SORT 29 (2) July-December 2005, 249-268

# Factorial experimental designs and generalized linear models

S. Dossou-Gbété and W. Tinsson

*Université de Pau et des Pays de l'Adour*

**Abstract**

This paper deals with experimental designs adapted to a generalized linear model. We introduce a special link function for which the orthogonality of design matrix obtained under Gaussian assumption is preserved. We investigate by simulation some of its properties.

*MSC:* 62J12, 62K15

*Keywords:* generalized linear model, exponential family, Fisher-Scoring algorithm, factorial designs, regular fraction.

## 1 Introduction

Experimental designs are usually used in a linear context, *i.e.* assuming that the mean response can be correctly fitted by a linear model (polynomial of degree one or two in most cases). This assumption is often associated with the normality of the observed responses. Some classical and efficient experimental designs are then well known in this context (see the books of Box and Draper (1987) or Khuri and Cornell (1996)). However, it is clear that these linear assumptions are inadequate for some applications.

Many books and papers deal with the question of relaxing the linear model and the gaussian model framework (see, for example, chapter 10 of the book of Khuri and Cornell (1996) for a synthesis). But there are two main difficulties with this approach. First, the choice of a good nonlinear model is not always easy. Second, assuming the

*Address for correspondence:* S. Dossou-Gbété and W. Tinsson Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques Appliquées, avenue de l'Université-64000 Pau-France.

Received: March 2004
Accepted: November 2005

nonlinear model is given, using a classical design (factorial, central composite, *etc.*) is not in general the best choice. This fact can be problematic when industrial results are first obtained with a classical design. If a linear model turns out to be inappropriate it is then impossible in general to make new experiments because they are too expensive.

Our goal in this paper is to propose another class of solutions. These solutions have to be, on the one hand, more general than the linear case and the gaussian framework and, on the other hand, easier to improve than nonlinear modeling.

This intermediate solution consists of the choice of a generalized linear model (see, for example, McCullagh and Nelder (1989) or Green and Silverman (1994)). In other words, we assume that the image of the mean response by a given "link function" can be modelled *via* a linear relationship. Such an assumption allows us to consider any responses with a distribution in the exponential family (Bernoulli, binomial, Poisson, Gamma, *etc.*) and then we do not have the restrictions of the classical linear case. These models have been studied in order to construct D-optimal designs (see the book of Pukelsheim (1993) for the general problem of optimality). The main problem of this approach is the fact that the information matrix depends on the unknown parameters of the model. Some authors have then developed Bayesian methods (see Chaloner and Larntz (1989)) or robust designs (see Chipman and Welch (1996) or Sebastiani and Settimi (1997)) but these are available only for logistic regression. Our goal in this paper is to propose a general method of analysis with a simple information matrix, independent of the parameters of the model. When there is no prior knowledge the canonical link function is classically used for the modelization of the mean. We prove in the following that if we use the alternative choice of an appropriate link function, called the *surrogate function*, then classical factorial designs can be advantageously used.

Our paper is organized as follows. Section 2 is devoted to notations and preliminary results concerning the generalized linear model, the Fisher scoring algorithm and factorial designs. Section 3 makes a link between these methods and the choice of an experimental design. At the end we present an example of application.

## 2 Experimental designs and GLM

### *2.1 The generalized linear model*

We consider in the following a generalized linear model as it was introduced by Nelder and Wedderburn (1972). Suppose that we have $n$ observed responses $y_i$ $(i = 1, \ldots, n)$ associated with the independent random variables $Y_i$ having the same distribution, a member of exponential family. Denoting $m_i = E(Y_i)$, we then have a generalized linear model if and only if:

$$\forall\ i = 1, \ldots, n\ ,\ g(m_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

where $\mathbf{x}_i \in \mathbb{R}^r$ is the vector of independant variables, $\boldsymbol{\beta} \in \mathbb{R}^r$ is the vector of unknown parameters of the model and $g$ is the link function (assumed to be bijective and differentiable). Because $Y_i$ $(i = 1, \ldots, n)$ is a member of exponential family we have the following classe of density functions:

$$f(y_i, \theta_i, \phi) = h(y_i, \phi) \exp\left(\frac{y_i\theta_i - v(\theta_i)}{\phi}\right) \text{ with } \phi \text{ known.} \tag{1}$$

We say that $\theta_i$ is the canonical parameter of the distribution (associated with $Y_i$) and that $\phi$ is a dispersion parameter. It is usual to use the canonical link function which means that:

$$\forall\ i = 1, \ldots, n\ ,\ g(m_i) = \theta_i.$$

Recall that for every element of an exponential family we have the following relations:

$$E(Y_i) = m_i = v'(\theta_i) \text{ and } \operatorname{Var}(Y_i) = \phi v''(\theta_i)\,. \tag{2}$$

Hence we can write $\operatorname{Var}(Y_i) = V(m_i)$ with $V(m_i) = \phi m_i'(\theta_i)$.

**Example 1** Consider the common case of binary responses. Every observed response $y_i$ is then a realization of a Bernoulli distribution of parameter $p_i$ (unknown in most cases). Such a distribution belongs to the exponential family because its density satisfies relation (1) with :

$$\theta_i = \ln\frac{p_i}{1-p_i}\ ,\ v(\theta_i) = -\ln\left(1 + e^{\theta_i}\right)\ ,\ \phi_i = 1 \text{ and } h(y_i, \phi_i) = \mathbb{I}_{\{0,1\}}(y_i, \phi_i)\,.$$

For the function $V$ and the canonical link function we have $m_i = p_i$ and $\operatorname{Var}(Y_i) = p_i(1-p_i)$ so:

$$V(t) = t(1-t) \text{ and } g(t) = \ln\frac{t}{1-t}.$$

### 2.2 Estimation of the parameters

For a given generalized linear model, our problem is then to estimate the unknown parameters for the specification of the mean. Using the maximum likelihood method, our goal is then to maximize the likelihood of the sample or (equivalently) its logarithm, that is:

$$L(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i=1}^{n} \frac{y_i \theta_i - v(\theta_i)}{\phi_i} + \sum_{i=1}^{n} \ln(h(y_i, \phi_i)). \tag{3}$$

The likelihood maximization involves a nonlinear equation for which the solution is not in closed form. Nelder and Wedderburn (1972) proposed the Fisher-scoring algorithm in order to find a numerical approximation of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$. Fisher-scoring is one of the best known quasi-Newton method to solve the likelihood maximization problem (see Smyth (2002)). For the implementation of this algorithm we have to choose an initial value $\boldsymbol{\beta}^{(0)}$ for the parameters of the model and then to apply iteratively the relation:

$$\forall\, k \in \mathbb{N}^*, \quad \boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left(\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X}\right)^{-1} \mathbf{q}^{(k)} \tag{4}$$

where $\boldsymbol{\beta}^{(k)} \in \mathbb{R}^r$ is an approximation of the solution at iteration $k$, $\mathbf{X}$ is the model matrix (with $n$ rows and $r$ columns), $\mathbf{W}^{(k)}$ and $\mathbf{q}^{(k)}$ depend on the vector $\boldsymbol{\beta}$ at iteration $k$ as follows:

$$\mathbf{W}^{(k)} = diag(\omega_i, i = 1, \ldots, n) \text{ with } \omega_i = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\partial m_i}{\partial \eta_i}\right)^2,$$
$$\eta_i = g(m_i) \text{ and } \mathbf{q}^{(k)} \text{ as } \frac{\partial L(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \beta_j} \text{ for } j\text{-th element}(j = 1, \ldots, r).$$

Note that the matrix $\mathbf{W}^{(k)}$ has to be computed at every iteration because it depends on $m_i$ and $m_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ depends on the value of the approximation of the solution at iteration $k$ (vector $\boldsymbol{\beta}^{(k)}$).

**Remark 1** It is also possible to find a vector $\mathbf{z}^{(k)}$ such that relation (4) becomes:

$$\boldsymbol{\beta}^{(k+1)} = \left(\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{z}^{(k)}.$$

In other words, the Fisher scoring algorithm is also an iteratively reweighted least squares method.

### 2.3 Factorial designs

We assume now that every variable is coded in such a way that its values always belong to the interval $[-1, 1]$ (this can be done in a very simple way by using a linear transformation, see chapter 2 of Khuri and Cornell (1996)). A complete factorial design, for $m$ factors, is then constituted by all the vertices of the cube $[-1, 1]^m$. Nevertheless,

using such designs is not possible when the number of factors $m$ becomes high (because of the $2^m$ experimental units). So we also consider in the following some regular fractions of these factorial designs (see Box and Hunter, 1961*a*, *b*). In other words, we are now working with configurations given by:

1) $2^{m-q}$ vertices of the cube $[-1, 1]^m$,
2) $n_0$ central replications of the experimental domain.

**Example 2** For $m = 3$ factors we can consider first the complete factorial design associated to the design matrix $\mathbf{D}_C$ (*i.e.* the $n \times m$ matrix with row $i$ made up from the $m$ coordinates of the $i$-th design point). Another choice is given by a regular fraction associated with the matrix $\mathbf{D}_F$ (with $n_0 = 1$ central point in our case):

$$\mathbf{D}_C = \begin{bmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \text{ and } \mathbf{D}_F = \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

This fraction is obtained by keeping the experimental units such that $x_1 x_2 x_3 = +1$ where $x_i$ denotes the i-th coordinate of each design point of the factorial part. We say in a classic way that this regular fraction is generated by the relation $123 = \mathbb{I}$ where 1, 2 and 3 are the three different columns of the design matrix and 123 is $1 \odot 2 \odot 3$ with $\odot$ the Hadamard product operator (also called elementwise product).

In the framework of linear models these factorial designs can be used in order to fit linear ($L$) or interaction ($I$) models such that:

$$(L) \ , \ \forall\, i = 1, \ldots, n \ , m_i = E\,(Y_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \ ,$$

$$(I) \ , \ \forall\, i = 1, \ldots, n \ , m_i = E\,(Y_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \sum\sum_{j<l} \beta_{jl} x_{ij} x_{il}.$$

We denote in the following by $\mathbf{D}$ the design matrix, by $\mathbf{D}_j$ $(1 \leq j \leq m)$ the $j$-th column of this matrix and we put $\mathbf{Q}_{jl} = \mathbf{D}_j \odot \mathbf{D}_l$ $(1 \leq j < l \leq m)$. The model matrix is then given by:

$$\mathbf{X} = [\mathbb{I}_n \mid \mathbf{D}_1 \ldots \mathbf{D}_m] \text{ for the model } (L),$$
$$\mathbf{X} = [\mathbb{I}_n \mid \mathbf{D}_1 \ldots \mathbf{D}_m \mid \mathbf{Q}_{12} \ldots \mathbf{Q}_{(m-1)m}] \text{ for the model } (I).$$

It is well known that the matrix model $\mathbf{X}$ is of full rank (*i.e.* $\mathbf{X}^T\mathbf{X}$ is regular) for the two models when the factorial design is complete. In the case of a regular fraction then it will be of resolution at least III for the model $(L)$ and at least V for the model $(I)$ in order the obtain a full rank matrix $\mathbf{X}$ (see Box and Hunter, 1961*a*, *b*). When a factorial regular design is used we have also an orthogonal configuration such that (for the two models):

$$\mathbf{X}^T\mathbf{X} = \text{diag}\left(2^{m-q} + n_0, 2^{m-q}, \cdots, 2^{m-q}\right).$$

**Example 3** (continuation) The complete factorial design associated with matrix $\mathbf{D}_C$ can be used to fit model $(L)$ or $(I)$. For the regular fraction associated with the matrix $\mathbf{D}_F$ it is a fraction of resolution III because it has only one generator (123) and this generator is a word of length 3. So such a fraction can be used to fit model $(L)$ but is not able to fit model $(I)$. In other words the following model matrix $\mathbf{X}_F^1$ for model $(L)$ is of full rank but $\mathbf{X}_F^2$ for model $(I)$ is not (because, for example, columns 2 and 7 are the same):

$$\mathbf{X}_F^1 = \left[\begin{array}{c|ccc} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{array}\right], \mathbf{X}_F^2 = \left[\begin{array}{c|ccc|ccc} 1 & 1 & -1 & -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

## 3 The surrogate link function

### 3.1 Modified Fisher-scoring method

Our goal is now to simplify the algorithm of Fisher scoring by dropping out the diagonal weighting matrix $\mathbf{W}$. This can be done by a judicious choice of the link function. In fact our objective is:

$$\mathbf{W} = I_d \Leftrightarrow \forall\, i = 1, \ldots, n\,, \frac{1}{\text{Var}\,(Y_i)}\left(\frac{\partial m_i}{\partial \eta_i}\right)^2 = 1. \tag{5}$$

But we know, from relation (2), that $\text{Var}\,(Y_i) = V\,(m_i)$. Then $m_i = g^{-1}\,(\eta_i)$ implies that:

$$(5) \Leftrightarrow \frac{\partial m_i}{\partial \eta_i} = \sqrt{V\,(m_i)} \Longleftrightarrow \frac{1}{g'(m_i)} = \sqrt{V\,(m_i)}.$$

Our proposal relies on the following lemma:

**Lemma 1** *The matrix* $\mathbf{W}$ *is the identity matrix if and only if the link function* $g$ *satisfies:*

$$\forall\, i = 1, \ldots, n\,,\; g'(m_i) = V^{-1/2}(m_i)\,.$$

*Such a function is then called the surrogate link function.*

Table 1 gives, for some exponential families of distributions, the surrogate link functions (depending on $t$) verifying the differential equations of lemma 1 (with the additive constant chosen to be zero). We also recall in this table the associated canonical link functions.

***Table 1***: *Surrogate link function for different distributions.*

| *Distribution of* $Y_i$ | *Function* $V$ | *Surrogate link fn.* | *Canonical link fn.* |
|---|---|---|---|
| Bernoulli $(p)$ | $t(1-t)$ | $\arcsin(2t-1)$ | $\ln\left(\frac{t}{1-t}\right)$ |
| Binomial $\mathcal{B}(n,p)$ | $t\left(1-\frac{t}{n}\right)$ | $\sqrt{n}\arcsin\left(\frac{2t}{n}-1\right)$ | $\ln\left(\frac{t}{n-t}\right)$ |
| Neg. Bin. $(n,p)$ | $t\left(\frac{t}{n}+1\right)$ | $\sqrt{n}\,\mathrm{arccosh}\left(\frac{2t}{n}+1\right)$ | $\ln\left(\frac{t}{n+t}\right)$ |
| Poisson $\mathcal{P}(\lambda)$ | $t$ | $2\sqrt{t}$ | $\ln t$ |
| Gamma $\mathcal{G}(a,p)$ | $\frac{t^2}{p}$ | $\sqrt{p}\ln t$ | $\frac{p}{t}$ |

**Remark 2** We have seen in Section 2.2 that the Fisher-scoring algorithm is in fact an iteratively reweighted least squares method. Then, the use of the surrogate link function allows us to have an iteratively unweighted least squares method.

The algorithm of Fisher scoring needs also the use of a vector $\mathbf{q}$ with $j$-th element $\partial L(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\phi})/\partial\beta_j$ for $j = 1, \ldots, r$ (see section 2.2). We have the following relation, using the chain rule:

$$\frac{\partial L}{\partial \beta_j} = \frac{\partial L}{\partial \theta_i}\frac{\partial \theta_i}{\partial m_i}\frac{\partial m_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Then we obtain immediately for the likelihood of every sample of the exponential family (see formula (3)):

$$\forall\; j = 1, \ldots, r\,,\; \frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - m_i)}{\mathrm{Var}\, Y_i}\frac{\partial m_i}{\partial \eta_i}\,[\mathbf{X}]_{ij}$$

where $[\mathbf{X}]_{ij}$ is the element of row $i$ and column $j$ of the matrix model $\mathbf{X}$. This general relation can be simplified in our case because we have:

$$\eta_i = g(m_i) \text{ with } g'(m_i) = V^{-1/2}(m_i) \text{ so } \frac{\partial m_i}{\partial \eta_i} = \sqrt{V(m_i)}.$$

Thus, we can state the following lemma:

**Lemma 2** *If the link function is the surrogate link function the vector* $\mathbf{q}$ *is then defined by:*

$$\forall\ j = 1, \ldots, r\ ,\ \frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} [\mathbf{X}]_{ij}\, y_i^* \text{ with } y_i^* = \frac{y_i - m_i}{\sqrt{V(m_i)}}.$$

We see from lemma 2 that the vector $\mathbf{q}$ has a very simple expression when the surrogate link function is used. It needs only the observations $y_i^*$ in their standardized and centred form.

**Example 4** (continuation) Consider a random binary phenomenon such that every observed response $y_i$ is a realization of a Bernoulli distribution with parameter $p_i$ (unknown). Here we make the assumption that this phenomenon depends on three factors and the true response is given by:

$$\forall\ i = 1, \ldots, n\ ,\ p_i = 0.2x_{i1} - 0.1x_{i2} - 0.1x_{i3} + 0.6$$

where $x_{i1}$, $x_{i2}$ and $x_{i3}$ are the coded levels for the three factors. In other words, we assume that the probabilities associated with each Bernoulli distribution can be correctly fitted by a Taylor series of order one in the experimental domain. We also assume that the effects of factors 2 and 3 are opposite (and lower) to the effect of the factor 1 on the response. In order to make a modelization of this phenomenon using the surrogate link function (such that $g(t) = \arcsin(2t - 1)$ in our case) we can consider the following model (with $m_i = \mathrm{E}(Y_i) = p_i$):

$$\forall\ i = 1, \ldots, n\ ,\ \arcsin(2m_i - 1) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

### *3.2 Application to factorial designs*

Consider a random phenomenon of $m$ factors that may be checked by the experimenter. We have seen that the choice of the surrogate link function allows us to put the matrix $\mathbf{X}^T\mathbf{X}$ in place of the initial matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ in the algorithm of Fisher scoring. The optimal situation is then reached when a complete factorial design or a well chosen regular

fraction is used, because we have seen in Section 2.3 that $\mathbf{X}^T\mathbf{X}$ is then a diagonal matrix (*i.e.* the design is orthogonal). Now we consider in the following the two non-linear models given below:

$$(L^*)\ ,\ \forall\ i = 1, \ldots, n\ ,\ g\left(m_i\right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}\ ,$$

$$(I^*)\ ,\ \forall\ i = 1, \ldots, n\ ,\ g\left(m_i\right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \sum\sum_{j<l} \beta_{jl} x_{ij} x_{il},$$

Models $(L^*)$ and $(I^*)$ are then two generalized linear models with a polynomial linear part of degree one for $(L^*)$ and of degree two with interactions for $(I^*)$. Using relation (4) and lemmas 1 and 2 we can state the following simplified iterative treatment when factorial designs are used:

**Proposition 3** *Consider the model* $(L^*)$ *or* $(I^*)$ *used with the surrogate link function. For a complete factorial design or a regular fraction of resolution at least III, the Fisher scoring algorithm is given for the model* $(L^*)$ *by:*

1) $\beta_0^{(k+1)} = \beta_0^{(k)} + \dfrac{1}{2^{m-q} + n_0} \displaystyle\sum_{i=1}^{n} y_i^*,$

2) $\forall\ j = 1, \ldots, m\ ,\ \beta_j^{(k+1)} = \beta_j^{(k)} + \dfrac{1}{2^{m-q}} \displaystyle\sum_{i=1}^{n} x_{ij} y_i^*.$

*For a complete factorial design or a regular fraction of resolution at least V, the algorithm of Fisher scoring for the model* $(I^*)$ *verifies, in addition to the two previous relations:*

3) $\forall\ j, l = 1, \ldots, m \text{ with } j < l\ ,\ \beta_{jl}^{(k+1)} = \beta_{jl}^{(k)} + \dfrac{1}{2^{m-q}} \displaystyle\sum_{i=1}^{n} x_{ij} x_{il} y_i^*.$

The implementation of the Fisher scoring algorithm is then very simple in our case because we only have to apply iteratively results from this proposition and no use of matrix calculus is needed (in particular we do not have to invert any matrix). Note also that factorial designs have only two levels, so the values for the coded variables $x_{ij}$ are only $-1$, 0 (if at least one central point is used) or 1. This algorithm has to be initialized by judicious values for $\beta^{(0)}$. This can be done, for example, by a classic linear regression on the transformed response (*i.e.* on the $g\left(y_i\right)$ with $g$ surrogate link function in place of the $y_i$). It can also be stopped using different criteria: when the likelihood seems to be constant (*i.e.* when $|L_{\max}^{(k+1)} - L_{\max}^{(k)}| < \varepsilon$ with $\varepsilon$ small positive) or when the estimated parameters seem to be constant (*i.e.* when $\|\beta^{(k)} - \beta^{(k+1)}\| < \varepsilon$ where $\|.\|$ is a chosen norm) for example.

**Example 5** (continuation) For the binary responses we assume that the experimenter has conducted the experiment according to a complete factorial design with two centre

points (the low number of factors allow us to consider the complete design in this case). We have then a total of 10 trials given in Table 2 with the probabilities $p_i$ associated for each experimental unit (column $p_i$) and simulated results for the different responses (column $y_i$).

***Table 2***: *Results for the complete factorial design.*

| Trial | Fac. 1 | Fac.2 | Fac. 3 | $p_i$ | $\mathbf{y}_i$ | $\hat{p}_i$ | $\hat{\mathbf{y}}_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0.60 | **1** | 0.54 *(0.75)* | **1** *(1)* |
| 2 | −1 | 1 | 1 | 0.20 | **0** | 0.27 *(0.00)* | **0** *(0)* |
| 3 | 1 | −1 | 1 | 0.80 | **1** | 1.00 *(1.00)* | **1** *(1)* |
| 4 | 1 | 1 | −1 | 0.80 | **1** | 1.00 *(1.00)* | **1** *(1)* |
| 5 | −1 | −1 | 1 | 0.30 | **0** | 0.03 *(0.00)* | **0** *(0)* |
| 6 | −1 | 1 | −1 | 0.30 | **0** | 0.03 *(0.00)* | **0** *(0)* |
| 7 | 1 | −1 | −1 | 1.00 | **1** | 0.59 *(1.00)* | **1** *(1)* |
| 8 | −1 | −1 | −1 | 0.60 | **1** | 0.60 *(0.75)* | **1** *(1)* |
| 9 | 0 | 0 | 0 | 0.60 | **1** | 0.57 *(0.75)* | **1** *(1)* |
| 10 | 0 | 0 | 0 | 0.60 | **0** | 0.57 *(0.75)* | **1** *(1)* |

If we have no information concerning the choice for the initial values of the algorithm, we can take, for example:

$$\beta_0^{(0)} = 1\ , \beta_1^{(0)} = \beta_2^{(0)} = \beta_3^{(0)} = 0.$$

Then the iterative treatment of Proposition 3 leads us very quickly (in two iterations) to the maximum likelihood solution:

$$\hat{\beta}_0 = 0.143\ , \hat{\beta}_1 = 1.376\ ,\ \hat{\beta}_2 = -0.719 \text{ and } \hat{\beta}_3 = -0.719.$$

In other words, the best fitted model satisfies ($\forall\ x_1, x_2 \in [-1, 1]$):

$$\hat{p}\,(x_1, x_2) = \frac{\sin\left(0.143 + 1.376x_1 - 0.719x_2 - 0.719x_3\right) + 1}{2}.$$

Predicted values of the probabilities $p_i$ are given in Table 2 (column $\hat{p}_i$) with the predicted responses (column $\hat{y}_i$), that is the values of $\hat{p}_i$ rounded to the nearest integer. We also present, in brackets, results obtained by the classical analysis with the canonical link function (these results come from the SAS software). We observe the good global quality of the results since observed responses $y_i$ and predicted responses $\hat{y}_i$ are always the same (except, of course, for the two last trials where it is impossible to predict at once 0 and 1). If we consider probabilities $p_i$ we note, on the one hand, that predictions are very good for half of the experiments (*i.e.* trials 1, 2, 8, 9 and 10). On the other hand, these results are not so good for trials 3 and 4 and they are bad for trials 5, 6 and 7. These problems of prediction are principally due to the small number of trials, and also to the nature of the responses which give poor information because we have only two levels.

We can finally note that the adjusted model allows us to find again the correct effect of each factor (*i.e.* factor 1 has a preponderant effect on the response and factors 2 and 3 have equal effects, opposite to factor 1).

### 3.3 Dispersion of the estimations

We know (see Green and Silverman (1994)) that asymptotically the maximum likelihood estimator of $\boldsymbol{\beta}$ has a Gaussian distribution and a dispersion given by:

$$\mathrm{Var}\hat{\boldsymbol{\beta}} = \phi\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}.$$

If $\phi$ is unknown then it can be estimated by means of Pearson statistics. This result is very interesting in our case because we know that $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is a diagonal matrix and the diagonal elements are given in the last subsection. So we have the following proposition:

**Proposition 4** *Consider the model* $(L^*)$ *or* $(I^*)$ *used with the surrogate link function and a complete factorial design or an appropriate regular fraction (of resolution at least III for* $(L^*)$ *and at least V for* $(I^*)$*). The maximum likelihood estimator* $\hat{\beta}$ *satisfies asymptotically the following properties:*

*1) its components are non-correlated,*

*2) its dispersion is given by:*

$$\mathrm{Var}\,\hat{\beta}_0 = \frac{\phi}{2^{m-q}+n_0} \text{ and } \forall\ j,l = 1,\ldots,m\ ,\ j<l\ ,\ \mathrm{Var}\,\hat{\beta}_j = \mathrm{Var}\,\hat{\beta}_{jl} = \frac{\phi}{2^{m-q}}.$$

**Remark 3** The dispersion parameter $\phi$ is needed in order to obtain these different dispersions. This is not a serious problem in practice because $\phi$ has often a simple form (for example, $\phi = 1$ for a binomial distribution, a Poisson distribution and a negative-binomial distribution).

### 3.4 Considering submodels

For $m$ quantitative factors, models $(L^*)$ and $(I^*)$ are not often the best choice because some linear effects or interactions may be sometimes removed. Then it is preferable to use a submodel. Propositions 3 and 4 follow from the orthogonal properties of the model matrix $\mathbf{X}$ when the model is complete. These results are then *a fortiori* true in the case of a submodel. The main problem for an experimenter is the validation of such a submodel. In other words, is the chosen submodel really better than the complete model? To answer to this question it is usually advised to use the notion of deviance (see Green

and Silverman [6]). For a given model and a submodel, the deviance is defined as:

$$D = -2\left[L_{\max}\left(\text{submodel}\right) - L_{\max}\left(\text{model}\right)\right] \tag{6}$$

where $L_{\max}(.)$ is the maximal value of the likelihood for the (sub)model (*i.e.* $L(\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator). The choice of the submodel is then a good alternative when the deviance $D$ is close to zero. In order to quantify this notion we usually use the following rule: when the model has $r$ unknown parameters and the submodel has $r' < r$ unknown parameters, the submodel is then a better one if and only if:

$$D < \chi^2_{p-p',0.05}$$

with $\chi^2_{p-p',0.05}$ the upper 5% point of a $\chi^2$ distribution with $(r - r')$ degrees of freedom.

## 4 Application to the geometric distribution

### 4.1 Utilization of a full model

We consider in this part responses with a binomial negative distribution and, more precisely, the particular case of the geometric distribution. It is, once again, a very classical situation and such a distribution is in the exponential family because its density satisfies relation (1) with:

$$\theta_i = \ln\left(1 - p_i\right)\ ,\ v\left(\theta_i\right) = -\ln\left(1 - e^{\theta_i}\right)\ ,\ \phi = 1 \text{ and } h\left(y_i, \phi\right) = \mathbb{I}_{\mathbb{N}}\left(y_i, \phi\right).$$

We can illustrate such model by considering experiments made in order to test the tensile strength of ropes. The experimenter makes identical tractions and the response is then the number of tractions endured by the rope before breaking. We assume in the following that the tensile strength of the rope depends mainly on five concentrations of chemicals (called now factors 1, 2, 3, 4 and 5).

From section 2.3 we consider the surrogate link function $g(t) = \operatorname{arccosh}(2t + 1)$ and we can use the interaction model (with $m_i = E(Y_i) = (1 - p_i)/p_i$):

$$\forall\ i = 1, \ldots, n\ ,\ \ \operatorname{arccosh}\left(2m_i + 1\right) = \beta_0 + \sum_{j=1}^{5} \beta_j x_{ij} + \sum\sum_{j<k} \beta_{jk} x_{ij} x_{ik}.$$

We consider in the following the experimental design obtained by the regular fraction of the factorial design such that $\mathbb{I}_{16} = 12345$. With the addition of three central replications, we have then a total of 19 experiments given in Table 3 (with 1 denoted by + and −1 by

−). Responses given in this table are obtained by simulation of geometric distributions with $p_i$ parameters such that :

$$\forall\, i = 1, \ldots, 19\ , \quad \operatorname{arccosh}(2m_i + 1) = \\ x_{i1} + 0.5x_{i2} + 0.5x_{i3} + 0.5x_{i4} - 0.5x_{i5} + x_{i1}x_{i2} + 0.5x_{i1}x_{i4} + 2. \tag{7}$$

In other words, we make two important assumptions in this part. First, we assume that there are only two interaction effects in this phenomenon and they are associated with the pair of factors $\{1, 2\}$ and $\{1, 4\}$. Secondly, we assume here that we are in "optimal" conditions because the true model uses the surrogate link function (simulations with another link function will be used later).
Now we can implement the Fisher-scoring algorithm with a set of simulated responses (given in column $y_i$ of Table 3). Concerning the initial values, we take:

$$\beta_0^{(0)} = \operatorname{arccosh}(2\overline{y} - 1) \text{ and all the others components of } \boldsymbol{\beta}^{(0)} \text{ are zero.}$$

So, the algorithm is initiated with the best choice for a constant model.

***Table 3***: *Results for the fractional factorial design, the full model and the submodel (in brackets).*

| Exp | f 1 | f 2 | f 3 | f 4 | f 5 | $p_i$ | $y_i$ | $\hat{p}_i$ | $\hat{y}_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | + | + | + | + | + | 0.02 | **40** | 0.03 *(0.02)* | **39** *(64)* |
| 2 | − | − | + | + | + | 0.60 | **1** | 0.51 *(0.50)* | **1** *(1)* |
| 3 | − | + | − | + | + | 0.94 | **0** | 0.99 *(0.94)* | **0** *(0)* |
| 4 | − | + | + | − | + | 0.94 | **0** | 0.99 *(0.99)* | **0** *(0)* |
| 5 | − | + | + | + | − | 0.60 | **1** | 0.51 *(0.63)* | **1** *(1)* |
| 6 | + | − | − | + | + | 0.60 | **2** | 0.34 *(0.52)* | **2** *(1)* |
| 7 | + | − | + | − | + | 0.94 | **0** | 0.99 *(0.97)* | **0** *(0)* |
| 8 | + | − | + | + | − | 0.11 | **7** | 0.13 *(0.10)* | **7** *(9)* |
| 9 | + | + | − | − | + | 0.28 | **4** | 0.21 *(0.29)* | **4** *(2)* |
| 10 | + | + | − | + | − | 0.02 | **72** | 0.01 *(0.01)* | **70** *(76)* |
| 11 | + | + | + | − | − | 0.04 | **20** | 0.05 *(0.05)* | **19** *(19)* |
| 12 | − | − | − | − | + | 0.94 | **0** | 0.99 *(0.89)* | **0** *(0)* |
| 13 | − | − | − | + | − | 0.60 | **1** | 0.51 *(0.44)* | **1** *(1)* |
| 14 | − | − | + | − | − | 0.28 | **5** | 0.17 *(0.26)* | **5** *(3)* |
| 15 | − | + | − | − | − | 0.94 | **0** | 0.99 *(0.97)* | **0** *(0)* |
| 16 | + | − | − | − | − | 0.94 | **0** | 0.99 *(0.64)* | **0** *(0)* |
| 17 | 0 | 0 | 0 | 0 | 0 | 0.42 | **0** | 0.44 *(0.41)* | **1** *(1)* |
| 18 | 0 | 0 | 0 | 0 | 0 | 0.42 | **2** | 0.44 *(0.41)* | **1** *(1)* |
| 19 | 0 | 0 | 0 | 0 | 0 | 0.42 | **1** | 0.44 *(0.41)* | **1** *(1)* |

The iterations continue until the likelihood increases by only a small amount (*i.e.* until $L_{\max}^{(k+1)} - L_{\max}^{(k)} < \varepsilon$ with $\varepsilon = 0.001$). Then, we obtain the following estimates after 10 iterations:

$$\begin{array}{lllll} \hat{\beta}_0 = 1.946 & \hat{\beta}_1 = 0.971 & \hat{\beta}_{12} = 1.097 & \hat{\beta}_{23} = -0.094 & \hat{\beta}_{34} = -0.145 \\ & \hat{\beta}_2 = 0.469 & \hat{\beta}_{13} = -0.186 & \hat{\beta}_{24} = -0.067 & \hat{\beta}_{35} = -0.233 \\ & \hat{\beta}_3 = 0.441 & \hat{\beta}_{14} = 0.435 & \hat{\beta}_{25} = 0.021 & \hat{\beta}_{45} = 0.072 \\ & \hat{\beta}_4 = 0.731 & \hat{\beta}_{15} = 0.113 & & \\ & \hat{\beta}_5 = -0.514 & & & \end{array}$$

The predicted probabilities $\hat{p}_i$ and the predicted mean responses (i.e. the rounded values to the nearest integer of $(1 - \hat{p}_i)/\hat{p}_i$) are then reported in Table 3 (results in brackets will be discussed in the next subsection). We note the global good fit of the model: observed responses and predicted responses are always very close. The maximum likelihood associated with this model is equal to −32.523.

### 4.2 Utilization of a submodel

Our goal in this part is to find a good submodel of the previous full polynomial model. Again we find that the submodel containing only the interactions $x_1x_2$ and $x_1x_4$ is interesting because it is associated with a maximum likelihood equal to −33.531. In other words, the deviance between the full model (with $r = 16$ parameters) and this submodel (with $r' = 8$ parameters) is:

$$D = -2\,(-33.531 + 32.523) = 2.016.$$

But we have $\chi^2_{8,0.05} = 15.51$ so results from Section 3.4 show us that this submodel is a good alternative to the full model. The Fisher-scoring algorithm leads us (after 8 iterations and until $L_{\max}^{(k+1)} - L_{\max}^{(k)} < \varepsilon$ with $\varepsilon = 0.001$) to the following estimates:

$$\begin{array}{llll} \hat{\beta}_0 = 2.039 & \hat{\beta}_1 = 0.987 & \hat{\beta}_4 = 0.612 & \hat{\beta}_{12} = 1.095 \\ & \hat{\beta}_2 = 0.396 & \hat{\beta}_5 = -0.518 & \hat{\beta}_{14} = 0.507 \\ & \hat{\beta}_3 = 0.429 & & \end{array}$$

In conclusion, the best fitted model is then given by the following formula (for every $x = (x_1, x_2, x_3, x_4, x_5) \in [-1, 1]^5$) :

$$\hat{p}\,(x) = \frac{2}{\cosh\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_{12} x_1 x_2 + \hat{\beta}_{14} x_1 x_4\right) + 1}$$

Results concerning this submodel are given in brackets in Table 3. Once again, we note the good quality of predicted probabilities and estimated responses.

### 4.3 Quality of the parameter estimations

Results from Sections 5.1 and 5.2 are obtained with only one simulation of the responses $y_i$ ($i = 1, \ldots, 19$). So it is natural to perform now a large number of simulations in order to evaluate the global quality of this method. Table 4 presents, for each parameter of the model, the basic statistical results (mean and dispersion) for 1000 simulations of the geometric distribution.

***Table 4***: *Simulation results.*

| *Param.* | *Mean* | *Variance* | *Param.* | *Mean* | *Variance* |
|---|---|---|---|---|---|
| $\hat{\beta}_0$ | 1.712 | 0.074 | $\hat{\beta}_4$ | 0.532 | 0.100 |
| $\hat{\beta}_1$ | 0.967 | 0.095 | $\hat{\beta}_5$ | −0.457 | 0.099 |
| $\hat{\beta}_2$ | 0.556 | 0.085 | $\hat{\beta}_{12}$ | 0.981 | 0.089 |
| $\hat{\beta}_3$ | 0.470 | 0.095 | $\hat{\beta}_{14}$ | 0.485 | 0.096 |

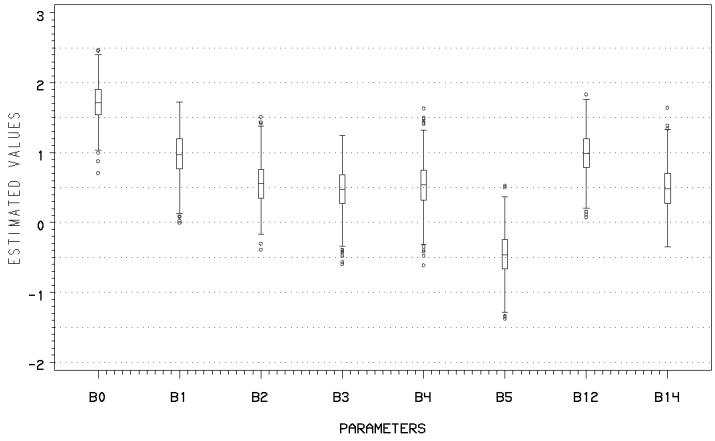


***Figure 1***: *Boxplots of values of the estimated parameters; the length of the whiskers is 1.5 times the interquartile range.*

A graphical representation of these results, using boxplots, is also given (Figure 1). We deduce from Table 4 and Figure 1 that this method of estimation is adequate concerning the stability of the estimated parameters (i.e. only a very few of these parameters are outside the whiskers). We have also a good convergence speed of this iterative method because, for the 1000 simulations, the algorithm needs an average of 8.7 iterations in order to converge (less than 10 iterations are needed in 94 % of the cases and the maximum does not exceed 20). Concerning the linear and interaction effects we note the good quality of the estimated values, with mean and median very close from the theoretical values of the model of Section 5.1. The only imprecision concerns the general mean effect $\beta_0$ which is underestimated.

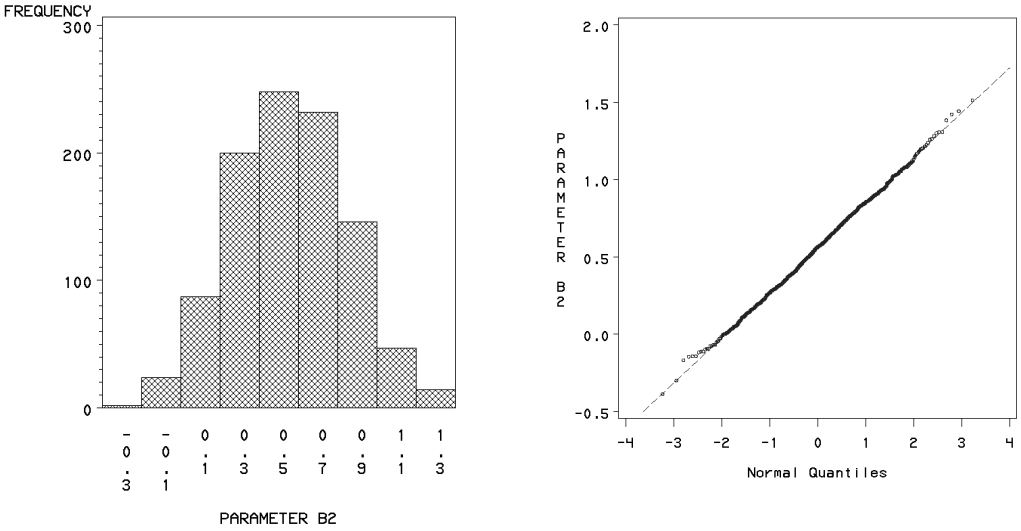


***Figure 2***: *Estimation of $\beta_2$ (histogram and QQ-plot).*

Another important problem concerns the validity of Proposition 4. The goal in statistical planning is to reduce the number of experiments so we must be very careful with asymptotic results. Nevertheless, some properties of this proposition are true in our case. First, we find again that dispersions of linear and interaction effects seem to be very close whereas the dispersion of the general mean effect is smaller (because of the three central replications). Secondly, Figure 2 (the histogram and QQ-plot for the estimated values of the parameter $\beta_2$) shows us that we can assume that this parameter follows a normal distribution, and we obtain similar results in the case of the other parameters).

We have the same problem for the choice of a submodel with the deviance criterion. We know that $D$ follows asymptotically a $\chi^2_{r-r'}$ distribution but is this result true for our 19 experiments ? We have computed, for each simulation, the deviance between the full model and the chosen submodel with only interactions $x_1x_2$ and $x_1x_4$. Figure 3 gives a graphical representation (the histogram and QQ-plot) for these deviances. The line of the QQ-plot represents the best fitted $\chi^2$ distribution and we find that it has 6 degrees of freedom (i.e. it is a gamma distribution with parameters $1/2$ and 3). In conclusion, we note that the deviance is close to a $\chi^2$ distribution but we have to be careful because the observed degrees of freedom (6) are smaller than the theoretical ones ($r - r' = 8$). This fact implies that theoretical results lead us to reject the submodel when $D > \chi^2_{8,0.05} = 15.51$ but it seems more adequate to reject it as soon as $D > \chi^2_{6,0.05} = 12.59$. Note that it has a weak influence for the validation of the submodel because, for our 1000 simulations, we have only 16 values of $D$ in the interval $[12.59, 15.51]$.

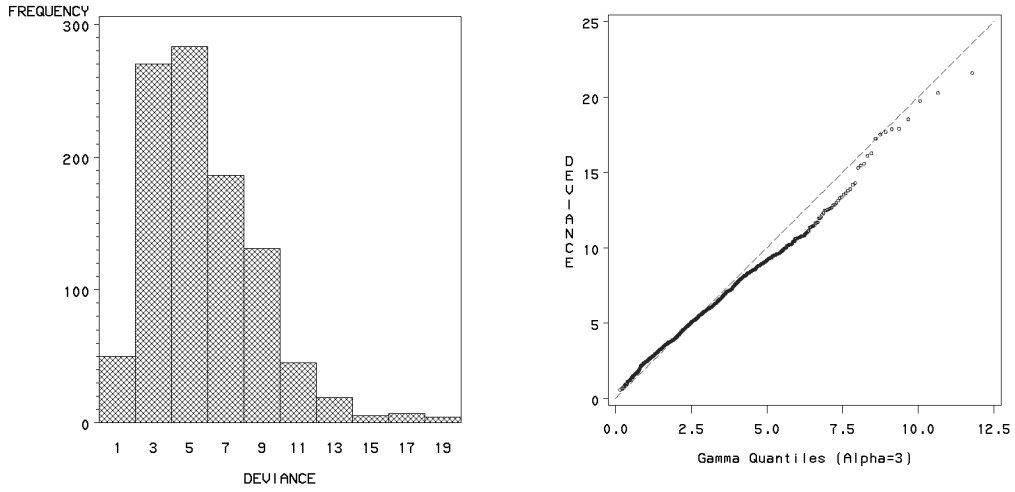


***Figure 3***: *Deviance (histogram and QQ-plot).*

### *4.4 Comparison with the classical method*

The previous Sections 5.1, 5.2 and 5.3 use simulated responses given by the model (7) (i.e. responses obtained with the surrogate link function). Now we are going to use other simulations in order to compare our method to a classical one. For the classical method we use the GENMOD procedure of the SAS program. In the case of a geometric distribution this procedure uses by default the logarithm function for the link. So, responses are now obtained by simulation of a geometric distribution with probabilites given by (with $m_i = E(Y_i) = (1 - p_i)/p_i$):

$$\begin{gathered} \forall\, i = 1, \ldots, 19\,, \quad \ln(m_i) = \\ x_{i1} + 0.5x_{i2} + 0.5x_{i3} + 0.5x_{i4} - 0.5x_{i5} + x_{i1}x_{i2} + 0.5x_{i1}x_{i4} + 2. \end{gathered} \tag{8}$$

This model is then the optimal choice concerning the classical method because it uses the same link function. The values of the $p_i$ are given for each point of the design in Table 5. For each of these parameters 1000 simulations have been made and the means of the estimated values for the $p_i$ parameters are given in Table 5. We can note that these two methods gives very close estimations. Note also that the convergence is obtained for the surrogate link function with a mean of 8.7 iterations (and the number of iterations is always between 2 and 34) but for the classical method the SAS software stops the algorithm, in most of the cases, after 50 iterations because the convergence is not reached.

Figure 4 and 5 allow us to compare these two methods with a graphical representation using boxplots of the 17 estimated probabilities associated with every experimental unit. Once again the two results seem to be very close. Figure 6 is a graphical representation for the estimated values of the model parameters. We note that the stability of the estimated parameters is again satisfactory (i.e. all the observed distributions are very close to a normal distribution).

***Table 5***: *Simulation results (mean values of $\hat{p}_i$) SLF: surrogate link function. CLF: classical link function.*

| Exp | $p_i$ | **SLF** $\hat{p}_i$ | CLF $\hat{p}_i$ | Exp | $p_i$ | **SLF** $\hat{p}_i$ | CLF $\hat{p}_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.06 | **0.10** | 0.11 | 10 | 0.06 | **0.10** | 0.11 |
| 2 | 0.32 | **0.42** | 0.41 | 11 | 0.10 | **0.16** | 0.16 |
| 3 | 0.56 | **0.64** | 0.61 | 12 | 0.44 | **0.51** | 0.50 |
| 4 | 0.44 | **0.54** | 0.52 | 13 | 0.32 | **0.41** | 0.40 |
| 5 | 0.32 | **0.43** | 0.42 | 14 | 0.22 | **0.31** | 0.31 |
| 6 | 0.32 | **0.39** | 0.40 | 15 | 0.44 | **0.52** | 0.51 |
| 7 | 0.44 | **0.53** | 0.50 | 16 | 0.44 | **0.52** | 0.50 |
| 8 | 0.15 | **0.22** | 0.23 | 17 | 0.27 | **0.32** | 0.35 |
| 9 | 0.22 | **0.29** | 0.30 | | | | |

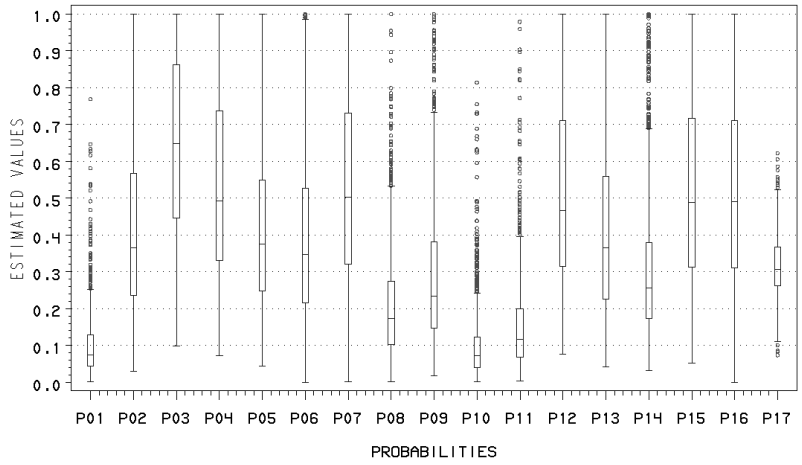


***Figure 4***: *Estimated probabilities for the surrogate link function.*

### *4.5 Conclusion*

In this paper we have presented a new method in order to extend the classical one associated with the analysis of a linear model. The constraint of this new method

concerns the utilization of the surrogate link function. Neverthless, the two examples of this paper have suggested that this choice for the link is a good choice. This method has two principal advantages:

1) the Fisher-scoring algorithm is now very easy to improve (and then computations can be done faster),
2) classical designs like factorial designs, well known in the linear case, can be used.

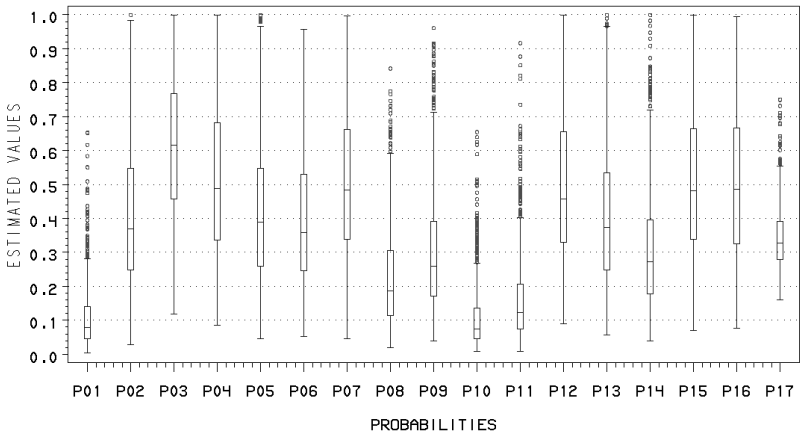


***Figure 5***: *Estimated probabilities for the canonical link function.*

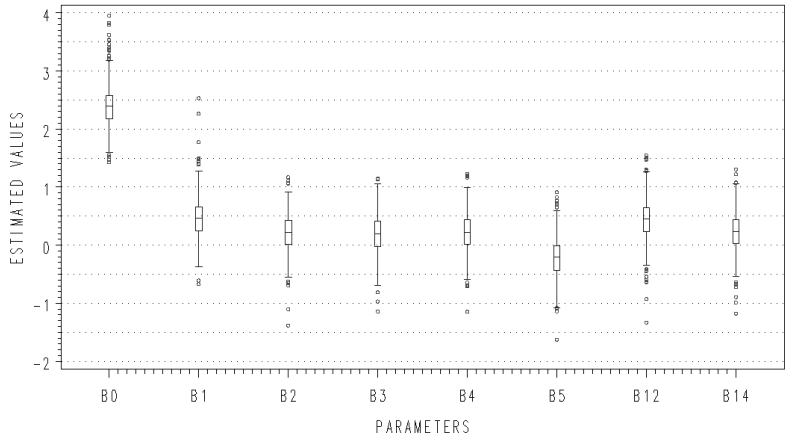


***Figure 6***: *Values of the estimated parameters (natural link function).*

## References

Box G. and Draper N. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley, New-York.

Box G. E. P. and Hunter J. S. (1961*a*). The $2^{k-p}$ fractionnal factorial designs, Part *I*. *Technometrics*, 3, 311-351.

Box G. E. P. and Hunter J. S. (1961*b*). The $2^{k-p}$ fractionnal factorial designs, Part *II*. *Technometrics*, 3, 449-458.

Chipman H. A. and Welch W. J. (1996). D-Optimal Design for Generalized Linear Models, Unpublised. http://www.stats.uwaterloo.ca/~hachipma/publications.html.

Chaloner K. and Larntz K. (1989). Optimal bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21, 191-208.

Green P.J. and Silverman B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Monographs on Statistics and Applied Probability, 58. London: Chapman & Hall.

Khuri A. and Cornell J. (1996). *Response Surfaces: Designs and Analyses*. Dekker, Statistics: textbooks and monographs, 152, New-York.

McCullagh P. and Nelder J. A. (1989). *Generalized Linear Models (second edition).* Monographs on Statistics and Applied Probability, 37. London: Chapman & Hall.

Nelder J. A. and Wedderburn R. W. M. (1972). Generalized linear models. *J. Roy. Stat. Soc.* A, 135, 370-384.

Pukelsheim F. (1993). *Optimal Design of Experiments.* New York: John Wiley.

Sebastiani P. and Settimi R. (1997), A note on D-optimal designs for a logistic regression model. *Journal of statistical planning and inference*, 59, 359-368.

Smyth G. K. (2002). Optimization. *Encyclopedia of Environmetrics*, 3, 1481-1487.