Journal of Statistical Software

April 2012, Volume 47, Code Snippet 2. *http://www.jstatsoft.org/*

Separation-Resistant and Bias-Reduced Logistic Regression: STATISTICA Macro

Kamil Fijorek
Cracow University of Economics

Andrzej Sokołowski
Cracow University of Economics

Abstract

Logistic regression is one of the most popular techniques used to describe the relationship between a binary dependent variable and a set of independent variables. However, the application of logistic regression to small data sets is often hindered by the complete or quasicomplete separation. Under the separation scenario, results obtained via maximum likelihood should not be trusted, since at least one parameter estimate diverges to infinity. Firth's approach to logistic regression is a theoretically sound procedure, which is guaranteed to arrive at finite estimates even in a separation case. Firth's procedure was also proved to significantly reduce the small sample bias of maximum likelihood estimates. The main goal of the paper is to introduce the STATISTICA macro, which performs Firth-type logistic regression.

Keywords: logistic regression, complete separation, STATISTICA.

1. Introduction

1.1. Logistic regression model

Logistic regression is a commonly used tool to describe the relationship between a binary outcome variable and a set of explanatory variables. It is routinely employed in many fields, e.g., medicine, social sciences, economics. The popularity of logistic regression stems mainly from its mathematical convenience and the relative ease of interpretation in terms of odds ratios (Hosmer and Lemeshow 2000; Long 1997; Greene 2003).

Assume that the dependent variable $y_i \in \{0, 1\}$ is a Bernoulli distributed variable with success probability $F(x_i^\top \theta)$, where $F(\circ)$ is the logistic distribution function, x_i is a p-dimensional vector of explanatory variables and $\theta \in \mathbb{R}^p$ is a p-dimensional parameter vector ($i = 1, \ldots, n$). The most frequently used estimation method of θ is the maximum likelihood estimation

(MLE). The MLE principle states that the estimate of θ is the value which maximizes the likelihood function (Hosmer and Lemeshow 2000). The likelihood function and its logarithm in case of logistic regression are given by the following equations:

$$L(\theta) = \prod_{i=1}^{n} F(x_i^\top \theta)^{y_i} [1 - F(x_i^\top \theta)]^{1-y_i}, \tag{1}$$

$$l(\theta) = \sum_{i=1}^{n} y_i \ln F(x_i^\top \theta) + (1 - y_i) \ln[1 - F(x_i^\top \theta)]. \tag{2}$$

In order to find the value of θ that maximizes $L(\theta)$, partial derivatives of a log-likelihood function with respect to θ are calculated:

$$U(\theta) = \sum_{i=1}^{n} [y_i - F(x_i^\top \theta)] x_i. \tag{3}$$

The solution to score equations $U(\theta) = 0$ gives the ML estimate of θ, i.e., $\hat{\theta}$.

In most cases, there is no analytical solution to score equations. Consequently, numerical methods are used to find $\hat{\theta}$. With starting value $\theta^{(1)}$, the maximum likelihood estimate $\hat{\theta}$ is obtained iteratively:

$$\theta^{(r+1)} = \theta^{(r)} + I^{-1}_{\theta^{(r)}} U\left(\theta^{(r)}\right), \tag{4}$$

where the superscript (r) refers to the r-th iteration and I_θ denotes the Fisher information matrix evaluated at θ (Greene 2003):

$$I_\theta = -\sum_{i=1}^{n} \frac{\partial^2 l_i(\theta)}{\partial\theta\partial\theta^\top} = \sum_{i=1}^{n} F(x_i^\top \theta)\left(1 - F(x_i^\top \theta)\right) x_i x_i^\top. \tag{5}$$

The desirable properties of ML estimates such as: consistency, efficiency and normality, are based on the assumption that the sample size (n) approaches infinity. However in many real life situations the large sample assumption is not satisfied, and as a result, ML estimates should not be trusted.

The bias of ML estimates in small samples can be substantial. Moreover, in small samples, there is a non-negligible probability of encountering the separation. From the geometrical point of view the separation occurs when there exists a hyperplane which separates successes and failures (complete separation), where the hyperplane itself may contain both successes and failures (quasicomplete separation). In that case, at least one parameter estimate diverges to infinity (Albert and Anderson 1984; Heinze and Schemper 2002). In practice, the separation phenomenon can be detected by tracking the magnitude of standard errors. The most common strategy to deal with separation is to remove any offending variable(s) from the model. However, this approach is seriously flawed since the omission of any important variable(s) is inappropriate.

1.2. Penalized maximum likelihood

Firth (1992a,b, 1993) derived the procedure that guarantees finite estimates of logistic regression parameters in case of separation; it was also proven to significantly reduce the small

sample bias of maximum likelihood estimates, i.e., the first-order term is removed from the asymptotic bias of maximum likelihood estimates. The procedure originally developed by Firth (1993) was further researched and popularized by the work of Heinze and Schemper (2002), Heinze and Ploner (2004), and Heinze (2006).

The basis of Firth's approach is the idea that the bias in $\hat{\theta}$ can be reduced by modifying the score equations. The modified equation has the following form:

$$U^*(\theta) = \sum_{i=1}^{n} \left\{ y_i - F(x_i^\top \theta) + h_i \left[\frac{1}{2} - F(x_i^\top \theta) \right] \right\} x_i, \tag{6}$$

where h_i is the i-th diagonal element of the H matrix $H = W^{\frac{1}{2}} X (X^\top W X)^{-1} X^\top W^{\frac{1}{2}}$ is a $n \times p$ data matrix and W is a $n \times n$ diagonal matrix with the i-th diagonal element $F(x_i^\top \theta) \left[1 - F(x_i^\top \theta)\right]$.

The modification to score equations can alternatively be introduced by penalizing the original likelihood function:

$$L^*(\theta) = L(\theta) |I_\theta|^{\frac{1}{2}}. \tag{7}$$

It is interesting that Firth's approach to logistic regression is identical to Bayesian logistic regression with noninformative Jeffreys prior.

Penalized maximum likelihood estimates (PMLE) can be found with the use of the numerical routine described above with $U\left(\theta^{(r)}\right)$ term replaced by $U^*\left(\theta^{(r)}\right)$.

1.3. Statistical inference

Estimation of standard errors can be based on the roots of the diagonal elements of $I_{\hat{\theta}}^{-1}$, which is a first-order approximation to $I_\theta^{*-1} = \left(-\frac{\partial^2 l^*(\theta)}{\partial\theta\partial\theta^\top}\right)^{-1}$ (Firth 1992a,b, 1993; Bull, Mak, and Greenwood 2002). According to the simulation study performed by the authors (results not shown), there is no clear advantage of using I_θ^{*-1} in place of I_θ^{-1} with respect to the number of iterations needed for convergence and the coverage of a Wald confidence interval. However, more extensive study regarding this subject would be useful. Appropriate simulation studies can be greatly facilitated by the recent work of (Chen, Ibrahim, and Kim 2008) from which the closed form of I_θ^{*-1} can be easily obtained.

Given the estimate of the covariance matrix $I_{\hat{\theta}}^{-1}$, one can compute Wald confidence intervals and p values based on the normal approximation to the distribution of the PML estimates. The $(1-\alpha)\%$ Wald confidence interval for θ_j, $(j = 1, \ldots, p)$ is given by:

$$\left(\hat{\theta}_j - z_{1-\frac{\alpha}{2}} \sqrt{(I_{\hat{\theta}}^{-1})_j};\quad \hat{\theta}_j + z_{1-\frac{\alpha}{2}} \sqrt{(I_{\hat{\theta}}^{-1})_j} \right), \tag{8}$$

where $\hat{\theta}_j$ is the PML estimate of j-th element of $\hat{\theta}$, $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution function and $(I_{\hat{\theta}}^{-1})_j$ is a j-th diagonal element of $I_{\hat{\theta}}^{-1}$.

It should be noted, however, that in small samples the coverage probability of the Wald confidence interval may deviate from its nominal value. This behavior was observed in simulation studies performed by (Heinze 1999), it was also shown there that profile likelihood confidence intervals are superior to Wald intervals. Simulation studies performed by the authors (results not shown) led to similar conclusions.

2. STATISTICA macro

The STATISTICA data analysis software system (StatSoft, Inc. 2010) offers a user-friendly module for maximum likelihood estimation of logistic regression coefficients. As was mentioned above, the maximum likelihood estimation is not resistant to the separation problem and, generally speaking, is not suitable for small datasets. This was the main motivation for the implementation of a STATISTICA macro performing Firth-type logistic regression.

The presented macro was written in SVB (the Visual Basic programming environment integrated with STATISTICA) in STATISTICA 9.1. SVB is similar to Microsoft Visual Basic 6, as well as the Visual Basic language available in Microsoft **Excel**. SVB includes a comprehensive library of optimized matrix procedures. The employment of built-in matrix functions resulted in a more readable macro code.

The macro code was partly based on a source code of the R (R Development Core Team 2012) package **logistf** (Heinze and Ploner 2004; Ploner, Dunkler, Southworth, and Heinze 2010). An alternative R implementation offering similar functionality is available in the package **brglm** (Kosmidis 2011), based on the recent work of Kosmidis and Firth (2009).

2.1. Starting a macro

- Launch STATISTICA and open a dataset of interest.
- Open the macro file named `SR_BR_LR.svb` (separation-resistant bias-reduced logistic regression), available along with this manuscript.
- Press the `F5` keyboard button or left-click the `Run Macro` button on the `Macro` toolbar (see Figure 1).
- The variable selection window should appear. The list of variables available for analysis is automatically loaded from the active workbook.
- Select appropriate variables and press the OK button. After a short time, an output window should appear.

If one plans to use the macro frequently, then a different approach should be taken, i.e., the macro should be installed. This process creates a button which automatically starts the macro without the need to manually reopen the macro file every time STATISTICA is restarted. To install the macro file:

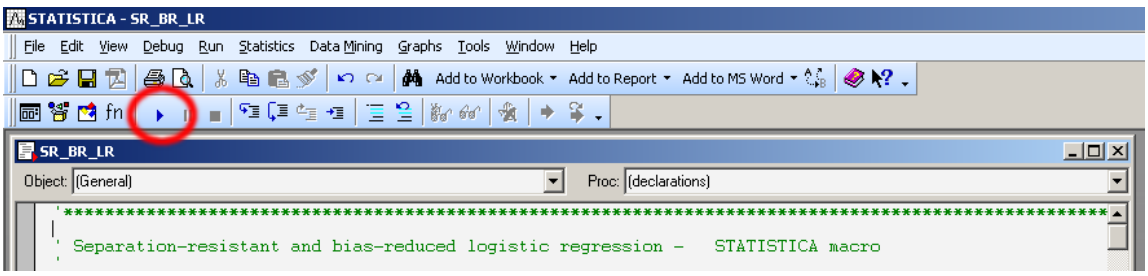

Figure 1: The localization of the `Run Macro` button on the `Macro` toolbar.

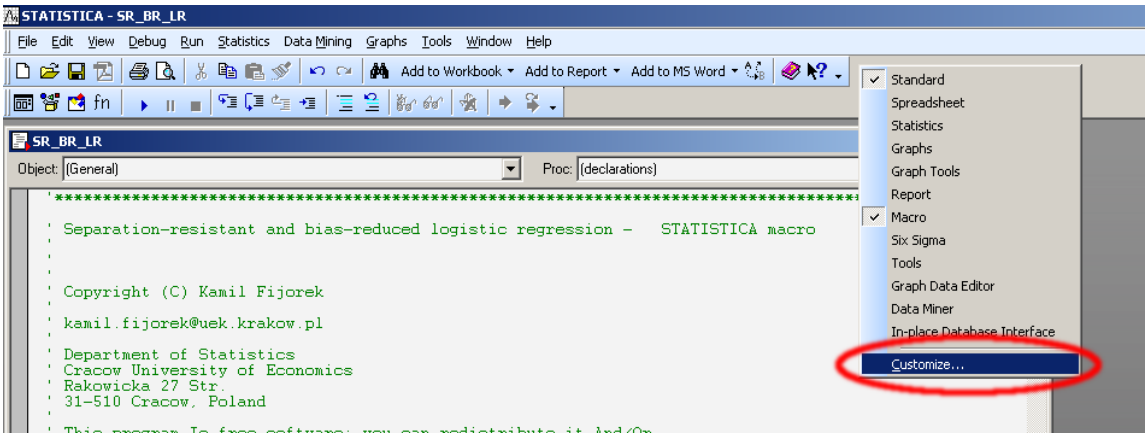

Figure 2: The localization of the `Customize` button.

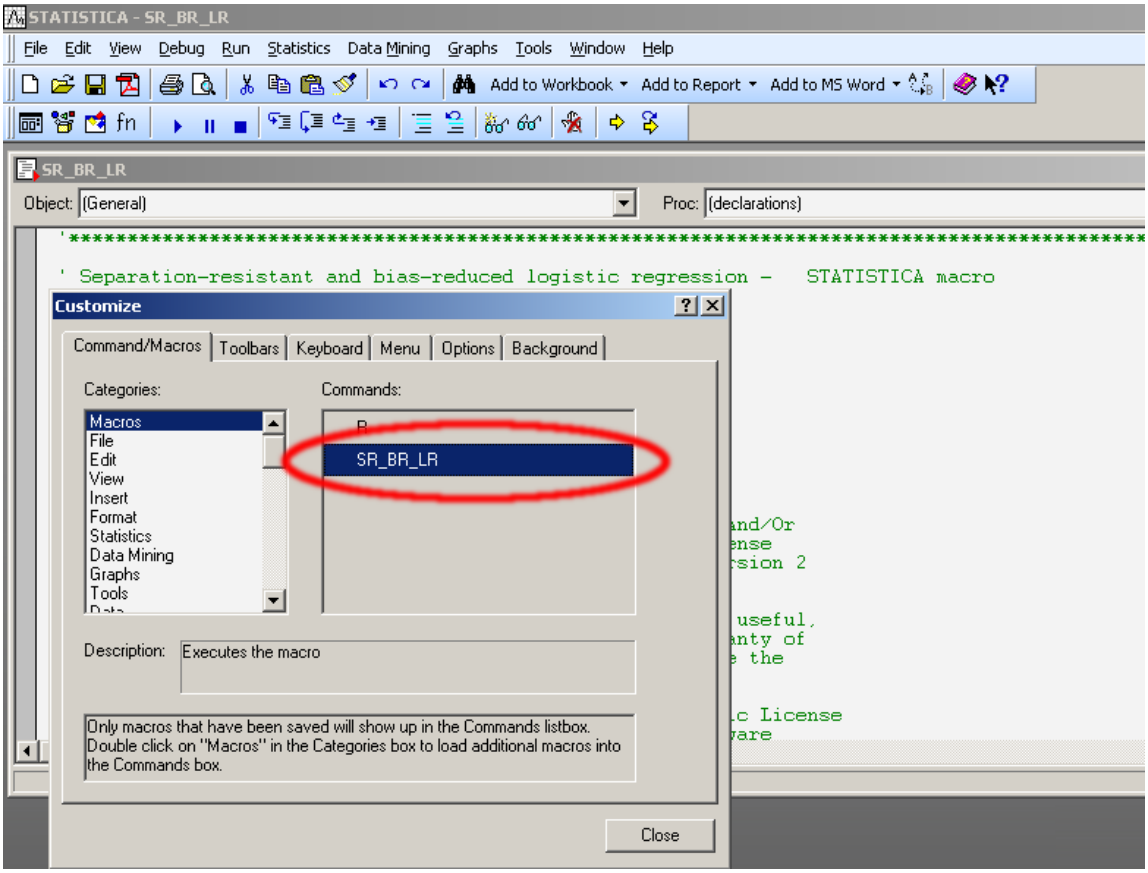

Figure 3: The localization of the `Commands` list.

Figure 4: The effect of the macro installation.

- Launch STATISTICA.
- Open the macro file named `SR_BR_LR.svb`.
- Right-click on any toolbar and select the `Customize` (see Figure 2).
- Go to the `Command/Macros` tab, select `Macros` from the `Categories` list, and from the `Commands` list, drag the macro name (i.e., `SR_BR_LR`) from its original position to the main toolbar (see Figure 3 and 4).
- Press the newly created button to start the macro.

2.2. Generated output

The output generated by the macro consists of a workbook with three spreadsheets. The first spreadsheet holds the original values of the dependent variable along with estimated probabilities of a success. The second spreadsheet holds the estimate of the covariance matrix, i.e., $I^{-1}_{\hat{\theta}}$. The third spreadsheet is of the most importance as it shows the following:

- The number of iterations needed for convergence.
- The value of log-likelihood at last iteration.
- Parameter estimates.
- Standard errors (SE) of parameter estimates.
- 95% Wald confidence intervals for parameters.
- Odds ratios for a unit increase in the independent variables.
- 95% Wald confidence intervals for odds ratios.
- P values for a hypothesis test that a given parameter is equal to zero.

In the current version of the macro, profile penalized likelihood confidence intervals are not implemented. They are, however, planned for future release.

3. Examples

3.1. Toy example

The toy dataset was constructed to demonstrate the case of the infiniteness of ML estimates in logistic regression. The data matrix X in the first column contains 1s and the only explanatory variable assumes 10 equidistant values from 1 to 10. The first 5 values of dependent variable (y) are failures (0) and the last 5 are successes (1). One can clearly see the complete separation,

	Convergence after 14 iterations. Log-likelihood: -1,08069806320775							
	1 Estimate	2 SE	3 95% CI - lower limit	4 95% CI - upper limit	**5 OR**	6 OR - 95% CI - lower limit	7 OR - 95% CI - upper limit	8 p
Intercept	-5,3385695	3,3227101	-11,8510814	1,1739423	-	-	-	0,1081221
X	0,9706490	0,5765404	-0,1593703	2,1006683	2,6396571	0,8526806	8,1716290	0,0922639

Figure 5: STATISTICA macro output – toy dataset.

i.e., all cases are failures where the value of the explanatory variable is less than 6 and all cases are successes where the value of the explanatory variable is greater than or equal to 6.

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{pmatrix}^\top$$
$$y = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}^\top$$

To estimate the logistic regression model on this dataset, we used both the package **logistf** and our macro. Below we show the necessary commands to create the toy dataset, estimate the model and visualize the results. Figure 6 presents the dataset along with the fitted logistic function. Figure 5 shows a screen capture of the output of our macro. The estimation results of both the package **logistf** and our macro are in very close agreement.

```
R> library("logistf")
R> X <- 1:10
R> Y <- rep(0:1, each = 5)
R> mod <- logistf(Y ~ X, firth = TRUE, pl = FALSE)
R> mod

logistf(formula = Y ~ X, pl = FALSE, firth = TRUE)
Model fitted by Penalized ML
Confidence intervals and p-values by Wald

               coef se(coef) lower 0.95 upper 0.95      z        p
(Intercept) -5.33857  3.32271  -11.85096     1.1738 2.5815 0.108122
X            0.97065  0.57654   -0.15935     2.1006 2.8344 0.092264

Likelihood ratio test=7.7588 on 1 df, p=0.0053453, n=10

R> mod$loglik[2]

[1] -1.0807

R> C1 <- coef(mod)
R> par(mar = c(2.5, 2.5, 1.5, 1.5))
R> plot(X, Y, pch = 19, cex = 1.5)
R> grid()
R> sek <- seq(0, 10, by = 0.01)
R> pr_1 <- 1 / (1 + exp(-cbind(1, sek) %*% C1))
R> lines(sek, pr_1, lwd = 2)
```

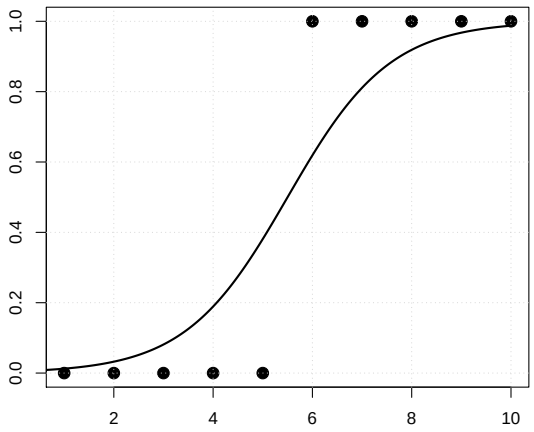

Figure 6: Logistic function fitted to the toy dataset.

3.2. Swiss banknotes example

For the next example we used the 'Swiss banknotes' dataset (Flury 1988). The dataset contains 7 variables, of which the first 6 variables are explanatory variables and the last variable is the outcome variable which describes whether the banknote is genuine or not. The dataset was taken from the R package **ncomplete** (Christmann and Rousseeuw 2001). The explanatory variables were standardized before applying estimation procedures.

The 'Swiss banknotes' dataset also exhibits complete separation. Consequently, penalized maximum likelihood estimates should be preferred. As before, estimation results of both the package **logistf** and our macro (see Figure 7) are in very close agreement.

```
R> data("Banknotes", package = "ncomplete")
R> bank <- cbind(scale(Banknotes[, 1:6]), Banknotes[, 7])
R> mod <- logistf(bank[, 7] ~ bank[, 1:6], firth = TRUE,
+    pl = FALSE, control = logistf.control(maxit = 100))
R> mod

logistf(formula = bank[, 7] ~ bank[, 1:6], pl = FALSE,
  control = logistf.control(maxit = 100), firth = TRUE)
Model fitted by Penalized ML
Confidence intervals and p-values by Wald

                  coef se(coef) lower 0.95 upper 0.95          z         p
(Intercept)  -0.221121  0.70080   -1.59466    1.15242  0.0995570 0.7523619
bank[, 1:6]1 -0.036786  0.53481   -1.08499    1.01142  0.0047312 0.9451615
bank[, 1:6]2 -0.264789  1.02000   -2.26395    1.73437  0.0673906 0.7951747
bank[, 1:6]3  0.661888  0.93900   -1.17852    2.50229  0.4968644 0.4808811
bank[, 1:6]4  2.747925  0.85890    1.06452    4.43133 10.2359251 0.0013773
bank[, 1:6]5  1.885411  0.81721    0.28371    3.48711  5.3228970 0.0210470
bank[, 1:6]6 -1.817096  0.71229   -3.21316   -0.42104  6.5079479 0.0107393

Likelihood ratio test=248.58 on 6 df, p=0, n=200
```

	Convergence after 51 iterations. Log-likelihood: -1,96038645591982							
	1 Estimate	2 SE	3 95% CI - lower limit	4 95% CI - upper limit	5 OR	6 OR - 95% CI - lower limit	7 OR - 95% CI - upper limit	8 p
Intercept	-0,22111737	0,70079674	-1,59467898	1,15244424	-	-	-	0,75236472
B_1	-0,03677491	0,53480855	-1,08499967	1,01144985	0,96389307	0,33790190	2,74958461	0,94517846
B_2	-0,26481983	1,01999595	-2,26401190	1,73437224	0,76734419	0,10393268	5,66537020	0,79515050
B_3	0,66191249	0,93900259	-1,17853258	2,50235757	1,93849615	0,30772998	12,21124889	0,48086613
B_4	2,74793173	0,85889907	1,06448956	4,43137391	15,61031221	2,89935866	84,04681021	0,00137731
B_5	1,88542048	0,81720817	0,28369247	3,48714849	6,58912446	1,32802446	32,69259143	0,02104655
B_6	-1,81708753	0,71228520	-3,21316651	-0,42100854	0,16249833	0,04022903	0,65638450	0,01073936

Figure 7: STATISTICA macro output – 'Swiss banknotes' dataset.

```
R> mod$loglik[2]

[1] -1.9604
```

Acknowledgments

The authors would like to thank the anonymous referees for their helpful comments.

References

Albert A, Anderson JA (1984). "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika*, **71**(1), 1–10.

Bull SB, Mak C, Greenwood CMT (2002). "A Modified Score Function Estimator for Multinomial Logistic Regression in Small Samples." *Computational Statistics & Data Analysis*, **39**, 57–74.

Chen MH, Ibrahim JG, Kim S (2008). "Properties and Implementation of Jeffreys's Prior in Binomial Regression Models." *Journal of the American Statistical Association*, **103**(484), 1659–1664.

Christmann A, Rousseeuw PJ (2001). "Measuring Overlap in Logistic Regression." *Computational Statistics & Data Analysis*, **37**, 65–75.

Firth D (1992a). "Bias Reduction, the Jeffreys Prior and GLIM." In R Gilchrist, G Tutz (eds.), *Advances in GLIM and Statistical Modelling*, pp. 91–100. Springer-Verlag.

Firth D (1992b). "Generalized Linear Models and Jeffreys Priors: An Iterative Weighted Leastsquares Approach." In Y Dodge, J Whittaker (eds.), *Computational Statistics*, volume 1, pp. 553–557. Physica-Verlag, Heidelberg.

Firth D (1993). "Bias Reduction of Maximum Likelihood Estimates." *Biometrika*, **80**(1), 27–38.

Flury B (1988). *Multivariate Statistics: A Practical Approach*. Chapman & Hall, London.

Greene WH (2003). *Econometric Analysis*. 5th edition. Prentice Hall, Upper Saddle River.

Heinze G (1999). "The Application of Firth's Procedure to Cox and Logistic Regression." *Technical Report 10*, Section of Clinical Biometrics, Department of Medical Computer Sciences, University of Vienna, Vienna, Austria.

Heinze G (2006). "A Comparative Investigation of Methods for Logistic Regression with Separated or Nearly Separated Data." *Statistics in Medicine*, **25**(24), 4216–4226.

Heinze G, Ploner M (2004). "A SAS Macro, S-PLUS Library and R Package to Perform Logistic Regression without Convergence Problems." *Technical report*, Section of Clinical Biometrics, Department of Medical Computer Sciences, Medical University of Vienna, Vienna, Austria.

Heinze G, Schemper M (2002). "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine*, **21**(16), 2409–2419.

Hosmer DW, Lemeshow S (2000). *Applied Logistic Regression*. John Wiley & Sons.

Kosmidis I (2011). "**brglm**: Bias Reduction in Binary-Response GLMs." R package version 0.5-6, URL http://CRAN.R-project.org/package=brglm.

Kosmidis I, Firth D (2009). "Bias Reduction in Exponential Family Nonlinear Models." *Biometrika*, **96**(4), 793–804.

Long JS (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, Thousand Oaks.

Ploner M, Dunkler D, Southworth H, Heinze G (2010). "**logistf**: Firth's Bias Reduced Logistic Regression." R package version 1.10, URL http://CRAN.R-project.org/package=logistf.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

StatSoft, Inc (2010). "STATISTICA Version 9.1." URL http://www.statsoft.com/.

A. Implementation details

At the variable selection step, only one dependent variable can be selected and at least one independent variable must be selected. The macro code performs a suite of data checks before attempting the estimation step. These data checks are as follows:

- No missing values are allowed.
- The dependent variable must be 0/1 coded.
- No constant variables are allowed.
- Independent variables are screened for extreme collinearity (the absolute value of Pearson's correlation coefficient between two variables greater than 0.99).

In case of any data check violation, the warning message is generated and the execution of the macro is terminated. After a successful data check, independent variables are internally standardized and the estimation procedure is started.

The convergence of the likelihood maximization procedure is declared if all of the following conditions are met:

- The value of the log-likelihood between two consecutive iterations is smaller than $1e-5$.
- The sum of absolute values of $U^*\left(\theta^{(r)}\right)$ is smaller than $1e-5$.
- The sum of absolute changes of $\theta^{(r)}$ between two consecutive iterations is smaller than $1e-5$.

In cases where the maximum number of iterations (500) is reached, no convergence is declared.

If the maximization routine overshoots the optimum (which can be detected be the decrease in the log-likelihood), step-halving is used (no more than 10 half-steps are calculated). Additionally the maximum step size is limited to 0.5. The vector of starting values $\theta^{(1)}$ is a vector of zeros.

All of the above-mentioned control parameters can be easily modified by the user within the macro source code.

Affiliation:

Kamil Fijorek
Department of Statistics
Cracow University of Economics
Rakowicka 27 Str.
31-510 Cracow, Poland
E-mail: kamil.fijorek@uek.krakow.pl

Andrzej Sokołowski
Department of Statistics
Cracow University of Economics
Rakowicka 27 Str.
31-510 Cracow, Poland
E-mail: andrzej.sokolowski@uek.krakow.pl

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 47, Code Snippet 2 *Submitted:* 2011-03-06
April 2012 *Accepted:* 2012-01-25