# PLS generalized linear regression
## Application to the analysis of life time data

**Philippe BASTIEN[1] , Michel TENENHAUS[2]**

(1) L'Oréal Research, Clichy, France (pbastien@recherche.loreal.com)
(2) HEC School of management, Jouy-en-Josas, France (tenenhaus@hec.fr)

### SUMMARY

Problems encountered in multiple regression due to multicolinearity or missing data can be overcome by using PLS regression. Several versions of the PLS regression algorithm exist. In this paper, we present a new version of this algorithm which can be extended to generalized linear regression models such as ordinal or multinomial logistic regression, generalized linear models, and Cox regression models. An application with discrete censored data concerning the occurrence of prematurely graying hair in more than 4000 adult males is presented. The Cox regression model has been used. For discrete data summarizing a continuous time process, this model is equivalent to a generalized linear model using a complementary log-log link function.

**key words** : PLS regression; Generalized linear regression; Life time data; Survival analysis; Cox model.

## I PLS generalized linear regression

The PLS regression method in its basic form ( PLS 1) applies for one single response variable Y and is non iterative (Tenenhaus 1998). It is particularly useful when the X variables are closely correlated with each other.

Marx (1996) proposed a generalization of the PLS algorithm to generalized linear models (Dobson 1990). He used the fact that, in the context of the exponential family, maximum likelihood estimates are obtained by an iterative weighted least squares procedure. His approach consisted of replacing the iterative weighted least squares step by a sequence of PLS regressions.

Esposito Vinci and Tenenhaus (2001) propose a much simpler approach consisting of adapting the PLS regression algorithm described by Wold, Martens and Wold, (1983) to the case of binary or ordinal logistic regression. In this paper, we present a generalization of this algorithm to generalized linear regression.

The generalized linear regression model is described below :

Let Y be a response function and $\tau$ some parameter related to the distribution of Y. For example $\tau$ can be the mean $\mu$, the probability $\pi$ of occurrence of an event, or a hazard function h(t). Let $x_1,..,x_p$ be the explanatory variables. Let $g$ be the link function.

The generalized linear regression is the following :

(1) $$g(\tau) = \sum_j \beta_j x_j$$

The model parameters $\beta_j$ are estimated by maximum likelihood.

The PLS generalized linear regression algorithm is described below:

Let $X_0$ be the matrix of the standardized input variables $x_{01},..,x_{0p}$.

***Determination of the first PLS component $t_1$***

1: For each *j = 1* to *p*, compute the regression coefficient $w_{1j}$ of $x_{0j}$ in the generalized linear regression model (1) of y on $x_{0j}$.
2: The vector $w_1 = (w_{11},\ldots,w_{1p})'$ is normalized.
3: $t_1 = X_0w_1/w_1'w_1$

### *Determination of the second PLS component $t_2$*

1: Compute the residual $X_1$ of the regression of $X_0$ on $t_1$.
2: Compute the regression coefficient $w_{2j}$ of $x_{1j}$ in the generalized linear regression of y on $t_1$ and $x_{1j}$, for $j = 1$ to $p$.
3: The vector $w_2 = (w_{21},\ldots,w_{2p})'$ is normalized.
4: $t_2 = X_1 w_2 / w_2' w_2$

This procedure is iterated for the other PLS components $t_h$.

When some data are missing, the calculation of the PLS components is modified by applying the NIPALS algorithm principle. The calculation of the numerator and denominator of $t_h = X_{h-1} w_h / w_h' w_h$ is carried out on the basis of the available data. The denominator is in fact calculated only on data available to the numerator.

At each step the generalized linear regression of y on components $t_1,\ldots, t_h$ is carried out. We stop the procedure and the component $t_h$ is not included in the model if it is not significant. The number *m* of PLS components $t_h$ can also be determined by cross-validation.. The final regression equation is obtained by expressing the generalized linear regression of y on $t_1,\ldots, t_m$ as a function of the original variables.
In the case of ordinary multiple regression, this algorithm gives usual PLS regression when there is no missing data. When some missing data are present this algorithm takes into account the correlation between the PLS components. This is not the case for the usual PLS regression.

## II Application

### *II.1 Material and objectives*

Begun in 1994 by Professor Serge Hercberg, the cohort study SUVIMAX *(«SUpplémentation en VItamines et Minéraux Antioxydants»)* assumed the task of evaluating the nutritional state of the population in France and assessing the influence of antioxidant minerals and vitamins on various indications of the state of health such as cancer or cardiovascular disease (Hercberg 1997). This study, scheduled for completion after 8 years in 2002, has enrolled more than 12000 volunteers aged between 35 and 65 years and representative of the population of France, half of whom received antioxidants and the other half a placebo.

We conducted a study of the state of health of hair and nails in 10323 subjects (4057 men and 6266 women) from this cohort. On the basis of responses to a questionnaire covering more than 150 items, an attempt was made to demonstrate the risk factors associated with the premature onset of graying hair in men.

The response variable is the age at which the first gray hairs appear in men. This is a discrete variable divided into 6 stages: less than 30 years, 31 to 35 years, 36 to 40 years, 46 to 50 years and over 50 years. The Cox regression model was used to fit the data. As the process of the advent of the first gray hairs is intrinsically continuous, the Cox regression model is equivalent to a generalized linear model using a complementary log-log link function (Allison, 1995, page 216).

Family factors, natural hair color, and hair thickness are predictor variables significantly related to the time at which the first gray hairs appear. A Cox model was used to relate the response variable to the predictors.

### *II 2 Discrete life time data*

The basic ideas described in Allison (1995) are simple. Each individual’s survival history is broken down into a set of discrete time units that are treated as distinct observations. After pooling these observations, the next step is to estimate a binary regression model to predict whether an event did or did not occur in each time unit. Even if multiple observations are created for a single individual, there is no concern about dependence here. The creation of multiple observations follows directly from factoring the likelihood function for the data. It follows immediately from the definition of conditional probability.

Two models for processing these data exist: one that assumed that events really occur at the same discrete time and another that assumed that ties result from imprecise measurement.

When events can only occur at regular discrete points in time, the appropriate model is a logit model.

Let $P_{it}$ be the conditional probability that individual i has an event at time t, given that the event has not already occurred in the case of that individual. The model states that $P_{it}$ is related to the predictors by the logit regression equation:

(2) $$\log \frac{P_{it}}{1-P_{it}} = \alpha_t + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

For most applications, however, ties occur because event times are measured coarsely even though events can actually occur at any point in time. If we now assume that events are generated by Cox's proportional hazards model it follows that:

(3) $$\log\left[-\log(1-P_{it})\right] = \alpha_t + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

i.e. the probability of an event at some well-defined interval of time is given by the complementary log-log model. Furthermore the β coefficients in the model have the same relative risk interpretation as in the underlying Cox proportional hazard model. The expression $-\log(1-P_{it})$ in formula 3 is the discrete version of the hazard function $h_i(t)$. Hence, formula 3 is the discrete version of the Cox proportional hazard model :

(4) $$\log\left[h_i(t)\right] = \alpha_t + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

## *II.3 Results*

All analyses have been carried out using SAS® software release 8.0.

### *II.3.1 The PLS Complementary log-log model*

The first two PLS components $t_1$,$t_2$ are significant and have been retained in the model. Table 1 below describes $t_1$ and $t_2$, based on all data, as a function of the original predictors.

**Table 1**: PLS components of model 3 in terms of the original variables.

| | $t_1$ | $t_2$ |
|---|---|---|
| Mother gray-haired under the age of 30 | 1.92 | -2.78 |
| Mother with no gray hair at 60+ years | -1.50 | -2.0 |
| Sister gray-haired under the age of 30 | 1.91 | -3.6 |
| Father gray-haired under the age of 30 | 2.23 | 0.35 |
| Father with no gray hair at 60+ years | -1.35 | 0.63 |
| Natural hair color<br>Red-blonde-light chestnut dark chestnut brown black | 0.45 | 0.06 |
| Hair thickness<br>Very fine – fine – medium – thick | 0.55 | 0.6 |

Results of the PLS Cloglog model based on all data are presented in Table 2.

**Table 2**: Parameter estimates of model 3 in terms of PLS components.

```
Parameter        DF    Estimate     Std Err    ChiSquare  Pr>Chi

INTERCEPT         1      0.4213      0.0483      76.2043  0.0001
T1                1      0.2247      0.0119     357.1172  0.0001
T2                1      0.0220      0.0108       4.1925  0.0406
TIME        1     1     -3.0941      0.0763    1644.8937  0.0001
TIME        2     1     -2.8345      0.0730    1509.6531  0.0001
TIME        3     1     -1.9391      0.0617     987.0238  0.0001
TIME        4     1     -1.1353      0.0575     389.4365  0.0001
TIME        5     1     -0.7232      0.0598     146.3576  0.0001
TIME        6     0      0.0000      0.0000            .       .
```

Intercept is an estimate of Time 6, the log-hazard for the onset of the first gray hairs over 50 years. For level j of the Time variable, the coefficient is the difference in the log-hazard of onset of the first gray hairs in interval j and the log-hazard of onset of the first gray hairs over 50 years.

The results of the Cloglog and PLS Cloglog models expressed according to the original predictors are given in table 3. The Cloglog parameters are estimated using observations without missing data. In order to compare PLS Cloglog and Cloglog models on the same basis, PLS Cloglog parameters have also been estimated on the observations without missing data. Finally PLS Cloglog model has been fitted on the whole data set. This possibility of working with missing data is a rather interesting feature of PLS methods.

**Table 3**: Parameter estimates of model 3 in terms of the original variables.

| | Cloglog<br>$\beta$<br>(exp $\beta$)<br>no missing | Cloglog<br>PLS<br>$t_1$<br>$t_1$ , $t_2$<br>no missing | Cloglog<br>PLS<br>$t_1$<br>$t_1$ , $t_2$<br>all data |
|---|---|---|---|
| Mother gray-haired<br>Under the age of 30 | 0.51<br>(1.65) | 0.58<br>0.50 | 0.42<br>0.37 |
| Mother with no gray hair<br>at 60+ years | -0.44<br>(0.65) | -0.39<br>-0.44 | -0.33<br>-0.38 |
| Sister gray-haired<br>under the age of 30 | 0.63<br>(1.88) | 0.73<br>0.63 | 0.42<br>0.35 |
| Father gray-haired<br>under the age of 30 | 0.82<br>(2.24) | 0.74<br>0.81 | 0.49<br>0.51 |
| Father with no gray hair<br>at 60+ years | -0.44<br>(0.64) | -0.40<br>-0.44 | -0.30<br>-0.29 |
| Natural hair color<br>Red-blonde-light chestnut<br>dark chestnut brown black | 0.14<br>(1.14) | 0.13<br>0.14 | 0.10<br>0.10 |
| Hair thickness<br>Very fine – fine – medium – thick | 0.13<br>(1.14) | 0.13<br>0.13 | 0.12<br>0.14 |

We can interpret the coefficients just as if this was a proportional hazards model. For instance, having a mother with gray hair under the age of 30 produce a 100(exp(0.51) – 1 ) = 65 percent increase in the hazard of occurrence of gray hair with respect to having a mother without gray hair under the age of 30, all the other variables being fixed.

In the classical multivariate approach, individuals for whom one or more variables are unknown are excluded from the analysis. They represent here more than half the population. In order to show convergence of the PLS Cloglog model coefficients towards those of the classic Cloglog model when the design matrix is not ill-conditioned, PLS model based on complete data has been carried out.
Results are displayed in the third column of Table 3. The coefficients of the PLS Cloglog model with two PLS components appear to be very close to those of the classic Cloglog model.

The loss of information due to a very high percentage of missing data demonstrate the advantage of using a PLS model here, allowing all the individuals to be taken into account. In an extreme case, the classic approach could collapse without the PLS model being really affected.
The last column of the table displays coefficients of the PLS Cloglog model based on all data. In spite of the high percentage of missing data results in the last two column are homogeneous.

### *II.3.2 The PLS Cox model*

Table 4 below presents the results of the classic Cox model (formula 4) and of the PLS Cox regression model in terms of the original predictors. The data are assumed here to be continuous without loss of generality. To take into account the ties, Efron’s approximation for tied event time was used.

The convergence of the coefficients of the PLS Cox model towards those of the classic model with two components is again observed. The remarks in the preceding paragraph remain valid here.

**Table 4**: Parameter estimates of model (4) in terms of the original variables

| | Cox<br>β<br>exp(β)<br>no missing | Cox<br>PLS<br>$t_1$<br>$t_1 , t_2$<br>no missing |
|---|---|---|
| Mother gray-haired<br>Under the age of 30 | 0.48<br>(1.62) | 0.54<br>0.47 |
| Mother without gray hair<br>at 60+ years | -0.42<br>(0.66) | -0.37<br>-0.41 |
| Sister gray-haired<br>Under the age of 30 | 0.58<br>(1.79) | 0.68<br>0.58 |
| Father gray-haired<br>Under the age of 30 | 0.77<br>(2.16) | 0.69<br>0.76 |
| Father without gray hair<br>at 60+ years | -0.41<br>(0.66) | -0.39<br>-0.42 |
| Hair color<br>Red – blond – light chestnut-dark chestnut<br>Brown black | 0.13<br>(1.14) | 0.12<br>0.12 |
| Hair thickness<br>Very fine – fine – medium – thick | 0.12<br>(1.13) | 0.12<br>0.12 |

## III Conclusion

In this article, it has been demonstrated that the PLS regression can be extended to generalized linear regression and, more specifically, to survival data analysis. Initially, only discrete data were handled in order to fit into the framework of generalized linear models with logit or complementary log-log models. Subsequently, the capacity of the algorithm for extension to any linear regression model outside the exponential family was used to extend the concept of PLS regression to the Cox regression model. This approach offers a true alternative to the generalized linear regression models in the event of missing data or strong colinearity.

## References

[1] Allison, Paul D.(1995) : *Survival Analysis Using the SAS System*: *A practical guide*, SAS® Inc, Cary, NC.

[2] Dobson, A.J. (1990), *An Introduction to generalized linear Model*, London : Chapman and Hall.

[3] Esposito Vinci, V. and Tenenhaus, M. (2001): PLS logistic regression, *2nd International Symposium on PLS and Related Methods*, Capri, October 1st-3rd, 2001.

[4] Hercberg, S. et al. (1997) : A primary prevention trial of nutritional doses of antioxidant vitamins and minerals on cardiovascular diseases and cancers in general populations : The SUVIMAX Study. Design, methods and participants characteristics. *Control Clin. Trials*.

[5] Marx, B.D. (1996) : Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression. *Technometrics,* vol. 38, n°4, pp. 374-381.

[6] Tenenhaus, M.(1998): *La régression PLS*. Paris : Technip.

[7] Wold, S., Martens, H. & Wold, H., (1983). The multivariate calibration problem in chemistry solved by the PLS method, in *Proc. Conf. Matrix Pencils*, Ruhe A. & Kågstrøm B. (Eds), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, pp. 286-293.