User's guide to the SAS macro FL

The following pages have been extracted from

Heinze G and Ploner M (2004). *Technical Report 2/2004: A SAS macro, S-PLUS library and R package to perform logistic regression without convergence problems.* Section of Clinical Biometrics, Department of Medical Computer Sciences, Medical University of Vienna, Vienna.

available at http://www.akh-wien.ac.at/imc/biometrie/programme/fl/tr2_2004.pdf
and http://www.meduniwien.ac.at/user/georg.heinze/techreps/tr2_2004.pdf

4 The SAS macro fl

4.1 Introduction

The SAS macro `%fl` was written to facilitate application of Firth's modified score procedure (Firth, 1993; Heinze & Schemper, 2002) in logistic regression analysis. The macro was written in 1999 and has now been subject to a complete revision. The new version 2004.07 outperforms any old version (Heinze, 1999; Heinze & Ploner, 2003) in terms of speed (by a factor of at least 3) by a more efficient use of matrix programming in PROC IML, and in terms of computational control by some new options. The macro is available at the websites `http://www.akh-wien.ac.at/imc/biometrie/programme/fl` and `http://www.meduniwien.ac.at/user/georg.heinze/programs/logistf`.

Given the X-matrix is of full rank (which is normally the case), Firth's procedure (FL) removes the problem of quasicomplete or complete separation in logistic regression analysis. Supplied with a SAS data set as input, the macro output contains FL-type logistic regression parameters, standard errors, confidence limits, p-values, the value of the maximized penalized log likelihood, and the number of iterations needed to arrive at the maximum. All output produced is stored in a SAS data set. Furthermore, parameter estimates, penalized log likelihood and covariance matrix are saved in a SAS data set that has the same structure as the output data set that can be obtained by SAS/PROC LOGISTIC (SAS Institute, 1999) using the OUTEST option in the PROC LOGISTIC statement of that procedure. The profile of the penalized log likelihood function for any parameter can be plotted to judge the adequacy of Wald confidence intervals. Efficient processing of multiple input data sets is possible through a BY variable, similarly to PROC LOGISTIC. Finally, the option of supplying offset values to the program enables computation of penalized likelihood ratio tests.

4.2 Working with the macro

In the following list of options available in `%fl`, *(new)* indicates options new in version 2004.07:

`%fl(<data=`*SAS data set*`,>`
`<y=`*variable*`,>`
`varlist=`*variables*`,`
`<noint=`*value*`,>`

<odds=*value*,>
<test=*variables*,>
<pl=*value*,>
<plint=*value*,> *(new)*
<print=*value*,>
<notes=*value*,> *(new)*
<profile=*variables*,
<profsel=*value*,>
<profser=*value*,>
<profn=*value*,>
<outprof=*SAS data set*,>
<outest=*SAS data set*,>
<outtab=*SAS data set*,>
<global=*SAS data set*,> *(new)*
<out=*SAS data set*,>
<pred=*variable*,>
<lower=*variable*,>
<upper=*variable*,>
<h=*variable*,>
<maxit=*value*,>
<maxhs=*value*,>
<epsilon=*value*,>
<maxstep=*value*,>
<standard=*value*,> *(new)*
<by=*variables*,>
<offset=*SAS data set*>);

These options can be categorized into basic options, output options, model fitting options, and options useful for simulation.

4.2.1 Basic options

- data=*SAS data set* names the input SAS data set. The default value is _LAST_.
- y=*variable* names the dependent variable containing values 1 and 0 only. The default value is Y.

- `varlist=`*variables* names a list of independent variables, separated by blanks. There is no default value. This option is required.

- `noint=`*value* suppresses estimation of an intercept if set to 1. The default value is 0.

- `odds=`*value* requests estimates and confidence limits for odds ratios if set to 1. The default value is 0.

- `test=`*variables* requests a likelihood ratio test for the null hypothesis that all parameters corresponding to variables specified in this option are zero.

- `pl=`*value* specifies the method of computing confidence limits and tests for parameters. If `pl`=1, then profile penalized likelihood confidence limits for parameters and penalized likelihood ratio tests will be computed. Otherwise, confidence limits and tests will be based on estimated variance (Wald method). Default value is 1.

- `plint=`*value* specifies the method of computing confidence limits and tests for the intercept. This option is only active if `pl`=1. If `plint=0`, which is the default value, confidence limits for the intercept will be base on the estimated variance (Wald method) instead of profile penalized likelihood. The value 1 requests computation of profile penalized likelihood confidence interval for the intercept. This option has been added as we frequently noticed numerical problems when estimating the profile likelihood confidence limits for the intercept parameter.

- `profile=`*variables* requests a plot of the profile penalized log likelihood function for all variables specified in this options. Of course, these variables have to appear also in the `varlist` option. The x-axis ranges for this plot are automatically chosen by the macro, but can also be specified by the user in terms of standard errors to the left and right from the point estimate (options `profsel` and `profser`, respectively). Also the number of profile likelihood evaluations (`profn`, default value=100) can be chosen. If the `profile` options is used, then the data set specified in option `outprof` will consist of the variables

  - `_name_` containing the covariate's name
  - `_b_` containing the values on the x-axis (the values for β)
  - `_profli_` containing the values of the profile penalized log likelihood

- `_normal_` containing the values $\ell_{\max} - 0.5(\beta - \hat{\beta})^2/\hat{\sigma}^2$ (where $\ell_{\max}$ is the maximized penalized log likelihood) which represent the Wald (normal) approximation to the profile penalized log likelihood
- `_refer_` containing the reference line (the values of β where the profile penalized log likelihood function and normal approximation crosses the reference line are the profile penalized likelihood and Wald confidence limits, respectively).
- any BY-variables specified

4.2.2 Model fitting options

- `maxit=`*value* specifies the maximum number of iterations. Default value is 50.
- `maxhs=`*value* specifies the maximum number of step–halvings allowed in one iteration. Default value is 5.
- `epsilon=`*value* specifies the maximum allowed sum of absolute changes in parameter values to declare convergence. Default value is 0.0001.
- `maxstep=`*value* specifies the maximum change of parameter values allowed in one iteration. Default value is 5.
- `standard=`*value* specifies if the data should be standardized prior to (profile) likelihood maximization. Standardization usually improves the numerical stability of likelihood maximization algorithms and may reduce the number of iterations. Default value is 1, requesting standardization. The only impact of standardization that can be seen by the user is that the value of the penalized likelihood is different from, yet proportional to, the value obtained by estimation from non-standardized data.

4.2.3 Output options

- `print=`*value* suppresses printed output if set to 0. Default value is 1.
- `outest=`*SAS data set* names a SAS data set containing parameter estimates, penalized log likelihood and covariance matrix. There is no default value. The data set contains a variable for the intercept parameter and one variable for each explanatory variable in the `varlist` option. The `outest` data set contains one

observation for each BY–group containing the FL–type estimates of the regression coefficients. Additionally, there are observations containing the rows of the estimated covariance matrix of the parameter estimators for each `by` group. The `outest` data set contains the following variables:

- any BY variables specified
- `_CODE_`, a variable containing the value -1 indicating a line with parameter estimates or the subsequent numbers of the covariates indicating lines containing corresponding rows of the estimated covariance matrix.
- if `noint`=0, the variable `INTERCPT`
- one variable for each explanatory variable in the `varlist` statement.
- `_LINK_`, a character variable of length 8 with the value LOGIT
- `_PENLIK_`, the maximized penalized log likelihood at the FL estimate (where `_CODE_`=-1) or, if `pl`=1, the maximized penalized log likelihood with the restriction that the corresponding parameter is 0 (where `_CODE_`> 0)
- `_LNLIKE_`, the log likelihood at the FL estimate (where `_CODE_`=-1) or, if `pl`=1, the log likelihood at the FL estimate maximizing the penalized likelihood with the restriction that the corresponding parameter is 0 (where `_CODE_`> 0)
- `_IT_`, the number of iterations needed to arrive at the maximum of the penalized likelihood
- `_RESP_`, the number of responses
- `_NORESP_`, the number of nonresponses
- `_TYPE_`, a character variable of length 8 with two possible values: PARMS for parameter estimates or COV for covariance estimates
- `_NAME_`, a character variable of length 8 containing the name ESTIMATE for parameter estimates or the name of each explanatory variable or INTERCPT for the covariance estimates

- `outtab=`*SAS data set* names a SAS data set containing parameter estimates, standard errors, Wald confidence limits and p–values. The default value is `_OUTTAB`. The data set contains one observation per explanatory variable or intercept parameter and BY–group. It contains the following variables:

  - any BY variables specified
  - `_RBY_`, the ascending rank of the corresponding BY-group
  - `_VAR_`, the subsequent number for each explanatory variable in the `varlist` option or 0 for the intercept parameter
  - `_NAME_`, a character variable of length 8 containing the name of each explanatory variable or INTERCPT
  - `BETA`, the FL parameter estimates
  - `STDERR`, the estimated standard error of the corresponding parameter estimate
  - `CI_LO`, the lower confidence limit for the parameter estimate
  - `CI_UP`, the upper confidence limit for the parameter estimate
  - `P_VALUE`, the p–value for $H_0 : \beta_r = 0$.
  - `_ITER_`, the number of iterations that the model fitting algorithm needed to arrive at the maximum of the penalized log likelihood.

- `global=`*SAS data set* names a SAS data set containing global likelihood ratio and Wald tests of the hypothesis that $beta = 0$. The default value is `_globtest`. The data set contains two observation per BY–group. It contains the following variables:
  - any BY variables specified
  - `_RBY_`, the ascending rank of the corresponding BY-group
  - `testtype`, a numerical variable indicating likelihood ratio test (1) or Wald test (2)
  - `type`, a character variable of length 16 containing the name of the test (‘”Likelihood ratio”’ or ‘”Wald”’)
  - `ChiSq`, the value of the χ^2-statistic
  - `df`, the number of degrees of freedom (the number of independent variables in the model)
  - `P_value`, the p–value

  If the options `odds` is set to 1, then this data set will also contain the following variables:
  - `ODDS`, the estimated odds ratio

  - `OR_LO`, the lower confidence limit for the odds ratio
  - `OR_UP`, the upper confidence limit for the odds ratio

- `out=`*SAS data set* creates a new SAS data set that contains all the variables in the input data set and the predicted probability of an event reponse, the lower and upper confidence intervals for the predicted probability, and the diagonal element of the hat matrix.

  - `pred=`*variable* assigns a name for the variable in the `out` data set containing the predicted probabilities of an event response. Default value is `_PRED_`.
  - `lower=`*variable* assigns a name for the variable in the `out` data set containing the lower confidence limit of the predicted probability. Default value is `_LOWER_`.
  - `upper=`*variable* assigns a name for the variable in the `out` data set containing the upper confidence limit of the predicted probability. Default value is `_UPPER_`.
  - `h=`*variable* assigns a name for the variable in the `out` data set containing the diagonal element of the hat matrix. Default value is `_H_`.

- `notes=`*value* If set to 1, requests notes in the SASLOG about the number of iterations needed for computation of profile likelihood confidence intervals and the profile of the penalized likelihood. Default value is 1.

4.2.4 Options useful for simulation

- `by=`*variables* requests separate analyses on observations in groups defined by the BY variable(s).

- `offset=`*SAS data set* names an input data set containing offset values of parameter estimates. This data set should contain the same variables as are specified in `varlist`, plus a variable named `INTERCEP` (unless `noint=1`) and, if the `by`–option is used, the variable(s) specified in this option. Therefore the `offset` data set should have as many observations as there are BY–groups in the input data set. If a particular parameter in a particular BY–group should be estimated, then its value should be missing in the `offset` data set, otherwise the parameter will be treated as fixed at the value found in the `offset` data set. If a variable contained in `varlist`

ist not defined in the `offset` data set, its parameter value will be estimated in any BY–group.

4.2.5 Titles

Titles 1–3 are not used by the macro. These titles can be set by the user in a statement before the macro call. Titles 4 and 5 are used by the macro. These titles are deleted on exit.

4.3 Printed output

If `print=0`, `%fl` does not produce any printed output. Otherwise, printed output includes the following:

- an initial page describing the macro and the names of the data set, the dependent and independent variables and any output data sets.
- if the `by` option is not used, a page containing the number of iterations needed to arrive at the maximum of the penalized log likelihood, the value of the penalized log likelihood, and the numbers of events, non-events and total observations
- for each BY–group a table containing the global likelihood ratio and Wald tests
- for each BY–group a table containing, for each parameter, the FL estimate, its estimated standard error, lower and upper confidence limits, and p–value
- if `odds=1`, for each BY–group a table containing, for each parameter, the FL odds ratio estimate ($\exp(\beta_r)$), lower and upper confidence limits, and p–value
- if the `test` option is used, an additional page containing the penalized log likelihood χ^2, associated degrees of freedom and p-value for the test that all parameters corresponding to variables in the `test` option are zero

4.4 Examples

We exemplify use of the `%fl` SAS macro program by means of the analysis of an endometrial cancer study and thank Dr. E. Asseryanis from the Vienna General Hospital for providing the data set. Purpose of this study of 79 patients was to explain the state of the endometrium by the putative risk factors (=covariates) neovasculization (NV),

pulsatility index of arteria uterina (*PI*), and endometrium height (*EH*). The state of the endometrium was histologically graded (*HG*) and classified as 0 (=0-II) and 1 (=III-IV) for 30 and 49 patients, respectively. *NV* is coded as 1 (present) for 13 patients and 0 (absent) for 66 patients. The two continuous covariates *PI* and *EH* range from 0 to 49 and from 0.27 to 3.61, with medians of 16 and 1.64, respectively. Suppose the data has been stored in a SAS data set `bsp.endo`. To obtain an FL analysis with a table containing variable names, parameter estimates, standard errors, profile penalized likelihood confidence limits and p–values based on penalized likelihood ratio tests, the macro call

```
title "Analysis of endometrian cancer study";
%fl(data=bsp.endo, y=hg, varlist=nv pi eh);
```

will produce the following four output pages:

```
                Analysis of endometrian cancer study

  FFFFF L               Logistic regression
  F     L               with Firth's bias reduction:
  FFF   L
  F     L               A solution to the problem of separation
  F     LLLLL           in logistic regression

  Author:               Georg Heinze
  Version:              2004.07

  Methods published in: Heinze, G. & Schemper, M. (2002). A solution
                        to the problem of separation in logistic
                        regression. Statistics in Medicine 21(16)
                        2409-2419.

  Data set:             BSP.ENDO
  Dependent variable:   HG
  Independent variables: NV PI EH

  Table with parameter estimates saved as _OUTTAB.
  Estimates and covariance matrix saved as _OUTEST.
```

Page 2:

Analysis of endometrian cancer study

Model fitting information

NOTE: Penalization computed from standardized data.

Iterations	Penalized log likelihood	Number of responses	Number of nonresponses	Number of observations
8	-24.9225	30	49	79

Page 3:

Analysis of endometrian cancer study

Testing global null hypothesis beta=0

Type	ChiSq	df	P_value
Likelihood Ratio	43.6558	3	<.0001
Wald	17.4797	3	0.0006

Page 4:

Analysis of endometrian cancer study

FL estimates, profile penalized likelihood confidence limits
and penalized likelihood ratio tests
NOTE: Confidence interval for Intercept based on Wald method.

Variable	Parameter estimate	Standard Error	Lower 95% c.l.	Upper 95% c.l.	Pr > Chi-Square
INTERCEP	3.77456	1.48869	0.85672	6.69239	0.0042
NV	2.92927	1.55076	0.60977	7.85456	0.0091
PI	-0.03475	0.03958	-0.12446	0.04046	0.3875
EH	-2.60416	0.77602	-4.36518	-1.23272	<.0001

To obtain estimated odds ratios and corresponding confidence intervals, the macro is called using the `odds` option

```
%fl(data=bsp.endo, y=y, varlist=nv eh pi, odds=1);
```

leading to the following output:

Page 1–4 omitted

Analysis of endometrian cancer study

FL odds ratio estimates, profile penalized likelihood confidence limits and penalized likelihood ratio tests

Variable	Odds ratio	Lower 95% c.l.	Upper 95% c.l.	Pr > Chi-Square
NV	18.7140	1.84000	2577.46	0.0091
PI	0.9658	0.88298	1.04	0.3875
EH	0.0740	0.01271	0.29	<.0001

The estimated probabilities of a high histological grading (III-IV) can be obtained by using the `out` option:

```
%fl(data=bsp.endo, y=hg, varlist=nv eh pi, out=prob, pred=prob,
    lower=p_lo, upper=p_up);
proc print data=prob;
var nv eh pi hg prob p_lo p_up;
run;
```

The output data set 'PROB' contains predicted probabilities as well as confidence limits:

Output omitted

OBS	NV	EH	PI	HG	PROB	P_LO	P_UP
1	0	1.64	13	0	0.27928	0.15998	0.44085
2	0	1.50	28	0	0.24885	0.10130	0.49335
3	0	2.02	29	0	0.07630	0.01839	0.26702
4	0	2.26	16	0	0.06496	0.01860	0.20297
5	0	1.33	11	0	0.48220	0.28478	0.68534
6	0	2.29	15	0	0.06237	0.01727	0.20115
7	0	3.14	8	0	0.00919	0.00074	0.10362
8	0	2.37	19	0	0.04489	0.01041	0.17354
9	0	2.33	12	0	0.06238	0.01637	0.21010
10	0	2.68	34	0	0.01230	0.00101	0.13286

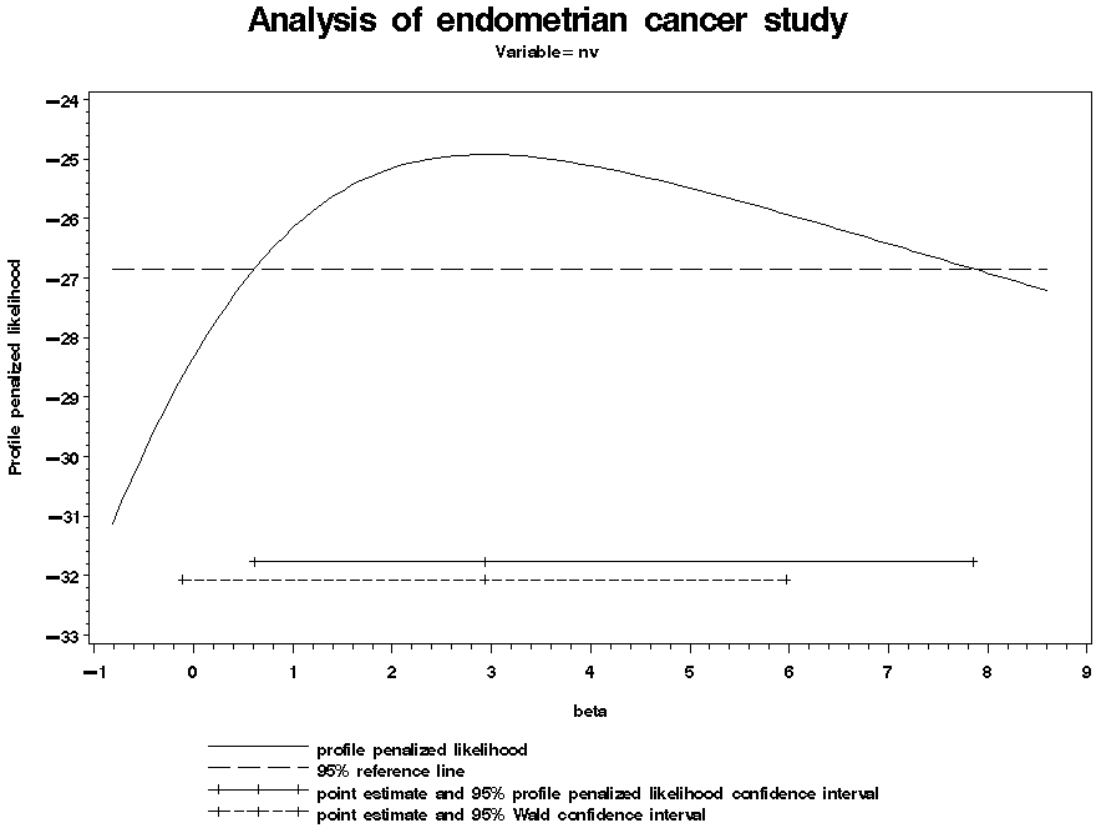
Output omitted

The `profile` option can be used to obtain a plot of the profile of the penalized log likelihood for a parameter.

```
%fl(data=bsp.endo, y=hg, varlist=nv eh pi, profile=nv);
```

The values of the penalized log likelihood are plotted against the parameter values of NV by the macro (see Fig. 2).

Figure 2: Profile penalized likelihood function for parameter β_{NV} of the endometrian cancer study.



The `test` option can be used to test the simultaneous effect of more than one independent variables on the outcome variable. In our example, to test the hypothesis $H_0 : \beta_{EH} = \beta_{PI} = 0$ the macro call

```
%fl(data=bsp.endo, y=hg, varlist=nv pi eh,
 test=eh pi);
```

is submitted. On the fifth (or sixth) page of output, the following table is displayed:

Tested variable (s)	Penalized log likelihood Chi-square	Degrees of freedom	Pr > Chi-square
EH PI	17.8667	2	0.0001

4.5 Troubleshooting

Some numerical problems may arise in computation of

- profile penalized likelihood (PPL) confidence intervals, particularly for the intercept parameter. If errors occur, we recommend the following procedure:
  1. Try `pl=0` to see if the problem is in the computation of the confidence intervals. If not, there may be multicollinearity in the design matrix.
  2. Try a smaller value for `maxstep`, 1 or 0.5, say.
  3. Is standardization requested (`standard=1`)?
  4. Try `plint=0`, which prevents computation of a PPL confidence interval for the intercept.
- profile of the penalized likelihood (option `profile`). If errors occur, we suggest to
  1. check if standardization is requested (`standard=1`)
  2. try a smaller value for `maxstep`, 1 or 0.5, say.

References

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

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

Firth, D. (1992a). Bias reduction, the Jeffreys prior and GLIM. In Fahrmeir, L., Francis, B., Gilchrist, R., & Tutz, G., editors, *Advances in GLIM and Statistical Modelling*, pages 91–100. Springer-Verlag, New York.

Firth, D. (1992b). Generalized linear models and Jeffreys priors: an iterative weighted least-squares approach. In Dodge, Y. & Whittaker, J., editors, *Computational Statistics*, volume 1, pages 553–557, Heidelberg. Physica-Verlag.

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

Foxman, B., Marsh, J., Gillespie, B., Rubin, N., Koopman, J. S., & Spear, S. (1997). Condom use and first-time urinary tract infection. *Epidemiology* **8**, 637–641.

Heinze, G. (1999). *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.

Heinze, G. & Ploner, M. (2002). SAS and SPLUS programs to perform Cox regression without convergence problems. *Computer methods and programs in Biomedicine* **67**, 217–223.

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

Heinze, G. & Schemper, M. (2001). A solution to the problem of monotone likelihood in Cox regression. *Biometrics* **57**, 114–119.

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

Insightful Corp. (2003). *S-PLUS 6.2*. Reinach, Switzerland.

Ploner, M. (2001). *Technical Report 2: An S-PLUS library to perform logistic regression without convergence problems*. Department of Medical Computer Sciences, Section of Clinical Biometrics, Vienna University, Vienna.

SAS Institute (1999). *SAS/STAT User's Guide, Version 8*. SAS Institute Inc., Cary, NC.

Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression.

Statistics in Medicine **2**, 71–78.

Venzon, D. J. & Moolgavkar, S. H. (1988). A method for computing profile-likelihood based confidence intervals. *Applied Statistics* **37**, 87–94.