses will become negligible. Thus, applying the Firth-type penalization particularly takes effect in small samples or in sparse data situations, as encountered in our examples. Furthermore, the Firth-type penalization allows the inclusion of continuous covariates in the model, which often leads to degenerate distributions of sufficient statistics in exact conditional analysis – the reason why LogXact could not be applied in the analysis of the lung transplant study. With maximum likelihood estimation, inference based on normal approximation using an infinite standard error could be replaced by inference using the profile likelihood. However, in sparse data situations as encountered in our examples, profile likelihood intervals must be assumed anticonservative and should not be trusted (Heinze, 2006).

SAS, SPLUS and R programs (Heinze and Ploner, 2003) for unconditional and conditional Firth-type analyses, comprising point and confidence interval estimation are available at: `http://www.muw.ac.at/msi/biometrie/programme/fl`. For unconditional logistic regression, the Firth-type method including profile penalized likelihood confidence intervals as suggested by Heinze and Schemper (2002) has also been implemented in the latest versions of LogXact (Cytel, 2005) and SAS (2008), in Ioannis Kosmidis' R package brglm (`http://cran.r-project.org/web/packages/brglm`), and in Joseph Coveney's Stata module FIRTHLOGIT (`http://ideas.repec.org/c/boc/bocode/s456948.html`).

The author is grateful to Karen Leffondré, Bordeaux, for help with translation of the abstract.

References

[1] Albert, A. and Anderson, J. A. (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71, 1–10.

[2] Bergmeister, H., Plasenzotti, R., Walter, I., Plass, C., Bastian, F., Rieder, E., Sipos, W., Kaider, A., Losert, U. and Weigel, G (2008) Decellularized, xenogeneic small-diameter arteries: transition from a muscular to an elastic phenotype in vivo. *J Biomed Mater Res B Appl Biomater*, 87, 95–104.

[3] Breslow, N. E. and Day, S. (1980) *Statistical methods in cancer research. Volume 1 - The analysis of case-control studies*. IARC Scientific Publications, Lyon.

[4] Bull, S. B., Greenwood, C. M. T. and Hauck, W. W. (1997) Jackknife bias reduction for polychotomous logistic regression. *Statistics in Medicine*, 16, 545–560.

[5] Cordeiro, G. M. and McCullagh, P. (1991) Bias correction in generalized linear models. *Journal of the Royal Statistical Society B*, 53, 629–643.

[6] Cytel Software Corporation (2005) *LogXact 7 manual*. Cytel, Cambridge, MA.

[7] Firth D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27–38.

[8] Heinze G. (2006) A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, 25, 4216–4226.

[9] Heinze, G. and Ploner, M. (2003) Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine*, 71, 181–187.

[10] Heinze, G. and Schemper, M. (2002) A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21, 2409–2419.

[11] SAS Institute Inc. (2008) *SAS/STAT 9.2 User's Guide*. SAS Institute Inc., Cary, NC.