Biometrical Letters
Vol. 45 (2008), No. 1, 69-80

Double minimization for logit models with an additive two factors structure

Inês Jorge Sequeira[1], João Tiago Mexia[1], Sandra Nunes[2]

[1]Department of Mathematics, Faculty of Science and Technology, New University of Lisbon,
Quinta da Torre 2829-516 Caparica, Portugal, e-mail: ijs@fct.unl.pt
[2]Department of Mathematics, EST/IPS, Campus do IPS, Estefanilha, 2910-761 Setúbal, Portugal,
e-mail: snunes@est.ips.pt

SUMMARY

Logit models may be used to express the incidence rate of diseases when the impact depends on two, or more, additive factors. A double minimization algorithm is presented for the adjustment of such models for the case of two additive factors. An application to the incidences of Tuberculosis and AIDS is presented.

Key words: Logit models, double minimization algorithm

1. Introduction

Nunes et al (2004) used a Zigzag algorithm to adjust logit models with anadditive two factors structure, but they did not establish the convergence of the algorithm to the absolute minimum of the objective function.

In this paper we develop an alternative to the Zigzag algorithm which does not present convergence problems. We will make the comparison between both algoritms. The paper is organized as follows. In Section 2 we describe the logit model. In Section 3 we briefly outline the Zigzag algorithm. The new algorithm and the key ideas of this paper are reported in Section 4. Lastly, we present an application to the incidence of Tuberculosis (TB) and of acquired immune deficiency syndrome (AIDS) in European Union countries and compare the results obtained with the Zigzag algorithm.

As we shall see both algorithms display a significant agreement. This is important since only for the new algorithm we have a proof of convergence. Actually the classical Zigzag algorithm is easier to apply but the new algorithm works well with, for instance, `Mathematica` software.

2. Logit Model

Logit models are a good choice to express the incidence rate of diseases when the impact depends on two (or more) additive factors. We take as dependent variable a binary variable,

$$\begin{cases} Y = 1 & \text{if the individual is infected with some disease} \\ Y = 0 & \text{if not} \end{cases}.$$

In our application the binary logistic regression is more adequate than the linear regression. Let us assume the logit model

$$y_{i,j} = \log it(p_{i,j}) = \ln \frac{p_{i,j}}{1 - p_{i,j}} = \alpha + \beta \cdot x_{i,j} \, ,$$

where

- $p_{i,j}$ represents the probability of an event Y occurring. In this case it represents the probability of an individual being infected with some disease in country i $(i = 1, \dots, m)$ during year j $(j = 1, \dots, n)$,
- $x_{i,j}$ represents the exposure,
- α and β are parameters, β representing the rate of variation of y with x.

Since the available information is the incidence rates for the pairs of countries and years we need to formulate our model in such a way that we may estimate both the regression coefficients α and β and the exposures $x_{i,j}$, $i = 1, \dots, m$, $j = 1, \dots, n$. To do this we assume that $x_{i,j} = f_i + g_j$, this is we assume for the exposures a two factor additive sub-model.

The two factors will be,

- a location factor whose levels $f_1, \dots, f_m$ correspond to countries,
- a time factor whose levels $g_1, \dots, g_n$ correspond to years.

We now have the model

$$y_{i,j} = \log it(p_{i,j}) = \alpha + \beta\left(f_i + g_j\right)$$

and the goal function

$$S = \sum_{i=1}^{m} \sum_{j=1}^{n} q_{i,j}\left(y_{i,j} - \alpha - \beta\left(f_i + g_j\right)\right)^2 ,$$

where the weights $q_{i,j}$ are the inverses of the variances of $y_{i,j}$:

$$Var(y_{i,j}) \approx \frac{1}{N_{i,j} \times p_{i,j}},$$

$i = 1, \dots, m$, $j = 1, \dots, n$, where $N_{i,j}$ represents the population in country i and in year j. Thus we must estimate α, β, $f_1, \dots, f_m$ and $g_1, \dots, g_n$.

3. Zigzag Algorithm

The Zigzag algorithm is an iterative algorithm. To initialize this algorithm we assume as initial values for $x_{i,j}$ the following ones

$$x_{i,j}(0) = y_{i,\bullet} + y_{\bullet,j} - y_{\bullet,\bullet}, \; i = 1, \dots, m, \; j = 1, \dots, n,$$

where

$$y_{i,\bullet} = \frac{1}{n} \sum_{j=1}^{n} y_{i,j} , \; y_{\bullet,j} = \frac{1}{m} \sum_{i=1}^{m} y_{i,j} , \; y_{\bullet,\bullet} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} y_{i,j} .$$

So we may write

$$S(0)=\sum_{i=1}^{m}\sum_{j=1}^{n}q_{i,j}\left(y_{i,j}-\alpha-\beta x_{i,j}(0)\right)^2=$$
$$=\sum_{i=1}^{m}\sum_{j=1}^{n}q_{i,j}\left(y_{i,j}-\alpha-\beta\left(f_i(0)+g_j(0)\right)\right)^2.$$

To lighten the notation let us put $x_{i,j}(0)=x_{i,j}$.

Then we are in conditions to describe the several steps of each iteration:

Step 1. In the first step we minimize S in order to the parameters (α,β), using the initials values of $x_{i,j}$. From this minimization we obtained the following estimates:

$$\hat{\alpha}(1)=\hat{\alpha}=y_\circ-\hat{\beta}x_\circ \text{ and } \hat{\beta}(1)=\hat{\beta}=\frac{s_{x,y}}{s_{x,x}},$$

where $y_\circ=\dfrac{\sum_{i=1}^{m}\sum_{j=1}^{n}q_{i,j}y_{i,j}}{q^+}$, $x_\circ=\dfrac{\sum_{i=1}^{m}\sum_{j=1}^{n}q_{i,j}x_{i,j}}{q^+}$ with $q^+=\sum_{i=1}^{m}\sum_{j=1}^{n}q_{i,j}$

and $s_{x,x}=\sum_{i=1}^{m}\sum_{j=1}^{n}q_{i,j}\left(x_{i,j}-x_\circ\right)^2$, $s_{x,y}=\sum_{i=1}^{m}\sum_{j=1}^{n}q_{i,j}\left(x_{i,j}-x_\circ\right)\left(y_{i,j}-y_\circ\right)$.

Step 2. In this step we minimize

$$S=\sum_{i=1}^{m}\sum_{j=1}^{n}q_{i,j}\left(y_{i,j}-\hat{\alpha}-\hat{\beta}\left(f_i+g_j\right)\right)^2$$

in order to the vectors $\underline{f}^m$ and $\underline{g}^n$. We then solve the system:

$$\begin{bmatrix} D_1 & Q \\ Q^T & D_2 \end{bmatrix}\begin{bmatrix} \underline{f}^m \\ \underline{g}^n \end{bmatrix}=\underline{V}^{m+n},$$

where $Q=\lfloor q_{i,j}\rfloor$, $D_1=D\left(\sum_{j=1}^{n}q_{1,j},\dots,\sum_{j=1}^{n}q_{m,j}\right)$ and $D_2=D\left(\sum_{i=1}^{m}q_{i,1},\dots,\sum_{i=1}^{m}q_{i,n}\right)$

and the components of $\underline{V}^{m+n}$ are:

$$V_i = \frac{1}{\hat{\beta}} \sum_{j=1}^{n} q_{i,j}(y_{i,j} - \hat{\alpha}),\ i = 1,\dots,m,$$

$$V_{m+j} = \frac{1}{\hat{\beta}} \sum_{i=1}^{m} q_{i,j} (y_{i,j} - \hat{\alpha}).\ \ j = 1,\dots,n.$$

The solutions of the previous system will be $\hat{f}_i(1)$, $i = 1,\dots,m$ and $\hat{g}_j(1)$, $j = 1,\dots,n$ and, consequently, $\hat{x}_{i,j}(1) = \hat{f}_i(1) + \hat{g}_j(1)$.

Step 3. In the third step we calculate the sum of square of residues

$$\hat{S}(1) = \hat{S} = \sum_{i=1}^{m} \sum_{j=1}^{n} q_{i,j} \left(y_{i,j} - \hat{\alpha}(1) - \hat{\beta}(1)\left(\hat{f}_i(1) + \hat{g}_j(1) \right) \right)^2 ,$$

where $\hat{\alpha}(1)$, $\hat{\beta}(1)$, $\hat{f}_i(1)$, $i = 1,\dots,m$ and $\hat{g}_j(1)$, $j = 1,\dots,n$ are the adjusted values obtained in iteration 1.

Step 4. In this last step we carry out the standardization, of the $x_{i,j}$, in order to keep unchanged the minimum and the maximum of $x_{i,j}$. We compute

$$\hat{w}_{i,j}(1) = \frac{b-a}{b(1) - a(1)} \times \left(\hat{x}_{i,j}(1) - a(1) \right) + a ,$$

where

$$a = \min\{x_{i,j}(0)\},\ b = \max\{x_{i,j}(0)\},$$

$$a(1) = \min\{x_{i,j}(1)\},\ b(1) = \max\{x_{i,j}(1)\}$$

with $i = 1,\dots,m$ and $j = 1,\dots,n$.

The values obtained from this standardization will be used in the next iteration. This procedure is repeated till the sum of squares of residues stabilizes, see for instance, Mexia *et al.* (1999).

4. Double Minimization Algorithm

To apply this algorithm we reparametrize the goal function taking

$$\begin{cases} \alpha_j = \alpha + \beta g_j \,, & j = 1, \dots, n \\ x_i = f_i \,, & i = 1, \dots, m \end{cases}$$

thus getting $S = \sum_{i=1}^{m} \sum_{j=1}^{n} q_{i,j} \left(y_{i,j} - \alpha_j - \beta x_i \right)^2$.

Nextly we obtain the minimum $s(\underline{x})$ of S for $\underline{x}$ known. This minimum will itself be minimized so that, as mentioned in the algorithm's name, we have a double minimization.

Step 1. As stated we now obtain $s(\underline{x})$. Since

$$S = \sum_{i=1}^{m} \sum_{j=1}^{n} q_{i,j} y_{i,j}^2 + \sum_{j=1}^{n} \left(\sum_{i=1}^{m} q_{i,j} \right) \alpha_j^2 + \left(\sum_{i=1}^{m} \sum_{j=1}^{n} q_{i,j} x_i^2 \right) \beta^2 - 2 \sum_{j=1}^{n} \left(\sum_{i=1}^{m} q_{i,j} y_{i,j} \right) \alpha_j$$
$$- 2 \left(\sum_{i=1}^{m} \sum_{j=1}^{n} q_{i,j} x_i y_{i,j} \right) \beta + 2 \sum_{j=1}^{n} \left(\sum_{i=1}^{m} q_{i,j} x_i \right) \alpha_j \beta,$$

we get

$$\frac{\partial S}{\partial \alpha_j} = 2 \left(\sum_{i=1}^{m} q_{i,j} \right) \alpha_j - 2 \sum_{i=1}^{m} q_{i,j} y_{i,j} + 2 \left(\sum_{i=1}^{m} q_{i,j} x_i \right) \beta, \; j = 1, \dots, n .$$

Solving the equation $\frac{\partial S}{\partial \alpha_j} = 0$, $j = 1, \dots, n$ we obtain

$$\alpha_{j\bullet} = y_{\bullet j} - \beta x_{\bullet j} \,, \; j = 1, \dots, n \qquad (1)$$

with $y_{\bullet j} = \frac{\sum_{i=1}^{m} q_{i,j} y_{i,j}}{\sum_{i=1}^{m} q_{i,j}}$ and $x_{\bullet j} = \frac{\sum_{i=1}^{m} q_{i,j} x_i}{\sum_{i=1}^{m} q_{i,j}}$, $j = 1, \dots, n$. Taking $\alpha_j = \alpha_{j\bullet}$ we get

$$\frac{\partial S}{\partial \beta} = 2\beta \sum_{j=1}^{n}\sum_{i=1}^{m} q_{i,j}\left(x_i - x_{\bullet j}\right)^2 - 2\sum_{j=1}^{n}\sum_{i=1}^{m} q_{i,j}\left(x_i - x_{\bullet j}\right)\left(y_{i,j} - y_{\bullet j}\right)$$

Solving the equation $\frac{\partial S}{\partial \beta} = 0$ we obtain

$$\beta_{\bullet} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} q_{i,j}\left(x_i - x_{\bullet j}\right)\left(y_{i,j} - y_{\bullet j}\right)}{\sum_{j=1}^{n}\sum_{i=1}^{m} q_{i,j}\left(x_i - x_{\bullet j}\right)^2} \tag{2}$$

Thus according to (1) and (2),

$$s(\underline{x}) = \sum_{i=1}^{m}\sum_{j=1}^{n} q_{i,j}\left(y_{i,j} - y_{\bullet j}\right)^2 - \frac{\left(\sum_{i=1}^{m} a_i x_i\right)^2}{\sum_{i=1}^{m}\sum_{j=1}^{n} q_{i,j}\left(x_i - x_{\bullet j}\right)^2}$$

with $a_i = \sum_{j=1}^{n} q_{i,j}\left(y_{i,j} - y_{\bullet j}\right),\ i = 1,\ldots,m$.

Step 2. In this step we minimize $s(\underline{x})$ as a function of $\underline{x}^m$. Let us observe that minimizing $s(\underline{x})$ is equivalent to maximizing

$$\frac{\left(\sum_{i=1}^{m} a_i x_i\right)^2}{\sum_{i=1}^{m}\sum_{j=1}^{n} q_{i,j}\left(x_i - x_{\bullet j}\right)^2},$$

which is equivalent to minimizing it's inverse.

If we multiply each x_i, $i = 1,\ldots,m$ by a non null constant c both the numerator, that we will represent by $g(\underline{x})$, and the denominator appear multiplied by c^2. This allows us to introduce the restriction

$\left(\sum_{i=1}^{m} a_i x_i\right)^2 = 1$ which is equivalent to having $\sum_{i=1}^{m} a_i x_i = 1$ or $\sum_{i=1}^{m} a_i x_i = -1$.

As $g(\underline{x})$ is an even function, $g(\underline{x}) = g(-\underline{x})$, we can assume either one of the conditions, so let us choose $\sum_{i=1}^{m} a_i x_i = 1$ and minimize the function $g(\underline{x})$ under this restriction.

In order to solve this problem of minimization it was necessary first to rewrite the function $g(\underline{x})$ in a quadratic form and nextly to incorporate the restriction.

Now taking $p_j = \sum_{i=1}^{m} q_{i,j}$, $j = 1,\dots,n$ we get

$$g(x) = \sum_{i=1}^{m}\left(\sum_{j=1}^{n}\left(q_{i,j} - \frac{q_{i,j}^2}{p_j}\right)\right)\cdot x_i^2 - \sum_{\substack{i=1\\ i\neq l}}^{m}\sum_{l=1}^{m}\left(\sum_{j=1}^{n}\frac{q_{i,j}q_{l,j}}{p_j}\right)\cdot x_i x_l$$

and putting $l_i = \sum_{j=1}^{n} q_{i,j}$ and $c_{i,l} = \sum_{j=1}^{n} \frac{q_{i,j}q_{l,j}}{p_j}$, $i,l = 1,\dots,m$ we have

$$g(\underline{x}) = \sum_{i=1}^{m} l_i x_i^2 - \sum_{i=1}^{m}\sum_{l=1}^{m} c_{il} x_i x_l$$

Nextly we incorporate the restriction, taking $x_m = \frac{1}{a_m}\left(1 - \sum_{l=1}^{m-1} a_l x_l\right)$ so the new function to be minimized will be

$$g^*(\underline{x}) = \sum_{i=1}^{m-1}\left(l_i - c_{i,i} + \frac{l_m - c_{m,m}}{a_m^2} a_i^2 + \frac{2c_{i,m}}{a_m} a_i\right) x_i^2 + \sum_{i=1}^{m-1}\sum_{\substack{l=1\\ i\neq l}}^{m-1}\left(\frac{2c_{i,m}}{a_m} a_l + \right.$$
$$\left. + \frac{l_m - c_{m,m}}{a_m^2} a_i a_l - c_{i,l}\right) x_i x_l - \sum_{i=1}^{m-1}\left(\frac{2c_{i,m}}{a_m} + 2\frac{l_m - c_{m,m}}{a_m^2} a_i\right) x_i + \frac{l_m - c_{m,m}}{a_m^2}$$

Lastly we may use Mathematica software to carry out this last minimization.

5. Application to the incidence of some diseases in European Countries

We applied both algorithms to data on the incidence of Tuberculosis (TB) and of AIDS in thirteen European countries (m=13) covering seven years (n=7), from 1997 to 2003. We used the data available in WHO/Europe's statistical databases.

We start with TB. To compare both algorithms we present in Table 1 the estimated values of $y_{i,j}$, $y^*_{i,j}$ obtained through Zigzag and Double Minimization. To lighten the presentation we only present the first and last year studied.

Table 1. TB: Comparison between the two algorithms

	1997		2003	
Country	ZigZag $\alpha+\beta(x_{ij})$	Double Minimization $\alpha_j+\beta(x_i)$	ZigZag $\alpha+\beta(x_{ij})$	Double Minimization $\alpha_j+\beta(x_i)$
Austria	-8.74355	-8.74355	-9.02797	-9.02797
Belgium	-8.89010	-8.89010	-9.17453	-9.17453
Denmark	-9.11910	-9.11910	-9.40352	-9.40352
Finland	-9.08677	-9.08677	-9.37119	-9.37119
Germany	-8.99053	-8.99053	-9.27495	-9.27495
Greece	-9.43489	-9.43489	-9.71932	-9.71932
Ireland	-9.00980	-9.00980	-9.29423	-9.29423
Italy	-9.33389	-9.33389	-9.61831	-9.61831
Luxembourg	-9.15778	-9.15778	-9.44220	-9.44220
Portugal	-7.55597	-7.55597	-7.84039	-7.84039
Spain	-8.36333	-8.36333	-8.64775	-8.64775
Sweden	-9.78557	-9.78557	-10.07000	-10.07000
UK	-9.02907	-9.02907	-9.31350	-9.31350

As we can see the values are the same and this happens for all the seven years.

In Figure 1 we present the adjusted coefficients for local factors f_i adjusted through both algorithms.

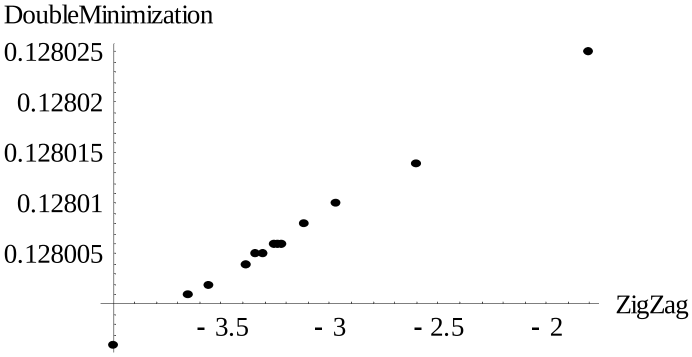

Figure 1. TB: adjusted coefficients for local factors f_i

This Figure clearly suggests a linear relation between the location effects adjusted by the two algorithms. Representing by f and f' those effects for the Zigzag and Double Minimization algorithms we adjusted the linear regression

$$f' = 0{,}128 + 0{,}000013 f \text{ with } R^2 = 0{,}998$$

which strongly validates the existence of the assumed linear relation. This is an additional reason to accept that, for this case, both algorithms led to equivalent results.

The estimated value of S and of the determination coefficient, for the logit model, are the same for both algorithms

$$\begin{cases} \hat{S} = 3440{,}38 \\ R^2 = 0{,}944 \end{cases}$$

and, as we can see, show a quite good adjustments.

In the case of the incidence of AIDS we made the same comparisons and the conclusions are the same.

The estimated values of $y_{i,j}$, $y^*_{i,j}$, are presented in Table 2. Once again the results are equal for both algorithms.

Table 2. AIDS: Comparison between the two algorithms

	1997		2003	
Country	ZigZag $\alpha+\beta(x_{ij})$	Double Minimization $\alpha_j+\beta(x_i)$	ZigZag $\alpha+\beta(x_{ij})$	Double Minimization $\alpha_j+\beta(x_i)$
Austria	-11.1303	-11.1303	-11.9147	-11.9147
Belgium	-10.9573	-10.9573	-11.7417	-11.7417
Denmark	-10.8563	-10.8563	-11.6407	-11.6407
Finland	-12.0946	-12.0946	-12.8790	-12.8790
Germany	-11.1801	-11.1801	-11.9645	-11.9645
Greece	-11.0507	-11.0507	-11.8351	-11.8351
Ireland	-11.5759	-11.5759	-12.3603	-12.3603
Italy	-9.7702	-9.7702	-10.5546	-10.5546
Luxembourg	-10.5483	-10.5483	-11.3327	-11.3327
Portugal	-8.8577	-8.8577	-9.6421	-9.6421
Spain	-9.1269	-9.1269	-9.9113	-9.9113
Sweden	-11.4857	-11.4857	-12.2701	-12.2701
UK	-10.7320	-10.7320	-11.5164	-11.5164

As for TB we have a well defined linear relation between the adjusted coefficients, obtained through both algorithms, for the location effects, as we can see in Figure 2.

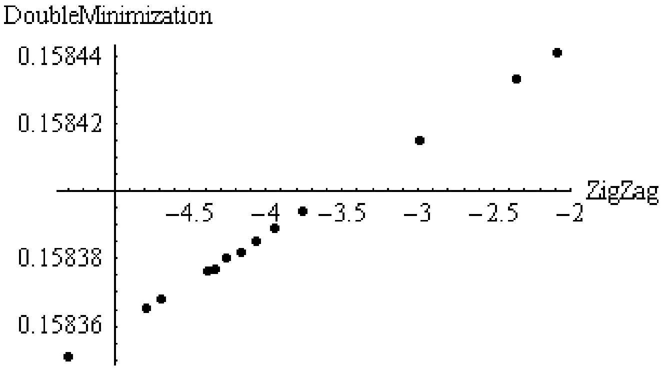

Figure 2. AIDS: adjusted coefficients for local factors f_i

We now got, $f' = 0{,}158 + 0{,}0000279 f$ with $R^2 = 0{,}999854$, thus also for AIDS the existence of the linear relation between f and f' is strongly validate. The adjustments of the logit model through both algorithms are very good

$$\begin{cases} \hat{S} = 1172{,}45 \\ R^2 = 0{,}904 \end{cases}$$

Thus both algorithms display a remarkable agreement. Moreover the Zigzag algorithm only required two iterations in either application. So, once again, this algorithm performed very well.

REFERENCES

Hosmer D.W., Lemeshow S. (2000): Applied Logistic Regression, Second Edition, Wiley Series in Probability and Statistics.

Mexia J.T., Pereira D., Baeta J. (1999): L_2 Environmental Indexes. Listy Biometryczne-Biometrical Letters 36 (2): 137-143.

Nunes S., Mexia J.T., Minder C. (2004): Logit Model for Tuberculosis in Europe (1995-2000). In: Proceedings of the 19th International Workshop on Statistical Modelling, Biggeri, A., Dreassi, E., Lagazio, C. and Marchi, M. (Eds.). Florence, Italy: 465-469.

Nunes S., Mexia J.T., Minder C. (2004): Logit Model for Tuberculosis Incidence in Europe (1995-2000). Analysis by Sex and Age Group. In: Colloquium Biometryczne 34: 147-159.

WHO/Europe - Data, Statistical data: European health for all database (HFA-DB): http://www.euro.who.int/hfadb

WHO/Europe - Data, Statistical data: Centralized information system for infectious diseases (CISID): http://data.euro.who.int/cisid