# Maximum likelihood estimation in the logistic regression model with a cure fraction

Diop, Aba
*Université Gaston Berger, Laboratoire d'Etudes et de Recherches en Statistiques et Développement*
*Route de Ngallele*
*Saint-Louis (234), Sénégal*
*E-mail: diopaba7@yahoo.fr*

Diop, Aliou
*Université Gaston Berger, Laboratoire d'Etudes et de Recherches en Statistiques et Développement*
*Route de Ngallele*
*Saint-Louis (234), Sénégal*
*E-mail: aliou.diop@ugb.edu.sn*

Dupuy, Jean-François
*Université de La Rochelle, Mathématiques Image et Applications*
*Avenue Michel Crépeau*
*La Rochelle (17042), France*
*E-mail: jean-francois.dupuy@univ-lr.fr*

## 1 Introduction

Let $(Y_1, S_1, \mathbf{X}_1, \mathbf{Z}_1), \ldots, (Y_n, S_n, \mathbf{X}_n, \mathbf{Z}_n)$ be independent and identically distributed copies of the random vector $(Y, S, \mathbf{X}, \mathbf{Z})$ defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For every individual $i = 1, \ldots, n$, $Y_i$ is a binary response variable indicating say, the infection status with respect to some disease (that is, $Y_i = 1$ if the $i$-th individual is infected, and $Y_i = 0$ otherwise), $S_i$ is a binary variable indicating whether individual $i$ is susceptible to the infection ($S_i = 1$) or immune ($S_i = 0$), and $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})'$ and $\mathbf{Z}_i = (1, Z_{i2}, \ldots, Z_{iq})'$ are corresponding random vectors of predictors, or covariates (both categorical and continuous predictors are allowed). We shall assume in the following that the predictors $\mathbf{X}_i$ are related to the infection status, while the predictors $\mathbf{Z}_i$ are related to immunity. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are allowed to share some common components.

As mentioned above, we consider the situation where the immunity status is unknown for an individual who has not yet developed infection at the time of analysis. That is, if $Y = 0$ for individual $i$, then the value of $S_i$ is unknown. This individual may be either immune to the infection ($S_i = 0$), or susceptible to the infection albeit still uninfected ($S_i = 1$).

The logistic regression model for the infection status assumes that the conditional probability $\mathbb{P}(Y = 1|\mathbf{X}_i, S_i)$ of infection is given by

$$(1) \qquad \log\left(\frac{\mathbb{P}(Y = 1|\mathbf{X}_i, S_i)}{1 - \mathbb{P}(Y = 1|\mathbf{X}_i, S_i)}\right) = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} := \beta' \mathbf{X}_i$$

if $\{S_i = 1\}$, and by

$$(2) \qquad \mathbb{P}(Y = 1|\mathbf{X}_i, S_i) = 0$$

if $\{S_i = 0\}$, where $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ is an unknown regression parameter measuring the association between potential predictors and the risk of infection (for a susceptible individual).

The statistical analysis of infection data with model (1) includes estimation and testing for $\beta$. Without immunity (that is, if $S_i = 1$ for every $i = 1, \ldots, n$), inference on $\beta$ from the sample $(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \ldots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)$ can be based, for example, on the maximum likelihood principle, applied to model (1). Asymptotic results (consistency and asymptotic normality) for the resulting estimator have been established, for example, by [4] and [2]. When immunity is present however, maximum likelihood estimation of $\beta$ is no longer straightforward, since $S_i$ is unknown for every $i$ such that $Y_i = 0$, $i = 1, \ldots, n$. If $Y_i = 0$, we do not know whether $\{S_i = 1\}$, so that (1) applies, or whether $\{S_i = 0\}$, so that (2) applies.

One solution is to consider every individual $i$ such that $\{Y_i = 0\}$ as being susceptible that is, to ignore a possible immunity of this individual. We may however expect this method to produce biased estimates of the association of interest (see [1]). Therefore in this paper, we aim at providing an alternative procedure for estimating $\beta$, which takes account of the possible immunity of those individuals who are still uninfected at the time of analysis. This method is described and its properties are investigated, in the next sections.

## 2 Méthodology and results

Recall that for each individual $i$ ($i = 1, \ldots, n$), the situation is as follows: either $\{Y_i = 1\}$ and we know that $\{S_i = 1\}$ (this individual is infected, and was therefore susceptible to the infection), or $\{Y_i = 0\}$ and we do not know whether this individual is immune ($S_i = 0$) or susceptible to the infection albeit still uninfected ($S_i = 1$). As mentioned above, the usual maximum likelihood estimation of $\beta$ in model (1) is not straightforward from these data. But if a model for immunity is available, we can nevertheless propose an estimation procedure for $\beta$.

A model for the immunity status is defined through the conditional probability $\mathbb{P}(S = 1|\mathbf{Z}_i)$ of being susceptible to the infection. A common choice for this is the logistic model (see, for example, [3] and [5, 6] who considered estimation in various survival regression models with a cure fraction):

$$(3) \qquad \log\left(\frac{\mathbb{P}(S = 1|\mathbf{Z}_i)}{1 - \mathbb{P}(S = 1|\mathbf{Z}_i)}\right) = \theta_1 + \theta_2 Z_{i2} + \ldots + \theta_q Z_{iq} := \theta' \mathbf{Z}_i$$

where $\theta = (\theta_1, \ldots, \theta_q)' \in \mathbb{R}^q$ is an unknown regression parameter (recall that $\mathbf{X}_i$ and $\mathbf{Z}_i$ may have some common components, so that the linear predictors $\beta' \mathbf{X}_i$ and $\theta' \mathbf{Z}_i$ eventually share some common covariates).

From (1), (2), and (3), a straightforward calculation yields that

$$\mathbb{P}(Y = 1|\mathbf{X}_i, \mathbf{Z}_i) = \frac{e^{\beta' \mathbf{X}_i + \theta' \mathbf{Z}_i}}{(1 + e^{\beta' \mathbf{X}_i})(1 + e^{\theta' \mathbf{Z}_i})}.$$

Let $\psi := (\beta', \theta')'$ denote the unknown $k$-dimensional ($k = p + q$) parameter in the conditional distribution of $Y$ given $\mathbf{X}_i$ and $\mathbf{Z}_i$. $\psi$ includes both $\beta$ (considered as the parameter of interest) and $\theta$ (considered as a nuisance parameter). Now, the likelihood for $\psi$ from the independent sample $(Y_i, S_i, \mathbf{X}_i, \mathbf{Z}_i)$ ($i = 1, \ldots, n$) (where $S_i$ is unknown when $Y_i = 0$) is as follows:

$$L_n(\psi) = \prod_{i=1}^{n} \left\{ \left[ \frac{e^{\beta' \mathbf{X}_i + \theta' \mathbf{Z}_i}}{(1 + e^{\beta' \mathbf{X}_i})(1 + e^{\theta' \mathbf{Z}_i})} \right]^{Y_i} \left[ 1 - \frac{e^{\beta' \mathbf{X}_i + \theta' \mathbf{Z}_i}}{(1 + e^{\beta' \mathbf{X}_i})(1 + e^{\theta' \mathbf{Z}_i})} \right]^{1 - Y_i} \right\}.$$

We define the maximum likelihood estimator $\widehat{\psi}_n := (\widehat{\beta}'_n, \widehat{\theta}'_n)'$ of $\psi$ as the solution (if it exists) of the $k$-dimensional score equation

$$(4) \qquad \dot{l}_n(\psi) = \frac{\partial l_n(\psi)}{\partial \psi} = 0,$$

where $l_n(\psi) := \log L_n(\psi)$ is the log-likelihood function. In the following, we shall be interested in the asymptotic properties of the maximum likelihood estimator $\widehat{\beta}_n$ of $\beta$, considered as a sub-component of $\widehat{\psi}_n$. We will however obtain consistency and asymptotic normality results for the whole $\widehat{\psi}_n$.

**Théorème 1** (Identifiability). *the model (1)-(2)-(3) is identifiable that is,* $L_1(\psi) = L_1(\psi^*)$ *almost surely implies* $\psi = \psi^*$.

**Théorème 2** (Existence et consistency). *the maximum likelihood estimator* $\widehat{\psi}_n$ *exists almost surely as* $n \to \infty$, *and converges almost surely to* $\psi_0$.

**Théorème 3** (Normalité asymptotique). *Let* $\widehat{\Sigma}_n = \mathbb{W}\mathbb{D}(\widehat{\psi}_n)\mathbb{W}'$ *and* $I_k$ *denote the identity matrix of order* $k$. *Then* $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0)$ *converges in distribution to the Gaussian vector* $\mathcal{N}(0, I_k)$.

# 3 Simulations study

We present below (partial results) a simulation study to assess the numerical properties of the estimate *widehat beta*$_n$ of the parameter $\beta$. We vary the sample size $n$, and percentage of immune individuals in the sample. The percentage of infected among susceptible is equal to 30%, whatever the immune fraction. The results given in the following table are based on 1500 simulated samples.

| | pourcentage d'immunes dans l'échantillon | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0% | | 25% | | 50% | | 75% | |
| n | $\widehat{\beta}_{1,n}$ | $\widehat{\beta}_{2,n}$ | $\widehat{\beta}_{1,n}$ | $\widehat{\beta}_{2,n}$ | $\widehat{\beta}_{1,n}$ | $\widehat{\beta}_{2,n}$ | $\widehat{\beta}_{1,n}$ | $\widehat{\beta}_{2,n}$ |
| 100 | -0.834 | 1.064 | -0.773 | 1.114 | -0.787 | 1.137 | -0.750 | 0.917 |
| | (0.258) | (0.301) | (0.583) | (0.412) | (0.825) | (0.603) | (0.921) | (0.858) |
| | [0.202] | [0.232] | [0.465] | [0.324] | [0.657] | [0.440] | [0.784] | [0.568] |
| | | 0.965* | | 0.109* | | 0.096* | | 0.121* |
| 500 | -0.807 | 1.012 | -0.783 | 1.111 | -0.788 | 1.129 | -0.791 | 1.120 |
| | (0.107) | (0.125) | (0.320) | (0.354) | (0.428) | (0.389) | (0.707) | (0.538) |
| | [0.085] | [0.099] | [0.264] | [0.227] | [0.352] | [0.270] | [0.603] | [0.407] |
| | | 1* | | 0.985* | | 0.85* | | 0.267* |
| 1000 | -0.801 | 1.004 | -0.794 | 1.058 | -0.798 | 1.060 | -0.797 | 1.108 |
| | (0.077) | (0.085) | (0.241) | (0.202) | (0.310) | (0.247) | (0.683) | (0.482) |
| | [0.062] | [0.068] | [ 0.201] | [0.147] | [0.253] | [0.178] | [0.569] | [0.354] |
| | | 1* | | 1* | | 1* | | 0.567* |
| 1500 | -0.805 | 1.003 | -0.801 | 1.040 | -0.799 | 1.040 | -0.802 | 1.057 |
| | (0.061) | (0.074) | (0.210) | (0.159) | (0.277) | (0.191) | (0.600) | (0.361) |
| | [0.048] | [0.059] | [0.176] | [0.119] | [0.228] | [0.141] | [0.493] | [0.276] |
| | | 1* | | 1* | | 1* | | 0.861* |

Note: $n$: sample size. $(\cdot)$: root mean square error. $[\cdot]$: mean absolute error. $^*$: empirical power ($^\dagger$: empirical size) of the Wald test at the level 5% for testing $H_0 : \beta_2 = 0$. For each percentage of immunes, the percentage of infected among the susceptibles is 30%. All results are based on 1500 replicates.

Figure 1 shows the histograms of 1500 values of $\widehat{\beta}_{2,n}$ obtained and the associated Q-Q plot. This figure allows us to assess graphically the approximation of the law of $\widehat{\beta}_{2,n}$ by a normal distribution.

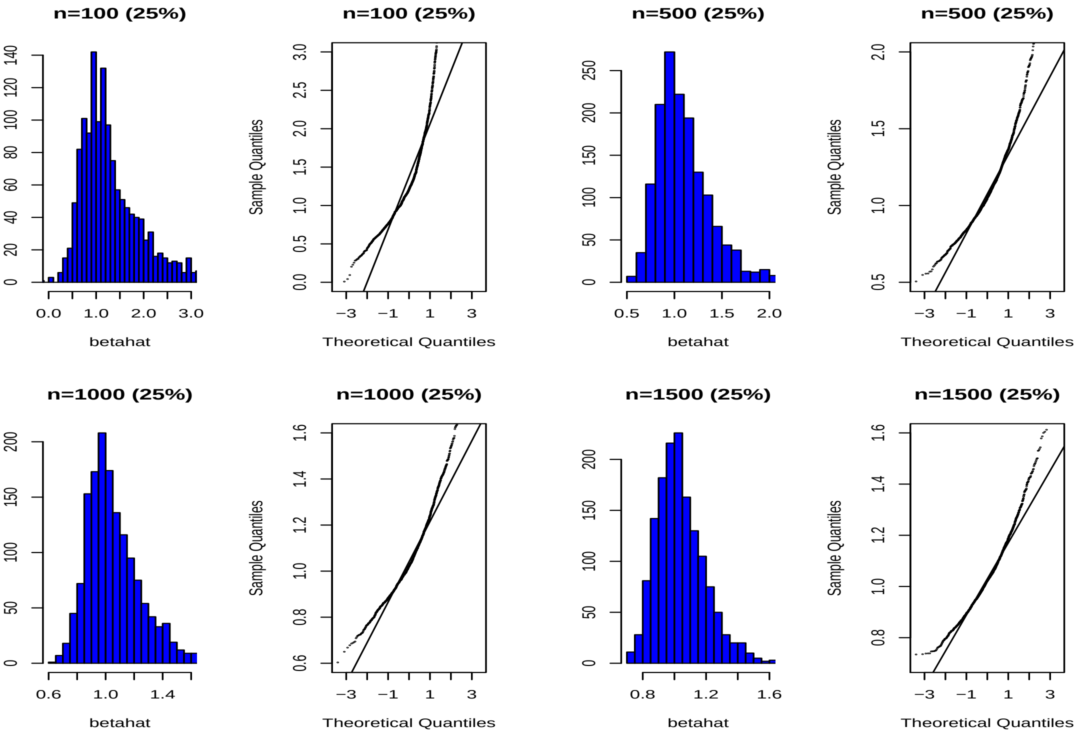

Figure 1: Histogrammes et Q-Q plots pour $\widehat{\beta}_{2,n}$.

All these results confirm the asymptotic properties stated in Section 2. Note the performance degradation of the estimator $\widehat{\beta}_n$ when the proportion of immune increases. The Gaussian approximation of Theorem 2 seems quite satisfied when the proportion of immune is moderate (about 25%) with a sample size greater than or equal to 500.

## References

[1] A. Diop, A. Diop, and J. F. Dupuy. Maximum likelihood estimation in the logistic regression model with a cure fraction. *Soumis*, 2010.

[2] L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

[3] H.-B. Fang, G. Li, and J. Sun. Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics*, 32(1):59–75, 2005.

[4] C. Gouriéroux and A. Monfort. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics*, 17(1):83–97, 1981.

[5] W. Lu. Maximum likelihood estimation in the proportional hazards cure model. *Annals of the Institute of Statistical Mathematics*, 60(3):545–574, 2008.

[6] W. Lu. Efficient estimation for an accelerated failure time model with a cure fraction. *Statistica Sinica*, 20:661–674, 2010.

## ABSTRACT

*Logistic regression is a method for the regression analysis of binary data. This model allows to study the relationship between a binary response* $Y$ *and a set of variables* $X_1, \ldots, X_p$*. In epidemiology, this variable* $Y$ *may be the occurrence of some outcome of interest (*$Y = 1$ *if the outcome occurred and* $Y = 0$ *otherwise). However, this model assumes that all individuals for which the answer* $Y$ *is* $0$ *are likely to develop the event of interest. Thus, it does not take account of a proportion of individuals in the study population, who are not likely to be infected (the so-called "cure fraction"). The immunity status is unknown unless the outcome of interest has been observed. In this work, we develop a maximum likelihood estimation procedure for this problem, based on a joint model for infection and immunity. We establish the identifiability of the proposed model and the consistency and asymptotic normality of the resulting estimator. We conduct a simulation study to investigate the finite-sample behavior of these estimators.*