# ELECTRONIC SPREADSHEETS AS A TEACHING AID TO A GENERALIZED LINEAR MODEL COURSE

Kaizô Iwakami Beltrão and Sonoe Sugahara
National School of Statistics, ENCE/IBGE, Brazil
kaizo@ibge.gov.br

*Using a software package to obtain parameters estimates for a given model can be straightforward and fast, but does not inform the user about the underlying process involved. On the other hand, using a spreadsheet to implement the operations, to eventually arrive to a final answer, assumes the analytical development of some previous steps necessary to attain this goal and therefore, besides making explicit the underlying GLM estimating procedure, reinforces previous knowledge and competences enriching the learning process.*

## INTRODUCTION

The use of regular statistical packages does not allow the students to fully understand the underlying principles governing generalized linear models (GLMs). As demonstrated by Nelder and Wedderburn (1972), GLM puts in a single framework many statistical methods usually taught under different disciplines, such as Planning of Experiments, Multiple and Polynomial Regressions, Logistic Regressions, Harmonic Analysis, etc. It is rather hard for the student to grasp GLM unifying capability when using a statistical package that does not spell out the subjacent process. On the other hand, using excel or other electronic spreadsheet, students have to explicitly handle all the model components and multiple links at a very basic level. Once adjusting a model for a given combination of probability distribution, link function and set of covariates, it is quite straightforward to go on to other combinations of these components.

## GENERALIZED LINEAR MODEL

To put into context we use Dobson (2002) notation. A GLM is defined for a set of independent random variables $Y_1$, $Y_2$,…$Y_N$, each one with a distribution from the exponential family and the following properties:

- variables $Y_i$ share the same distribution which has the canonical form $f(y;\theta_i) = s(y)t(\theta_i)e^{y\,b(\theta_i)}$ [a non restricted form of exponential family distribution considers *a(y)* instead of *y*] and depends on just one parameter $\theta_\iota$ ,$\forall i$, *i=1…N*. This expression can also be written as $f(y;\theta_i) = \exp\left[y\,b(\theta_i) + c(\theta_i) + d(y)\right]$ where $d(y) = \ln s(y)$ and $c(\theta_i) = \ln t(\theta_i)$. Note that functions *b(.)*, *c(.)* and *d(.)* do not carry the *i* subscript, because they are all the same;
- there is a vector of parameters, $\beta$ and for each $Y_i$ there is a vector of co-variates (explanatory variables), $x_i^t = \left(x_{i1}\, x_{i2}\, x_{i3}\, x_{i4} \cdots x_{ip}\right)$, $\forall i$, *i=1…N*;
- there is a monotonic, differentiable function *g* called link function such that $g(\mu_i) = X_i^T \beta$ where $E(Y_i) = \mu_i$, $\forall i$, *i=1…N*.

The maximum likelihood estimator (MLE) *b* for $\beta$ does not always present a closed form (unless *Y* is normally distributed) and it is (in general) obtained through an iterative process up to convergence: in step *m* one looks after the solution to the equation $X^TWXb^{(m)} = X^TWz$ where *W* is a *N*x*N* diagonal matrix with elements $w_{ii} = \dfrac{1}{\mathrm{var}(Y_i)}\left(\dfrac{\partial\mu_i}{\partial\eta_i}\right)^2$ and *z* is a *N* dimensional vector with elements $z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + \left(y_i - \mu_i\right)\left(\dfrac{\partial\eta_i}{\partial\mu_i}\right)$. The matrix $\Im = X^TWX$ is called the information matrix. Note that at step m, *W* and *z* are functions of the unknown parameter $\beta$ and are evaluated using the estimator for $\beta$ from the previous step, $b^{(m-1)}$.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

The estimation process is equivalent to a Weighted Least Square Procedure, where the weights may depend on the parameter being estimated. For a given data set of *N* observations $Y_i$, $\forall i$, *i=1…N.*, and *p* explanatory variables $x_i$, $\forall i$, *i=1…N*, the steps to follow in order to adjust a GLM are:

1. assume a distribution belonging to the exponential family consistent with the data characteristics (e.g., continuous x discrete, left or right bounded or not);
2. assume a link function whose domain is defined as the space of possible values of $\mu_i$ (to insure that all values obtained are permissible) and the image takes values in $\Re$. For example, with a binomial distribution, we should have $g:[0,1] \to \Re$;
3. obtain the analytical expression for *W* and *z* for the assumptions made in steps 1 and 2;
4. perform the iterative process: assume an initial value for *b*, $b^{(0)}$, and compute the *W* matrix, the information matrix $\Im$ and vector *z*. A new estimate for *b* is obtained by $b^{(m)} = (X^TWX)^{-1}X^TWz$, in this case with *m=1*, whenever the matrix is not invertible, a generalized inverse can be used;
5. with this new estimate for *b*, the process is repeated until convergence is attained.

EXAMPLES

As an example of the use of the excel program for estimating a GLM via an iterative process, the data on beetle mortality in Dobson book is presented (example 7.3.1 in section 7.3). The table below shows numbers of beetles dead after five hours exposure to gaseous carbon disulphide at various concentrations.

Table 1. Beetle mortality data

| I | Number killed, $y_i$ | Dose, $x_i$ ($\log_{10}$CS2mgl$^{-1}$) | Number of beetles, $n_i$ |
|---|---|---|---|
| 1 | 6 | 1.6907 | 59 |
| 2 | 13 | 1.7242 | 60 |
| 3 | 18 | 1.7552 | 62 |
| 4 | 28 | 1.7842 | 56 |
| 5 | 52 | 1.8113 | 63 |
| 6 | 53 | 1.8369 | 59 |
| 7 | 61 | 1.8610 | 62 |
| 8 | 60 | 1.8839 | 60 |

The binomial distribution $P(Y=y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$ is a natural candidate and possible link functions are the logistic, the probit, the complementary log-log and the tangent function. For our example in the use of excel, selected analytical equations for each one these link functions is presented in Table 2.

Table 2. Analytical functions involved in MLG – for selected link function

| | **logit** | **probit** | **complementary log- log** | **tangent** |
|---|---|---|---|---|
| $\eta = g(\theta)$ | $\ln\frac{\theta}{1-\theta}$ | $\Phi^{-1}(\theta)$ | $\ln[-\ln(1-\theta)]$ | $tg\pi\left(\theta-\frac{1}{2}\right)$ |
| $\theta = g^{-1}(\eta)$ | $\frac{\exp(\eta)}{1+\exp(\eta)}$ | $\Phi(\eta)$ | $1-\exp[-\exp(\eta)]$ | $\frac{1}{\pi}arctg(\eta)+\frac{1}{2}$ |
| $\frac{\partial\eta}{\partial\mu}$ | $\frac{\theta}{\theta(1-\theta)}$ | $\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)$ | $-\frac{1}{(1-\theta)\ln(1-\theta)}$ | $\pi\sec^2\left(\pi\left(\theta-\frac{1}{2}\right)\right)$ |
| $w_{ii}$ | $n\theta(1=\theta)$ | $\frac{n}{\theta(1=\theta)}\left(\frac{\sqrt{2\pi}}{\exp\left(-1/2\,y^2\right)}\right)^2$ | $\frac{n(1-\theta)}{\theta}\ln^2(1-\theta)$ | $\frac{n\cos^4\left(\pi(\theta-\frac{1}{2}\right)}{\theta(1-\theta)\pi^2}$ |

Note: ln =natural logarithm, $\Phi$ =cumulative distribution of the standardized normal.

To proceed to the implementation of the abovementioned steps it is necessary to employ the following excel functions (Microsoft Office Online, 2010):

- MMULT(.,.) – to multiply two matrices;
- TRANSPOSE(.) – to transpose a matrix
- MINVERSE(.) – to invert a matrix
- NORMSDIST(.) – to obtain the standard normal cumulative distribution
- NORMSINV – to obtain the inverse of the standard normal cumulative distribution
- And other functions such as ln(.), tan(.), atan(.), cos(.) etc.

Tables 3 and 4 present, respectively standard residuals for each of the link functions considered and estimated parameters with corresponding deviance. Standardized residuals are also displayed in Figure 1. Among the regular link functions, complementary log-log presents the smallest deviance. This is due to the fact that an asymmetric function was needed due to the data pattern. Taking this fact into consideration a modification of the tangent was used (noted by tangent*), substituting the centralizing constant ½ for an offset 0,575. The result was better than using the complementary log-log. The model tangent* has a draw back, though. The domain of the link function proposed does not coincide with the interval [0,1] and therefore, could produce out of range values, though, not with this dataset.

Table 3. Standardized Residuals under alternative link functions

| I | logit | probit | Complementary log-log | tangent | tangent* |
|---|---|---|---|---|---|
| 1 | 1,4093 | 1,4848 | 0,1825 | 0,5093 | -1,0083 |
| 2 | 1,1011 | 0,7678 | 0,5681 | 1,9388 | 0,6537 |
| 3 | -1,1763 | -1,4353 | -0,7932 | 0,4939 | 0,4283 |
| 4 | -1,6124 | -1,5889 | -0,6355 | -1,5925 | -0,1893 |
| 5 | 0,5944 | 0,7344 | 1,2430 | 0,2196 | 0,0770 |
| 6 | -0,1281 | -0,1407 | -0,5413 | 0,3326 | -1,0683 |
| 7 | 1,0914 | 0,8907 | -0,1212 | 1,9341 | 0,3679 |
| 8 | 1,1331 | 0,8844 | 0,2298 | 2,0762 | 0,1792 |

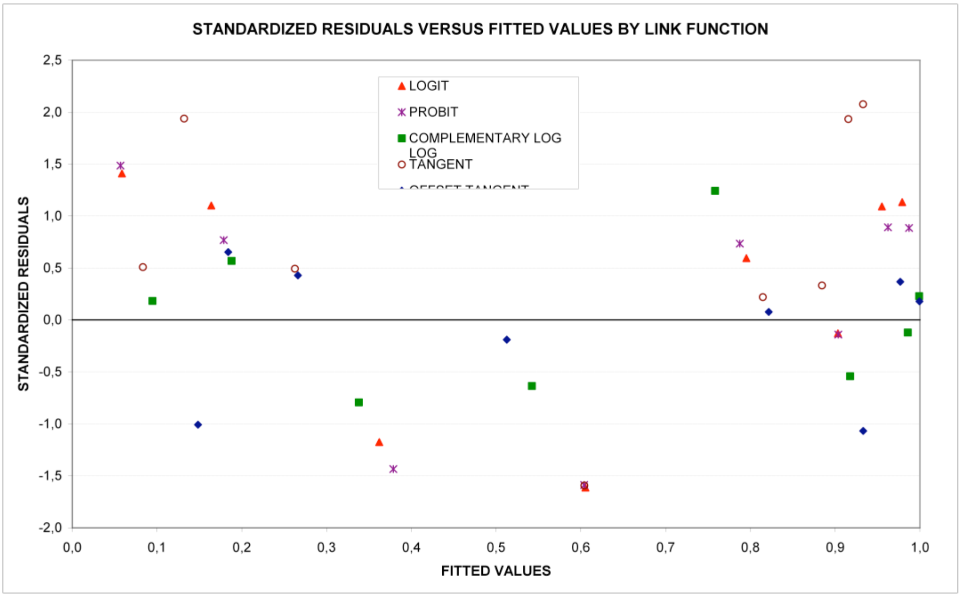

Figure 1. Standardized Residuals versus Fitted Values by Link Function

Table 4. Estimated Parameters and Deviance under alternative link functions

| | **logit** | **probit** | **Complementary log-log** | **tangent** | **tangent*** |
|---|---|---|---|---|---|
| $b_1$ | -60,7175 | -34,9353 | -39,5723 | -77,3200 | -77,7425 |
| $b_2$ | 34,2703 | 19,7279 | 22,0412 | 43,5260 | 43,4616 |
| DEVIANCE | 11,2322 | 10,1198 | 3,4464 | 20,1582 | 2,9640 |

## ADVANTAGES OF EXCELL VIS-À-VIS STATISTICAL SOFTWARE SUCH AS SPSS OR SAS

First and foremost, the analytical derivation of all the formulae involved is an important step in the cognitive process. Spreadsheets, as opposed to statistical packages, are easily available since they are part of the standard windows or Linux set up. Students are usually familiar with spreadsheets and their graphic capabilities are also well known to regular users. Another point that should be mentioned is the fact that usually in less developed countries statistical packages are expensive and available only at school or major companies. This makes it difficult for student to work on GLM cases at home or in smaller firms, for example.

The abovementioned steps have to be followed one by one which reinforces the understanding of the underlying general properties of GLM. Also important is the process of choosing an initial value. Sometimes convergence depends on initial values, sometimes they don't. To expose students to prepared sets with both characteristics re-introduces topics such as optimization process, local maxima and minima etc. The student can fully experience the iterative approach process, besides the initial value, the speed of convergence can be checked and compared across link functions, for example.

As one of the steps the student has to invert a matrix. Theory states that whenever an inverse does not exist, a generalized inverse can be used instead. Usually to deal with this situation, softwares assume a restriction, such as a vertices restriction (the first or the last parameter in a given subset is set to zero) or a zero sum restriction (the sum of the parameters in a given subset are set to zero). Including one of these restrictions in the spreadsheet re-introduce topics from Linear Algebra such as linear dependency, generalized inverse etc.

Though one could program the iterative process in a statistical package using weighted least square, and the students would have to develop the analytical formulae to define the weights, this procedure is much lengthier and not as user friendly and easily followed as in a spreadsheet.

Note that the iterative process described is equivalent to obtaining a maximum of the likelihood function using Newton-Raphson , with

$b^{(m)} = b^{(m-1)} + (\Im)^{-1} U^{(m-1)}$, where $U^{(m-1)}$ is the score function evaluated at sep *m-1*, and several topics included in Numerical Calculus can be revised at this point.

## CONCLUSION

The practical exercise to write down the analytical formulae for the information matrix, score function etc under various probability distributions and different link functions in order to prepare the iterative process in the spreadsheets helps the student to better understand the underlying properties of the GLM, as already mentioned. On the other hand, since the student has also to deal with matrix inversion, iterative processes, derivatives, integrals, non-linear optimization etc, it brings together Statistical and other Mathematical Disciplines like Linear Algebra, Calculus and Numerical Calculus and this reinforces the learning process.

## REFERENCES

Dobson, A. (2002). *An introduction to generalized linear models* (second edition). Boca Raton: Chapman & Hall/CRC.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*, Second Edition, London: Chapman and Hall.

Microsoft Office Online. (2010). *Microsoft Office Excel/Function reference*. Available at http://office.microsoft.com/en-us/excel/CH100645021033.aspx

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, 135*, 370-384.