Maximum likelihood estimation in the logistic regression model with a cure fraction
Aba Diop, Aliou Diop, Jean-Francois Dupuy

To cite this version:

HAL Id: hal-00512311
https://hal.archives-ouvertes.fr/hal-00512311v1
Submitted on 30 Aug 2010 (v1), last revised 11 Jan 2011 (v3)
Maximum likelihood estimation in the logistic regression model with a cure fraction

Aba Diop  
Laboratoire Mathématiques, Image et Applications, Université de La Rochelle, Avenue Michel Crépeau, La Rochelle, France, and Laboratoire d’Etudes et de Recherches en Statistiques et Développement, Université Gaston Berger, Saint Louis, Sénégal  
Email: aba.diop@univ-lr.fr

Aliou Diop  
Laboratoire d’Etudes et de Recherches en Statistiques et Développement, Université Gaston Berger, Saint Louis, Sénégal.  
Email: alioudiop52@yahoo.fr

Jean-François Dupuy†  
Laboratoire Mathématiques, Image et Applications, Université de La Rochelle, Avenue Michel Crépeau, La Rochelle, France  
Email: jean-francois.dupuy@univ-lr.fr

Abstract. Logistic regression is widely used in medical studies to investigate the relationship between a binary response variable $Y$ and a set of potential predictors $X$. The binary response may represent, for example, the occurrence of some outcome of interest ($Y = 1$ if the outcome occurred and $Y = 0$ otherwise). In this paper, we consider the problem of estimating the logistic regression model with a cure fraction. A sample of observations is said to contain a cure fraction when a proportion of the study subjects (the so-called cured individuals, as opposed to the susceptibles) cannot experience the outcome of interest. One problem arising then is that it is usually unknown who are the cured and the susceptible subjects, unless the outcome of interest has been observed. In this setting, a logistic regression analysis of the relationship between $X$ and $Y$ among the susceptibles is no more straightforward. We develop a maximum likelihood estimation procedure for this problem. We establish the consistency and asymptotic normality of the resulting estimator, and we conduct a simulation study to investigate its finite-sample behavior.

Keywords: Logistic regression, Cure fraction, Maximum likelihood estimation, Consistency, Asymptotic normality, Simulations

1. Introduction

Logistic regression is widely used to model binary response data in medical studies. An example of a binary response variable is the infection status (infected vs uninfected) with respect to some disease. A logistic regression model can be used to investigate the relationship between the infection status and various potential predictors. If $Y_i$ denotes the infection status for the $i$-th individual in a sample of size $n$ ($Y_i = 1$ if the individual is infected, and $Y_i = 0$ otherwise), and $X_i$ denotes the corresponding ($p$-dimensional, say)

†Corresponding author
predictor, the logistic regression model expresses the relationship between \( Y_i \) and \( X_i \) in term of the conditional probability \( \mathbb{P}(Y_i = 1 | X_i) \) of infection, as:

\[
\log \left( \frac{\mathbb{P}(Y_i = 1 | X_i)}{1 - \mathbb{P}(Y_i = 1 | X_i)} \right) = \beta^\prime X_i,
\]

where \( \beta \in \mathbb{R}^p \) is an unknown parameter to be estimated. An extensive literature has been devoted so far to the statistical inference in logistic regression models. Estimation and testing procedures for this class of models are now well established and are available in standard statistical softwares. In particular, the maximum likelihood estimator of \( \beta \) is obtained by solving the following score equation:

\[
\sum_{i=1}^{n} X_i \left( Y_i - \frac{e^{\beta^\prime X_i}}{1 + e^{\beta^\prime X_i}} \right) = 0.
\]

Asymptotic results (consistency and asymptotic normality) for this estimator were given by Gouriéroux and Monfort (1981) and Fahrmeir and Kaufmann (1985), among others. We refer the reader to Hosmer and Lemeshow (2000) and Hilbe (2009) for detailed treatments and numerous examples.

In this paper, we consider the problem of estimation in the logistic regression model with a cure fraction. In medical studies, it often arises that a proportion of the study subjects cannot experience the outcome of interest (such as the occurrence of an infection), due to some acquired immunity for example. Such individuals are said to be cured, or immune. The population under study can then be considered as a mixture of cured and susceptible subjects, where a subject is said to be susceptible if he would eventually experience the outcome of interest. One problem arising in this setting is that it is usually unknown who are the susceptible, and the cured subjects (unless the outcome of interest has been observed). Consider, for example, the occurrence of infection from some disease to be the outcome of interest. Then, if a subject is uninfected, the investigator does usually not know whether this subject is immune to the infection, or susceptible albeit still uninfected.

The problem of statistical inference with a cure fraction has recently attracted much attention. Over the past few years for example, estimation in survival regression models from survival data with a cure fraction has given rise to an extensive literature (recent references include Peng (2003), Lu and Ying (2004), Fang et al. (2005), Sugimoto and Hamasaki (2006), Lu (2008, 2010); see also the references therein). But to the best of our knowledge, despite the widespread use of logistic regression in medical applications, there has not been yet any investigation about inference in the logistic model from binary response data with a cure fraction. In this paper, we intend to fill this gap by developing an appropriate estimation method for this problem. The estimator we propose is obtained by maximizing a likelihood function derived from a joint regression model for the binary response of interest and the cure indicator, considered as a random variable whose distribution is modeled by a logistic regression. We prove the almost sure asymptotic existence, the consistency, and the asymptotic normality of this estimator. Then, we investigate its finite-sample properties via simulations.

The rest of this paper is organized as follows. In Section 2, we describe the problem of logistic regression with a cure fraction, and we propose an estimation method adapted to this setting. In Section 3, we derive the asymptotic properties of the resulting estimator. Section 4 describes a detailed simulation study, where we numerically investigate the small
Maximum likelihood estimation in the logistic regression model with a cure fraction

2. Logistic regression with a cure fraction

2.1. Notations and the model set-up
Let \((Y_1, S_1, X_1, Z_1), \ldots, (Y_n, S_n, X_n, Z_n)\) be independent and identically distributed copies of the random vector \((Y, S, X, Z)\) 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 \(X_i = (1, X_{i2}, \ldots, X_{ip})'\) and \(Z_i = (1, Z_{i2}, \ldots, Z_{iq})'\) are corresponding random vectors of predictors, or covariates (both categorical and continuous predictors are allowed). We shall assume in the following that the predictors \(X_i\) and \(Z_i\) are related to the infection status, while the predictors \(S_i\) are related to immunity. \(X_i\) and \(Z_i\) are allowed to share some common components.

As mentioned in the introduction, 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 \(\mathbb{P}(Y = 1|X_i, S_i)\) of infection is given by

\[
\log \left( \frac{\mathbb{P}(Y = 1|X_i, S_i)}{1 - \mathbb{P}(Y = 1|X_i, S_i)} \right) = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} := \beta' X_i \tag{1}
\]

if \(\{S_i = 1\}\), and by

\[
\mathbb{P}(Y = 1|X_i, S_i) = 0 \tag{2}
\]

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 \((Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)\) 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 Gouriéroux and Monfort (1981) and Fahrmeir and Kaufmann (1985). 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 (such a method will be evaluated in the simulation study described in section 4). Therefore in this paper, we aim at providing an \textit{ad hoc} 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.
2.2. The proposed estimation procedure

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 \(\mathbb{P}(S = 1|Z_i)\) of being susceptible to the infection. A common choice for this is the logistic model (see, for example, Fang et al. (2005) and Lu (2008, 2010) who considered estimation in various survival regression models with a cure fraction):

\[
\log \left( \frac{\mathbb{P}(S = 1|Z_i)}{1 - \mathbb{P}(S = 1|Z_i)} \right) = \theta_1 + \theta_2 Z_{i2} + \ldots + \theta_q Z_{iq} := \theta'Z_i
\]

(3)

where \(\theta = (\theta_1, \ldots, \theta_q)' \in \mathbb{R}^q\) is an unknown regression parameter (recall that \(X_i\) and \(Z_i\) may have some common components, so that the linear predictors \(\beta'X_i\) and \(\theta'Z_i\) eventually share some common covariates). From (1), (2), and (3), a straightforward calculation yields that

\[
\mathbb{P}(Y = 1|X_i, Z_i) = \frac{e^{\beta'X_i + \theta'Z_i}}{(1 + e^{\beta'X_i})(1 + e^{\theta'Z_i})}.
\]

Let \(\psi := (\beta', \theta')'\) denote the unknown \(k\)-dimensional \((k = p + q)\) parameter in the conditional distribution of \(Y\) given \(X_i\) and \(Z_i\). \(\psi\) 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 \((Y_i, S_i, X_i, Z_i)\) \((i = 1, \ldots, n)\) (where \(S_i\) is unknown when \(Y_i = 0\)) as is as follows:

\[
L_n(\psi) = \prod_{i=1}^{n} \left\{ \frac{e^{\beta'X_i + \theta'Z_i}}{(1 + e^{\beta'X_i})(1 + e^{\theta'Z_i})} Y_i \left[ 1 - \frac{e^{\beta'X_i + \theta'Z_i}}{(1 + e^{\beta'X_i})(1 + e^{\theta'Z_i})} \right]^{1-Y_i} \right\}
\]

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

\[
\dot{l}_n(\psi) = \frac{\partial L_n(\psi)}{\partial \psi} = 0,
\]

(4)

where \(\dot{l}_n(\psi) := \log L_n(\psi)\) is the log-likelihood function. In the following, we shall be interested in the asymptotic properties of the maximum likelihood estimator \(\hat{\beta}_n\) of \(\beta\), considered as a sub-component of \(\hat{\psi}_n\). We will however obtain consistency and asymptotic normality results for the whole \(\hat{\psi}_n\). Before proceeding, we need to set some further notations.

2.3. Some further notations

Define first the \((p \times n)\) and \((q \times n)\) matrices

\[
X = \begin{pmatrix}
1 & 1 & \cdots & 1 \\
X_{12} & X_{22} & \cdots & X_{n2} \\
\vdots & \vdots & \ddots & \vdots \\
X_{1p} & X_{2p} & \cdots & X_{np}
\end{pmatrix}
\quad \text{and} \quad
Z = \begin{pmatrix}
1 & 1 & \cdots & 1 \\
Z_{12} & Z_{22} & \cdots & Z_{n2} \\
\vdots & \vdots & \ddots & \vdots \\
Z_{1q} & Z_{2q} & \cdots & Z_{nq}
\end{pmatrix}
\]
and let $\mathcal{W}$ be the $(k \times 2n)$ block-matrix defined as

$$
\mathcal{W} = \begin{bmatrix}
    X & 0_{pn} \\
    0_{qn} & Z
\end{bmatrix},
$$

where $0_{a \times b}$ denotes the $(a \times b)$ matrix whose components are all equal to zero (for any positive integer values $a, b$). Let also $C(\psi)$ be the $2n$-dimensional column vector defined as

$$
C(\psi) = \left((A^\beta(\psi) - B^\beta(\psi))^\prime, (A^\theta(\psi) - B^\theta(\psi))^\prime\right)^\prime,
$$

where $A^\beta(\psi) = (A^\beta_i(\psi))_{1 \leq i \leq n}$, $B^\beta(\psi) = (B^\beta_i(\psi))_{1 \leq i \leq n}$, $A^\theta(\psi) = (A^\theta_i(\psi))_{1 \leq i \leq n}$, and $B^\theta(\psi) = (B^\theta_i(\psi))_{1 \leq i \leq n}$ are $n$-dimensional column vectors with respective elements

$$
A^\beta_i(\psi) = \frac{1 + e^{\beta^\prime X_i + \theta^\prime Z_i}}{1 + e^{\beta^\prime X_i + \theta^\prime Z_i} Y_i}, \quad B^\beta_i(\psi) = \frac{e^{\beta^\prime X_i + \theta^\prime Z_i}}{(1 + e^{\beta^\prime X_i + \theta^\prime Z_i})(1 + e^{\beta^\prime X_i + \theta^\prime Z_i})},
$$

$$
A^\theta_i(\psi) = \frac{1 + e^{\beta^\prime X_i + \theta^\prime Z_i}}{1 + e^{\beta^\prime X_i + \theta^\prime Z_i} Y_i}, \quad B^\theta_i(\psi) = \frac{e^{\beta^\prime X_i + \theta^\prime Z_i}}{(1 + e^{\theta^\prime Z_i})(1 + e^{\beta^\prime X_i + \theta^\prime Z_i})}.
$$

Then, simple algebra shows that the score equation can be rewritten as

$$
\hat{I}_n(\psi) = \mathcal{W} C(\psi) = 0.
$$

If $M = (M_{ij})_{1 \leq i \leq n, 1 \leq j \leq b}$ denotes some $(a \times b)$ matrix, we will denote by $M_{\bullet j}$ its $j$-th column ($j = 1, \ldots, b$) that is, $M_{\bullet j} = (M_{1j}, \ldots, M_{nj})^\prime$. Then, it will be useful to rewrite the score vector as

$$
\hat{I}_n(\psi) = \sum_{j=1}^{2n} \mathcal{W}_{\bullet j} C_j(\psi).
$$

We shall further note $\hat{I}_n(\psi)$ the $(k \times k)$ matrix of second derivatives of $I_n(\psi)$ that is, $\hat{I}_n(\psi) = \partial^2 I_n(\psi)/\partial \psi \partial \psi^\prime$. Let $D(\psi) = (D_{ij}(\psi))_{1 \leq i, j \leq 2n}$ be the $(2n \times 2n)$ block matrix defined as

$$
D(\psi) = \begin{bmatrix}
    D_1(\psi) & D_3(\psi) \\
    D_3(\psi) & D_2(\psi)
\end{bmatrix},
$$

where $D_1(\psi), D_2(\psi),$ and $D_3(\psi)$ are $(n \times n)$ diagonal matrices, with $i$-th diagonal elements ($i = 1, \ldots, n$) respectively given by

$$
D_{1,ii}(\psi) = \frac{e^{\beta^\prime X_i + \theta^\prime Z_i}}{(1 + e^{\beta^\prime X_i + \theta^\prime Z_i})^2(1 + e^{\beta^\prime X_i + \theta^\prime Z_i})},
$$

$$
D_{2,ii}(\psi) = \frac{e^{\beta^\prime X_i + \theta^\prime Z_i}}{(1 + e^{\theta^\prime Z_i})^2(1 + e^{\beta^\prime X_i + \theta^\prime Z_i})},
$$

$$
D_{3,ii}(\psi) = \frac{e^{\beta^\prime X_i + \theta^\prime Z_i}}{(1 + e^{\beta^\prime X_i + \theta^\prime Z_i})(1 + e^{\theta^\prime Z_i})^2(1 + e^{\beta^\prime X_i + \theta^\prime Z_i})}.
$$

Then, some algebra shows that $\hat{I}_n(\psi)$ can be expressed as

$$
\hat{I}_n(\psi) = -D(\psi) \mathcal{W}^\prime.
$$

Note that the size of $C(\psi), \mathcal{W},$ and $D(\psi)$ depends on $n$. However, in order to simplify notations, $n$ will not be used as a lower indice for these vector and matrices.
3. Asymptotic theory

In this section, we establish rigorously the existence, consistency and asymptotic normality of the maximum likelihood estimator \( \hat{\beta}_n \) of \( \beta \) in model (1), obtained from a sample of binary response data with a cure fraction. Some regularity conditions are needed to ensure these results. Assume

C1 The covariates are bounded that is, there exists a finite positive constant \( c_1 \) such that 
\[ |X_{ij}| < c_1 \text{ and } |Z_{ik}| < c_1 \text{ for every } i = 1, 2, \ldots, j = 1, \ldots, p, k = 1, \ldots, q. \]

C2 Let \( \psi_0 = (\beta_0', \theta_0')' \) denote the true parameter value. \( \beta_0 \) and \( \theta_0 \) lie in the interior of known compact sets \( B \subset \mathbb{R}^p \) and \( G \subset \mathbb{R}^q \) respectively.

C3 The Hessian matrix \( \hat{I}_n(\psi) \) is negative definite and of full rank, for every \( n = 1, 2, \ldots \) Let \( \lambda_n \) and \( \Lambda_n \) be respectively the smallest and largest eigenvalues of \( \hat{W}D(\psi_0)\hat{W}' \). There exists a finite positive constant \( c_2 \) such that 
\[ \lambda_n / \Lambda_n < c_2 \text{ for every } n = 1, 2, \ldots \]

In the sequel, the space \( \mathbb{R}^k \) of \( k \)-dimensional (column) vectors will be provided with the euclidean norm, and the space \( \mathbb{R}^{k \times k} \) of \( (k \times k) \) real matrices will be provided with the spectral norm (we will use the same notation \( \| \cdot \| \) for both). We first prove the following result:

**Theorem 1 (Existence and consistency).** Under the conditions C1-C3, the maximum likelihood estimator \( \hat{\psi}_n \) exists almost surely as \( n \to \infty \), and converges almost surely to \( \psi_0 \), if and only if \( \lambda_n \) tends to infinity as \( n \to \infty \).

**Proof of Theorem 1.** The principle of the proof is similar to Gouriéroux and Monfort (1981) but the technical details are different. Three lemmas are needed. The first lemma essentially provides an intermediate technical result. Its proof is postponed to the appendix.

**Lemma 1.** Let \( \phi_n : \mathbb{R}^k \longrightarrow \mathbb{R}^k \) be defined as 
\[ \phi_n(\psi) = \psi + (\hat{W}D(\psi_0)\hat{W}')^{-1}\hat{I}_n(\psi). \]
Then there exists an open ball \( B(\psi_0, r) \) (with \( r > 0 \)) such that \( \phi_n \) satisfies the Lipschitz condition on \( B(\psi_0, r) \) that is,
\[ \|\phi_n(\psi_1) - \phi_n(\psi_2)\| \leq c \|\psi_1 - \psi_2\| \text{ for all } \psi_1, \psi_2 \in B(\psi_0, r), \quad (5) \]
and \( 0 < c < 1 \).

**Lemma 2.** The maximum likelihood estimator \( \hat{\psi}_n \) exists almost surely as \( n \to \infty \), and converges almost surely to \( \psi_0 \), if and only if \( (\hat{W}D(\psi_0)\hat{W}')^{-1}\hat{I}_n(\psi_0) \) converges almost surely to 0.

**Proof of Lemma 2.** We first prove that the condition is sufficient. Thus, we assume that 
\( (\hat{W}D(\psi_0)\hat{W}')^{-1}\hat{I}_n(\psi_0) \) converges almost surely to 0.
Define \( \eta_n(\psi) = \psi - \phi_n(\psi) = -(\hat{W}D(\psi_0)\hat{W}')^{-1}\hat{I}_n(\psi) \) and let \( \epsilon \) be an arbitrary positive value. Then for almost every \( \omega \in \Omega \), there exists an integer value \( n(\epsilon, \omega) \) such that for any \( n \geq n(\epsilon, \omega) \), 
\[ \|\eta_n(\psi_0)\| \leq \epsilon \text{ or equivalently, } 0 \in B(\eta_n(\psi_0), \epsilon). \]
In particular, let \( \epsilon = (1 - c)s \) with \( 0 < c < 1 \) such as in Lemma 1. Since \( \phi_n \) satisfies the Lipschitz condition (5) (by
Lemma 1), the lemma 2 of Gouriéroux and Monfort (1981) ensures that there exists an element of $B(\psi_0, s)$ (let denote this element by $\hat{\psi}_n$) such that $\eta_n(\hat{\psi}_n) = 0$ that is,

$$(\mathbb{W} \mathbb{D}(\psi_0) \mathbb{W}')^{-1} \hat{l}_n(\hat{\psi}_n) = 0.$$ 

The condition C3 implies that $\hat{l}_n(\hat{\psi}_n) = 0$ and that $\hat{\psi}_n$ is the unique maximizer of $l_n$. To summarize, we have shown that for almost every $\omega \in \Omega$ and for every $s > 0$, there exists an integer value $n(s, \omega)$ such that if $n \geq n(s, \omega)$, then the maximum likelihood estimator $\hat{\psi}_n$ exists, and $\|\hat{\psi}_n - \psi_0\| \leq s$ (that is, $\hat{\psi}_n$ converges almost surely to $\psi_0$).

We now prove that the condition that $\eta_n(\psi_0)$ converges almost surely to 0 is necessary. We use a proof by contradiction.

Assume that as $n \to \infty$, $\hat{\psi}_n$ exists and converges almost surely to $\psi_0$, but $\eta_n(\psi_0)$ does not converge almost surely to 0. Then there exists a set $\tilde{\Omega} \subset \Omega$ with $\mathbb{P}(\tilde{\Omega}) > 0$, such that if $\omega \in \tilde{\Omega}$, there exists $\epsilon > 0$ such that for every $m \in \mathbb{N}$, there exists $n \geq m$ with $\|\eta_n(\psi_0)\| > \epsilon$. Now, let $t = \frac{\epsilon}{\epsilon + 1}$, with $d > 1$ sufficiently large so that $t \leq r$, where $r$ is such as in Lemma 1. Then for every $\psi \in B(\psi_0, t)$, the following holds:

$$\|\eta_n(\psi_0) - \eta_n(\psi)\| = \|\psi_0 - \phi_n(\psi_0) - \psi + \phi_n(\psi)\| \leq \|\psi_0 - \psi\| + \|\phi_n(\psi) - \phi_n(\psi_0)\| \leq t(1 + c) = \frac{\epsilon}{d},$$

where the second to third line follows by Lemma 1. Therefore, for every $\psi \in B(\psi_0, t)$,

$$\epsilon < \|\eta_n(\psi_0)\| \leq \|\eta_n(\psi_0) - \eta_n(\psi)\| + \|\eta_n(\psi)\| \leq \|\eta_n(\psi)\| + \frac{\epsilon}{d},$$

and we conclude that for every $\psi \in B(\psi_0, t)$, $\|\eta_n(\psi)\| > \epsilon(1 - \frac{1}{d}) > 0$. Since $\eta_n(\hat{\psi}_n) = 0$, $\hat{\psi}_n$ cannot belong to $B(\psi_0, t)$ for large $n$, which implies that $\hat{\psi}_n$ does not converge almost surely to $\psi_0$. This is the desired contradiction.

□

**Lemma 3.** $(\mathbb{W} \mathbb{D}(\psi_0) \mathbb{W}')^{-1} \hat{l}_n(\psi_0)$ converges almost surely to 0 if and only if $\lambda_n$ tends to infinity as $n \to \infty$.

**Proof of Lemma 3.** We first prove that the condition is sufficient that is, we assume that $\lambda_n$ tends to infinity as $n \to \infty$. Define the $(2n \times k)$ matrix $\mathbb{V} = (\mathbb{D}(\psi_0))^{\frac{1}{2}} \mathbb{W}'$ and the $2n$-dimensional vector $U = (\mathbb{D}(\psi_0))^{\frac{1}{2}} \mathbb{C}(\psi_0)$. Then

$$E[U] = 0 \quad \text{and} \quad \text{var}[U] = I_{2n}, \quad (6)$$

where $I_{2n}$ denotes the identity matrix of order $2n$. To see this, note that

$$E[U] = E[E[(\mathbb{D}(\psi_0))^{\frac{1}{2}} \mathbb{C}(\psi_0)|X, Z]]$$

$$= E[(\mathbb{D}(\psi_0))^{\frac{1}{2}} E[\mathbb{C}(\psi_0)|X, Z]]$$

$$= E[(\mathbb{D}(\psi_0))^{\frac{1}{2}} \mathbb{E}[(A^\theta(\psi_0) - B^\theta(\psi_0))', (A^\theta(\psi_0) - B^\theta(\psi_0))'|X, Z]].$$
For every $i = 1, \ldots, n$, $E[A_i^\beta(\psi_0) - B_i^\beta(\psi_0)|X_i, Z_i] = E[A_i^\beta(\psi_0) - B_i^\beta(\psi_0)|X_i, Z_i]$ by independence between the individuals, and

$$E[A_i^\beta(\psi_0) - B_i^\beta(\psi_0)|X_i, Z_i] = \frac{1 + e^\psi_iZ_i}{1 + e^\psi_iX_i + e^\psi_iZ_i}P(Y_i = 1|X_i, Z_i) - B_i^\beta(\psi_0)$$

$$= B_i^\beta(\psi_0) - B_i^\beta(\psi_0)$$

$$= 0.$$

Similarly, $E[A_i^\beta(\psi_0) - B_i^\beta(\psi_0)|X_i, Z_i] = 0$ for every $i = 1, \ldots, n$ and thus, $E[C(\psi_0)|X, Z] = 0$ and $E[U] = 0$.

Next, $\text{var}[U|X, Z] = \text{var}[\text{var}[U|X, Z]]$ since $E[U|X, Z] = 0$. Moreover,

$$\text{var}[U|X, Z] = \left(\text{var}(\underline{\psi|(\psi_0)})^{1/2}\right)^2\left(\text{var}(C(\psi_0)|X, Z)\right)^{-1/2},$$

with $\text{var}[C(\psi_0)|X, Z] = \text{var}\left(A^\beta(\psi_0), A^\beta(\psi_0)^T\right)^T|X, Z|$ a $(2n \times 2n)$ block-matrix of the form

$$\begin{bmatrix}
V_1 & V_3 \\
V_2 & V_3
\end{bmatrix}$$

where $V_1$, $V_2$, and $V_3$ are $(n \times n)$ matrices. The $i$-th diagonal elements ($i = 1, \ldots, n$) of $V_1$, $V_2$, and $V_3$ are $\text{var}[A_i^\beta(\psi_0)|X, Z]$, $\text{var}[A_i^\beta(\psi_0)|X, Z]$, and $\text{cov}[A_i^\beta(\psi_0), A_i^\beta(\psi_0)|X, Z]$ respectively. Similar calculations as above yield: $\text{var}[A_i^\beta(\psi_0)|X, Z] = D_{1,ii}(\psi_0)$, $\text{var}[A_i^\beta(\psi_0)|X, Z] = D_{2,ii}(\psi_0)$, and $\text{cov}[A_i^\beta(\psi_0), A_i^\beta(\psi_0)|X, Z] = D_{i,j}(\psi_0)$. Note also that $V_1$, $V_2$, and $V_3$ are diagonal matrices, by independence between the individuals. It follows that $\text{var}[C(\psi_0)|X, Z] = D(\psi_0)$ and thus, $\text{var}[U|X, Z] = I_{2n}$ and $\text{var}[U] = I_{2n}$.

By Gouriéroux and Monfort (1981) (proof of Lemma 4), if (6) holds, $\Lambda_n/\lambda_n < c_2$ for every $n = 1, 2, \ldots$, and $\lambda_n$ tends to infinity as $n \rightarrow \infty$, then

$$(\mathcal{V}'\mathcal{V})^{-1}\mathcal{V}'U \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty$$

that is, $(\mathbb{W}(\psi_0)\mathbb{W})^{-1}\hat{I}_n(\psi_0)$ converges almost surely to 0.

We now prove that the condition is necessary. Assume that $\lambda_n$ does not tend to infinity as $n \rightarrow \infty$. By Gouriéroux and Monfort (1981) (proof of Lemma 4), $(\mathcal{V}'\mathcal{V})^{-1}\mathcal{V}'U$ (and therefore $(\mathbb{W}(\psi_0)\mathbb{W})^{-1}\hat{I}_n(\psi_0)$) cannot converge to 0, which concludes the proof.

Finally, Theorem 1 follows by Lemma 2 and Lemma 3.

We now turn to the convergence in distribution of the proposed estimator, which is stated by the following theorem:

**Theorem 2 (Asymptotic normality).** Assume that the conditions C1-C3 hold and that $\hat{\psi}_n$ converges almost surely to $\psi_0$. Let $\Sigma_n = \mathbb{W}(\hat{\psi}_n)\mathbb{W}'$ and $I_k$ denote the identity matrix of order $k$. Then $\Sigma_n^{-1/2}(\hat{\psi}_n - \psi_0)$ converges in distribution to the Gaussian vector $\mathcal{N}(0, I_k)$.

**Proof of Theorem 2.** A Taylor expansion of the score function is as

$$0 = \hat{I}_n(\hat{\psi}) = \hat{I}_n(\psi_0) + \hat{I}_n(\hat{\psi}_n)(\hat{\psi}_n - \psi_0).$$
where $\tilde{\psi}_n$ lies between $\psi_n$ and $\psi_0$, and thus $\tilde{l}_n(\psi_0) = \tilde{l}_n(\psi_n)(\psi_n - \psi_0)$. Let $\hat{\Sigma}_n := -\tilde{l}_n(\psi_n) = \mathbb{WD}(\tilde{\psi}_n)\mathbb{W}'$ and $\Sigma_{n,0} := \mathbb{WD}(\psi_0)\mathbb{W}'$. Now,

$$\hat{\Sigma}_n^n(\tilde{\psi}_n - \psi_0) = [\hat{\Sigma}_n^n, \hat{\Sigma}_n^n, \hat{\Sigma}_n^n] \left[ \hat{\Sigma}_n^n, \hat{\Sigma}_n^n, \hat{\Sigma}_n^n \right] \hat{\Sigma}_n^n(\tilde{\psi}_n - \psi_0).$$

(7)

The two terms in brackets in (7) converge almost surely to $I_k$. To see this, we show for example that $\|\hat{\Sigma}_n^n - \hat{\Sigma}_n^n, - I_k\| \overset{a.s.}{\to} 0$ as $n \to \infty$. First, note that

$$\|\hat{\Sigma}_n^n - \hat{\Sigma}_n^n, - I_k\| \leq \Lambda_n^{-2} \|\hat{\Sigma}_n^n - \hat{\Sigma}_n^n, - I_k\| \|\Lambda_n^{-2} \left( \hat{\Sigma}_n^n - \hat{\Sigma}_n^n, - I_k\right)\|,$$

(8)

and

$$\Lambda_n^{-2} \|\hat{\Sigma}_n^n, - \hat{\Sigma}_n^n\| = \Lambda_n^{-2} \|\mathbb{WD}(\psi_0) - \mathbb{D}(\tilde{\psi}_n)\mathbb{W}'\|.$$

Note also that $\tilde{\psi}_n$ converges almost surely to $\psi_0$ (that is, for every $\omega \in \Omega$, where $\Omega \subset \Omega$ and $\mathbb{P}(\Omega) = 1$). Let $\omega \in \Omega$. By the same arguments as in the proof of Lemma 1, for every $\epsilon > 0$, there exists a positive $n(\epsilon, \omega) \in \mathbb{N}$ such that if $n \geq n(\epsilon, \omega)$, then $\Lambda_n^{-2} \|\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\tilde{\psi}_n))\mathbb{W}'\| \leq \epsilon$. Hence $\Lambda_n^{-2} \|\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\tilde{\psi}_n))\mathbb{W}'\|$ converges almost surely to 0. By continuity of the map $x \mapsto x^2$, $\Lambda_n^{-2} \left( \hat{\Sigma}_n^n - \hat{\Sigma}_n^n, - I_k\right)$ converges also almost surely to 0. Moreover, for sufficiently large, there exists a positive constant $c_4 < \infty$ such that almost surely, $\|\hat{\Sigma}_n^n\| \leq c_4 \lambda_n^{-2}$. It follows from (8) and the condition C3 that $\|\hat{\Sigma}_n^n, - \hat{\Sigma}_n^n, - I_k\|$ converges almost surely to 0. The almost sure convergence to 0 of $\|\hat{\Sigma}_n^n, - \hat{\Sigma}_n^n, - I_k\|$ follows by similar arguments.

It remains us to show that $\Sigma_{n,0}^{-2}(\tilde{\Sigma}_n^n(\tilde{\psi}_n - \psi_0))$ converges in distribution to $\mathcal{N}(0, I_k)$, or equivalently, that $(\mathbb{W})^{-2} \mathbb{W}U$ converges in distribution to $\mathcal{N}(0, I_k)$. Following Eicker (1966), this convergence holds if we can check the following conditions: i) $\sup_{1 \leq i \leq 2n} \mathbb{V}_i \hat{\Sigma}_n^n \hat{\Sigma}_n^n \mathbb{V}_i' \to 0$ as $n \to \infty$, ii) $\sup_{1 \leq i \leq 2n} \mathbb{E}[U_i^2(1_{\{U_i > 0\}})] \to 0$ as $\alpha \to \infty$, iii) $\inf_{1 \leq i \leq 2n} \mathbb{E}[U_i^2] > 0$, where $\mathbb{V}_i$ and $U_i$ respectively denote the $i$-th raw of $\mathbb{V}$ and the $i$-th component of $U$, $i = 1, \ldots, 2n$. Condition i) follows by noting that

$$0 \leq \max_{1 \leq i \leq 2n} \mathbb{V}_i \hat{\Sigma}_n^n \hat{\Sigma}_n^n \mathbb{V}_i' \leq \max_{1 \leq i \leq 2n} \|\mathbb{V}_i\| \|\hat{\Sigma}_n^n\| \|\hat{\Sigma}_n^n\| = \max_{1 \leq i \leq 2n} \frac{1}{\lambda_n} \|\mathbb{V}_i\|^2,$$

and that $\|\mathbb{V}_i\|$ is bounded above, by C1 and C2. Moreover, $\frac{1}{\lambda_n}$ tends to 0 as $n \to \infty$, since $\tilde{\psi}_n$ converges almost surely to $\psi_0$. Condition ii) follows by noting that the components $U_i$ of $U$ are bounded under C1 and C2. Finally, for every $i = 1, \ldots, 2n$, $\mathbb{E}[U_i^2] = \mathbb{E}[U_i^2] = 1 > 0$. To summarize, we have proved that $\Sigma_{n,0}^{-2}(\tilde{\Sigma}_n^n(\tilde{\psi}_n - \psi_0))$ converges in distribution to $\mathcal{N}(0, I_k)$. This result, combined with Slutsky's theorem and equation (7), implies that $\tilde{\Sigma}_n^n(\tilde{\psi}_n - \psi_0)$ converges in distribution to $\mathcal{N}(0, I_k)$.

□
4. A simulation study

4.1. Study design

In this section, we investigate the numerical properties of the maximum likelihood estimator \( \hat{\beta}_n \), under various conditions. The simulation setting is as follows. We consider the following models for the infection status:

\[
\begin{align*}
\log \left( \frac{\Pr(Y = 1|X_i, S_i)}{1 - \Pr(Y = 1|X_i, S_i)} \right) &= \beta_1 + \beta_2 X_{i2} & \text{if } S_i = 1 \\
\Pr(Y = 1|X_i, S_i) &= 0 & \text{if } S_i = 0
\end{align*}
\]

and the immunity status:

\[
\log \left( \frac{\Pr(S = 1|Z_i)}{1 - \Pr(S = 1|Z_i)} \right) = \theta_1 + \theta_2 Z_{i2},
\]

where \( X_{i2} \) is normally distributed with mean 0 and variance 1, and \( Z_{i2} \) is normally distributed with mean 1 and variance 1. An i.i.d. sample of size \( n \) of the vector \((Y, S, X, Z)\) is generated from this model, and for each individual \( i \), we get a realization \((y_i, s_i, x_i, z_i)\), where \( s_i \) is considered as unknown if \( y_i = 0 \). A maximum likelihood estimator \( \hat{\beta}_n \) of \( \beta = (\beta_1, \beta_2) \) is obtained from this incomplete dataset by solving the score equation (4), using the \texttt{optim} function of the software \texttt{R}. An estimate is also obtained for \( \theta = (\theta_1, \theta_2) \), but \( \theta \) is not the primary parameter of interest hence we only focus on the simulation results for \( \hat{\beta}_n \).

The finite-sample behavior of the maximum likelihood estimator \( \hat{\beta}_n \) was assessed for several sample sizes \((n = 100, 500, 1000, 1500)\) and various values for the percentage of immunes in the sample, namely 25%, 50%, and 75%. The case where it is known that there are no immunes in the sample was also considered. In this case, there is no missing information about the infection status and therefore, this case provides a benchmark for evaluating the performance of the proposed estimation method. We also considered different values for the proportion of infected individuals among the susceptibles. The desired proportions of immunes and infected were obtained by choosing appropriate values for the parameters \( \beta \) (the parameter of interest) and \( \theta \) (the nuisance parameter). The following values were considered for \( \beta \): i) model \( \mathcal{M}_1: \beta = (-.8, 1) \) (using these values, approximately 30% of the susceptibles are infected), ii) model \( \mathcal{M}_2: \beta = (1, .7) \) (approximately 70% of the susceptibles are infected), iii) model \( \mathcal{M}_3: \beta = (-.8, 0) \) (approximately 30% of the susceptibles are infected), iv) model \( \mathcal{M}_4: \beta = (1, 0) \) (approximately 70% of the susceptibles are infected).

4.2. Results

For each configuration (sample size, percentage of immunes, percentage of infected among susceptibles) of the design parameters, \( N = 1500 \) samples were obtained. Based on these 1500 repetitions, we obtain averaged values for the estimates of \( \beta_1 \) and \( \beta_2 \), which are calculated as \( N^{-1} \sum_{j=1}^{N} \hat{\beta}_{1,n}^{(j)} \) and \( N^{-1} \sum_{j=1}^{N} \hat{\beta}_{2,n}^{(j)} \), where \( \hat{\beta}_{1,n}^{(j)} = (\hat{\beta}_{1,n}^{(j)}, \hat{\beta}_{2,n}^{(j)}) \) is the estimate obtained from the \( j \)-th simulated sample. For each of the parameters \( \beta_1 \) and \( \beta_2 \), we also obtain the empirical root mean square and mean absolute errors, based on the \( N \) samples. When \( \beta_2 \neq 0 \) (respectively \( \beta_2 = 0 \)), we obtain the empirical power (respectively the empirical size) of the Wald test at the 5% level for testing \( H_0: \beta_2 = 0 \) (models \( \mathcal{M}_1 \) and \( \mathcal{M}_2 \), see Tables 1 and 2) (respectively models \( \mathcal{M}_3 \) and \( \mathcal{M}_4 \), see Tables 3 and 4). The null hypothesis \( H_0: \beta_2 = 0 \) is the hypothesis that the predictor \( X_2 \) does not influence the risk of infection of susceptible individuals. The results are summarized in Tables 1, 2, 3, 4.
Maximum likelihood estimation in the logistic regression model with a cure fraction

From these tables, it appears that the proposed maximum likelihood estimator $\hat{\beta}_n$ provides a reasonable approximation of the true parameter value, even when the percentage of immunes is high. While the bias of $\hat{\beta}_n$ stays limited, its variability increases with the immune fraction, sometimes drastically when the sample size is small. Consequently, when the sample size is small ($n = 100$) and/or the immune proportion is very high (75%), the power of the Wald test for nullity of the regression coefficient $\beta_2$ can be low, compared to the case where there are no immunes. But we note that for moderately large to large sample sizes ($n \geq 500$), the dispersion indicators and the power of the Wald test indicate good performance of the maximum likelihood estimate, even when the immune proportion is up to 50%. From Tables 3 and 4, the level of the Wald test for nullity of $\beta_2$ is globally respected except, for every immune proportion, when the sample size is small ($n = 100$).

We compare these results to the ones obtained from a "naive" method where: i) we consider every individual $i$ such that $\{Y_i = 0\}$ as being susceptible but uninfected, that is we ignore the eventual immunity of this individual, ii) we apply a usual logistic regression analysis to the resulting dataset. We hope that this comparison will convince the reader that it is necessary to take account of the immunity, when fitting a logistic regression model to a sample of binary data (arising from an infection setting, for example) with a cure fraction. The results of the "naive" analysis for model $M_1$ are given in Table 5 (the results for models $M_2, M_3, M_4$ yield similar observations and thus, they are not given here. However, the complete simulation study is available from a web-based supplementary material at the following address: http://perso.univ-lr.fr/jfdupuy/supplementary.pdf).

From this table, it appears that ignoring the immunity present in the sample results in strongly biased estimates of $\beta$. The bias of the intercept estimate increases with the immune proportion. At the same time, the estimate of the regression coefficient $\beta_2$ is biased towards 0 for all values of the immune percentage and sample size. This results in a very low power for the Wald test of nullity of $\beta_2$, and in a wrong interpretation of the relationship between the covariate $X_2$ and the binary response $Y$.

The quality of the gaussian approximation to the large-sample distribution of $\hat{\beta}_{2,n}$ was also investigated. For each configuration of the design parameters, histograms of the $\hat{\beta}_{2,n}^{(j)}$ ($j = 1, \ldots, N$) are obtained, along with the corresponding QQ-plots. These plots are pictured on Figures 1 to 16.

From these figures, it appears that the normal approximation stated in Theorem 2 is reasonably satisfied when the proportion of immunes is moderate (25%), provided that the sample size is sufficiently large ($n \geq 500$, say). Consider the case when $\beta_2 \neq 0$. When the immune fraction is large (50%), the normal approximation still appears reasonable, provided that the sample size is at least 1000, or eventually 1500. When the immune proportion is very large (75%), the distribution of $\hat{\beta}_{2,n}$ can be highly skewed, in particular when the sample size is small. Consider the case when $\beta_2 = 0$. Then the finite-sample distribution of $\hat{\beta}_{2,n}$ appears to be symmetric, with heavy tails however, especially when the sample size is small. When the immune fraction is about 50% and the sample size is greater than or equal to 500, the normal distribution appears to fit reasonably well the distribution of $\hat{\beta}_{2,n}$. 

Table 5 about here

Figures 1 to 16 about here
Overall, these results indicate that a reliable statistical inference on the regression effect in the model (1) with a cure fraction should be based on a sample having, at least, a moderately large size \((n \geq 500, \text{say})\) when the immune fraction is moderate (25%), or a large size \((n \geq 1000, \text{say})\) when the immune proportion is large (50%).

5. Discussion and perspectives

In this paper, we have considered the problem of estimating the logistic regression model from a sample of binary response data with a cure fraction. The estimator we propose is obtained by maximizing a likelihood function, which is derived from a joint regression model for the binary response of interest and the cure indicator, considered as a random variable whose distribution is modeled by a logistic regression. We have established the existence, consistency, and asymptotic normality of this estimator, and we have investigated its finite-sample properties via simulations.

Several open questions now deserve attention. The estimation approach proposed here relies on our ability to correctly specify the model for the binary immunity status. It is therefore of interest to investigate the effect of a misspecification of this model (and in particular, of the link function). The techniques and results by Czado and Santner (1992) may be useful for that purpose. Another issue of interest deals with the inference in the logistic regression model with a cure fraction, in a high-dimensional setting. We have established the theoretical properties of our estimator in a low-dimensional setting that is, when a small number of potential predictors are involved. Several recent contributions (see for example Huang et al. (2008) and Meier et al. (2008)) have considered the problem of estimation in the logistic model (without cure fraction) when the predictor dimension is much larger than the sample size (this problem arises, for example, in genetic studies where high-dimensional data are generated using microarray technologies). Extending our methodology to this setting constitutes another topic for further research.

Appendix

Proof of Lemma 1. Recall that \(I_k\) denotes the identity matrix of order \(k\). Then we write:

\[
\frac{\partial \phi_n(\psi)}{\partial \psi^r} = \|I_k - (WD(\psi_0)W')^{-1}WD(\psi)W'\| \\
= \| (WD(\psi_0)W')^{-1}W(D(\psi_0) - D(\psi))W'\| \\
\leq \| (WD(\psi_0)W')^{-1}\| \|W(D(\psi_0) - D(\psi))W'\| \\
= \lambda^{-1}_n \|W(D(\psi_0) - D(\psi))W'\|.
\]

Next, define \(S = \{(i,j) \in \{1, 2, \ldots, 2n\}^2 | D_{ij}(\psi_0) \neq 0\}\). Then the following holds:

\[
\|W(D(\psi_0) - D(\psi))W'\| = \left\| \sum_{i=1}^{2n} \sum_{j=1}^{2n} W_{ij}W'_{ij}(D_{ij}(\psi) - D_{ij}(\psi_0)) \right\| \\
\leq \sum_{(i,j) \in S} \|W_{ij}W'_{ij}D_{ij}(\psi_0)\| \left| \frac{D_{ij}(\psi) - D_{ij}(\psi_0)}{D_{ij}(\psi_0)} \right|.
\]
From C1 and C2, there exists a real constant $c_3 > 0$ such that $D_{ij}(\psi_0) > c_3$ for every $(i, j) \in S$. Moreover, $D_{ij}(\cdot)$ is uniformly continuous on $B \times G$, thus for every $\epsilon > 0$, there exists a positive $r$ such that for all $\psi \in B(\psi_0, r)$, $|D_{ij}(\psi) - D_{ij}(\psi_0)| < \epsilon$. It follows that

$$\|W(D(\psi_0) - D(\psi))W'\| \leq \frac{\epsilon}{c_3} \sum_{(i, j) \in S} \|W_iW_j' D_{ij}(\psi_0)\|$$

$$\leq \frac{\epsilon}{c_3} \text{tr} \left( \sum_{(i, j) \in S} W_iW_j' D_{ij}(\psi_0) \right)$$

$$= \frac{\epsilon}{c_3} \text{tr} \left( \sum_{i=1}^{2n} \sum_{j=1}^{2n} W_iW_j' D_{ij}(\psi_0) \right)$$

$$= \frac{\epsilon}{c_3} \text{tr} (WD(\psi_0)W')$$

$$\leq \frac{\epsilon}{c_3} \Lambda_n k.$$ 

This in turn implies that $\frac{\|\delta\phi_n(\psi)\|}{\delta \psi} \leq \frac{\epsilon \Lambda_n k}{c_3} < \frac{\epsilon c_2 k}{c_3}$. Now, choosing $\epsilon = c \frac{c_2 k}{c_3}$ with $0 < c < 1$, we get that $\frac{\|\delta\phi_n(\psi)\|}{\delta \psi} \leq c$ for all $\psi \in B(\psi_0, r)$, and the result follows.

References


Table 1. Simulation results for model $\mathcal{M}_1$: $\beta = (-0.8, 1)$

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
<td></td>
<td>25%</td>
<td></td>
<td>50%</td>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{1,n}$</td>
<td>$\hat{\beta}_{2,n}$</td>
<td>$\hat{\beta}_{1,n}$</td>
<td>$\hat{\beta}_{2,n}$</td>
<td>$\hat{\beta}_{1,n}$</td>
<td>$\hat{\beta}_{2,n}$</td>
<td>$\hat{\beta}_{1,n}$</td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>-0.834</td>
<td>1.064</td>
<td>-0.773</td>
<td>1.114</td>
<td>-0.787</td>
<td>1.137</td>
<td>-0.750</td>
</tr>
<tr>
<td></td>
<td>(0.258)</td>
<td>(0.301)</td>
<td>(0.583)</td>
<td>(0.412)</td>
<td>(0.825)</td>
<td>(0.603)</td>
<td>(0.921)</td>
</tr>
<tr>
<td></td>
<td>[0.202]</td>
<td>[0.232]</td>
<td>[0.465]</td>
<td>[0.324]</td>
<td>[0.657]</td>
<td>[0.440]</td>
<td>[0.784]</td>
</tr>
<tr>
<td></td>
<td>0.965*</td>
<td>0.109*</td>
<td>0.096*</td>
<td>0.121*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>-0.807</td>
<td>1.012</td>
<td>-0.783</td>
<td>1.111</td>
<td>-0.788</td>
<td>1.129</td>
<td>-0.791</td>
</tr>
<tr>
<td></td>
<td>(0.107)</td>
<td>(0.125)</td>
<td>(0.320)</td>
<td>(0.354)</td>
<td>(0.428)</td>
<td>(0.389)</td>
<td>(0.707)</td>
</tr>
<tr>
<td></td>
<td>[0.085]</td>
<td>[0.099]</td>
<td>[0.264]</td>
<td>[0.227]</td>
<td>[0.352]</td>
<td>[0.270]</td>
<td>[0.603]</td>
</tr>
<tr>
<td></td>
<td>1*</td>
<td>0.985*</td>
<td>0.85*</td>
<td>0.267*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>-0.801</td>
<td>1.004</td>
<td>-0.794</td>
<td>1.058</td>
<td>-0.798</td>
<td>1.060</td>
<td>-0.797</td>
</tr>
<tr>
<td></td>
<td>(0.077)</td>
<td>(0.085)</td>
<td>(0.241)</td>
<td>(0.202)</td>
<td>(0.310)</td>
<td>(0.247)</td>
<td>(0.683)</td>
</tr>
<tr>
<td></td>
<td>[0.062]</td>
<td>[0.068]</td>
<td>[0.201]</td>
<td>[0.147]</td>
<td>[0.253]</td>
<td>[0.178]</td>
<td>[0.569]</td>
</tr>
<tr>
<td></td>
<td>1*</td>
<td>1*</td>
<td>1*</td>
<td>0.567*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1500</td>
<td>-0.805</td>
<td>1.003</td>
<td>-0.801</td>
<td>1.040</td>
<td>-0.799</td>
<td>1.040</td>
<td>-0.802</td>
</tr>
<tr>
<td></td>
<td>(0.061)</td>
<td>(0.074)</td>
<td>(0.210)</td>
<td>(0.159)</td>
<td>(0.277)</td>
<td>(0.191)</td>
<td>(0.600)</td>
</tr>
<tr>
<td></td>
<td>[0.048]</td>
<td>[0.059]</td>
<td>[0.176]</td>
<td>[0.119]</td>
<td>[0.228]</td>
<td>[0.141]</td>
<td>[0.493]</td>
</tr>
<tr>
<td></td>
<td>1*</td>
<td>1*</td>
<td>1*</td>
<td>0.861*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: $n$: sample size. (·): root mean square error. [·]: mean absolute error. *: empirical power 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.
### Table 2. Simulation results for model $M_2$: $\beta = (1.7)$

<table>
<thead>
<tr>
<th>n</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.026</td>
<td>0.720</td>
<td>0.945</td>
<td>0.723</td>
<td>0.949</td>
<td>0.740</td>
<td>0.834</td>
<td>0.647</td>
</tr>
<tr>
<td></td>
<td>(0.246)</td>
<td>(0.273)</td>
<td>(0.780)</td>
<td>(0.534)</td>
<td>(0.988)</td>
<td>(0.788)</td>
<td>(1.549)</td>
<td>(1.455)</td>
</tr>
<tr>
<td></td>
<td>[0.196]</td>
<td>[0.215]</td>
<td>[0.655]</td>
<td>[0.376]</td>
<td>[0.829]</td>
<td>[0.555]</td>
<td>[1.326]</td>
<td>[0.933]</td>
</tr>
<tr>
<td></td>
<td>0.746*</td>
<td>0.132*</td>
<td>0.118*</td>
<td>0.088*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>1.003</td>
<td>0.712</td>
<td>1.098</td>
<td>0.717</td>
<td>1.112</td>
<td>0.721</td>
<td>0.840</td>
<td>0.652</td>
</tr>
<tr>
<td></td>
<td>(0.107)</td>
<td>(0.115)</td>
<td>(0.651)</td>
<td>(0.247)</td>
<td>(0.672)</td>
<td>(0.279)</td>
<td>(0.969)</td>
<td>(0.534)</td>
</tr>
<tr>
<td></td>
<td>[0.086]</td>
<td>[0.091]</td>
<td>[0.518]</td>
<td>[0.202]</td>
<td>[0.534]</td>
<td>[0.230]</td>
<td>[0.802]</td>
<td>[0.421]</td>
</tr>
<tr>
<td></td>
<td>1*</td>
<td>0.503*</td>
<td>0.418*</td>
<td>0.168*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1.003</td>
<td>0.707</td>
<td>1.078</td>
<td>0.711</td>
<td>1.096</td>
<td>0.719</td>
<td>0.842</td>
<td>0.657</td>
</tr>
<tr>
<td></td>
<td>(0.071)</td>
<td>(0.082)</td>
<td>(0.590)</td>
<td>(0.215)</td>
<td>(0.571)</td>
<td>(0.224)</td>
<td>(0.796)</td>
<td>(0.439)</td>
</tr>
<tr>
<td></td>
<td>[0.057]</td>
<td>[0.065]</td>
<td>[0.428]</td>
<td>[0.168]</td>
<td>[0.441]</td>
<td>[0.181]</td>
<td>[0.670]</td>
<td>[0.352]</td>
</tr>
<tr>
<td></td>
<td>1*</td>
<td>0.779*</td>
<td>0.675*</td>
<td>0.205*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1500</td>
<td>1.001</td>
<td>0.701</td>
<td>1.035</td>
<td>0.705</td>
<td>1.069</td>
<td>0.709</td>
<td>0.887</td>
<td>0.655</td>
</tr>
<tr>
<td></td>
<td>(0.064)</td>
<td>(0.065)</td>
<td>(0.450)</td>
<td>(0.163)</td>
<td>(0.466)</td>
<td>(0.177)</td>
<td>(0.604)</td>
<td>(0.312)</td>
</tr>
<tr>
<td></td>
<td>[0.050]</td>
<td>[0.052]</td>
<td>[0.344]</td>
<td>[0.135]</td>
<td>[0.358]</td>
<td>[0.144]</td>
<td>[0.502]</td>
<td>[0.257]</td>
</tr>
<tr>
<td></td>
<td>1*</td>
<td>0.986*</td>
<td>0.926*</td>
<td>0.300*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** *: empirical power 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 70%.
Table 3. Simulation results for model $M_3$: $\beta = (-0.8, 0)$

<table>
<thead>
<tr>
<th>n</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-0.815</td>
<td>-0.001</td>
<td>-0.721</td>
<td>-0.007</td>
<td>-0.734</td>
<td>0.000</td>
<td>-0.746</td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td>(0.224)</td>
<td>(0.229)</td>
<td>(0.465)</td>
<td>(1.341)</td>
<td>(0.800)</td>
<td>(2.109)</td>
<td>(1.966)</td>
<td>(3.258)</td>
</tr>
<tr>
<td></td>
<td>[0.177]</td>
<td>[0.179]</td>
<td>[0.377]</td>
<td>[0.762]</td>
<td>[0.636]</td>
<td>[1.111]</td>
<td>[1.516]</td>
<td>[1.715]</td>
</tr>
<tr>
<td></td>
<td>0.052*</td>
<td>0.077*</td>
<td>0.069*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>-0.801</td>
<td>-0.001</td>
<td>-0.748</td>
<td>0.007</td>
<td>-0.750</td>
<td>0.001</td>
<td>-0.775</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>(0.097)</td>
<td>(0.099)</td>
<td>(0.280)</td>
<td>(0.415)</td>
<td>(0.520)</td>
<td>(0.469)</td>
<td>(1.209)</td>
<td>(0.711)</td>
</tr>
<tr>
<td></td>
<td>[0.078]</td>
<td>[0.080]</td>
<td>[0.241]</td>
<td>[0.231]</td>
<td>[0.422]</td>
<td>[0.241]</td>
<td>[1.007]</td>
<td>[0.363]</td>
</tr>
<tr>
<td></td>
<td>0.041*</td>
<td>0.058*</td>
<td>0.052*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>-0.803</td>
<td>-0.001</td>
<td>-0.759</td>
<td>0.008</td>
<td>-0.763</td>
<td>0.005</td>
<td>-0.793</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>(0.067)</td>
<td>(0.066)</td>
<td>(0.221)</td>
<td>(0.237)</td>
<td>(0.367)</td>
<td>(0.266)</td>
<td>(1.154)</td>
<td>(0.312)</td>
</tr>
<tr>
<td></td>
<td>[0.053]</td>
<td>[0.053]</td>
<td>[0.182]</td>
<td>[0.137]</td>
<td>[0.299]</td>
<td>[0.140]</td>
<td>[0.911]</td>
<td>[0.175]</td>
</tr>
<tr>
<td></td>
<td>0.042*</td>
<td>0.045*</td>
<td>0.037*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1500</td>
<td>-0.801</td>
<td>0.000</td>
<td>-0.782</td>
<td>0.009</td>
<td>-0.784</td>
<td>0.003</td>
<td>-0.783</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>(0.053)</td>
<td>(0.054)</td>
<td>(0.208)</td>
<td>(0.168)</td>
<td>(0.328)</td>
<td>(0.212)</td>
<td>(1.149)</td>
<td>(0.258)</td>
</tr>
<tr>
<td></td>
<td>[0.042]</td>
<td>[0.043]</td>
<td>[0.178]</td>
<td>[0.099]</td>
<td>[0.267]</td>
<td>[0.102]</td>
<td>[0.901]</td>
<td>[0.144]</td>
</tr>
<tr>
<td></td>
<td>0.051*</td>
<td>0.048*</td>
<td>0.027*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *: 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%.
Table 4. Simulation results for model $M_4$: $\beta = (1, 0)$

<table>
<thead>
<tr>
<th>n</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
<th>$\hat{\beta}_{1,n}$</th>
<th>$\hat{\beta}_{2,n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.030</td>
<td>0.001</td>
<td>1.110</td>
<td>0.007</td>
<td>1.154</td>
<td>0.017</td>
<td>0.913</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.233)</td>
<td>(0.234)</td>
<td>(0.852)</td>
<td>(0.969)</td>
<td>(1.211)</td>
<td>(1.347)</td>
<td>(1.775)</td>
<td>(1.640)</td>
</tr>
<tr>
<td></td>
<td>[0.182]</td>
<td>[0.187]</td>
<td>[0.684]</td>
<td>[0.587]</td>
<td>[0.995]</td>
<td>[0.792]</td>
<td>[1.450]</td>
<td>[0.865]</td>
</tr>
<tr>
<td></td>
<td>0.058*</td>
<td>0.072*</td>
<td>0.083*</td>
<td></td>
<td>0.066*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>1.007</td>
<td>-0.005</td>
<td>1.105</td>
<td>0.020</td>
<td>1.123</td>
<td>0.054</td>
<td>0.915</td>
<td>-0.009</td>
</tr>
<tr>
<td></td>
<td>(0.103)</td>
<td>(0.103)</td>
<td>(0.609)</td>
<td>(0.293)</td>
<td>(0.690)</td>
<td>(0.318)</td>
<td>(0.817)</td>
<td>(0.370)</td>
</tr>
<tr>
<td></td>
<td>[0.081]</td>
<td>[0.082]</td>
<td>[0.492]</td>
<td>[0.180]</td>
<td>[0.562]</td>
<td>[0.208]</td>
<td>[0.614]</td>
<td>[0.215]</td>
</tr>
<tr>
<td></td>
<td>0.046*</td>
<td>0.050*</td>
<td>0.063*</td>
<td></td>
<td>0.051*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1.003</td>
<td>0.000</td>
<td>1.091</td>
<td>-0.003</td>
<td>1.101</td>
<td>0.033</td>
<td>0.934</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.071)</td>
<td>(0.070)</td>
<td>(0.521)</td>
<td>(0.198)</td>
<td>(0.578)</td>
<td>(0.210)</td>
<td>(0.757)</td>
<td>(0.256)</td>
</tr>
<tr>
<td></td>
<td>[0.057]</td>
<td>[0.055]</td>
<td>[0.437]</td>
<td>[0.125]</td>
<td>[0.455]</td>
<td>[0.135]</td>
<td>[0.600]</td>
<td>[0.142]</td>
</tr>
<tr>
<td></td>
<td>0.051*</td>
<td>0.045*</td>
<td>0.042*</td>
<td></td>
<td></td>
<td>0.039*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1500</td>
<td>1.003</td>
<td>0.001</td>
<td>1.073</td>
<td>0.009</td>
<td>1.115</td>
<td>0.015</td>
<td>0.934</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>(0.057)</td>
<td>(0.057)</td>
<td>(0.480)</td>
<td>(0.132)</td>
<td>(0.501)</td>
<td>(0.139)</td>
<td>(0.633)</td>
<td>(0.175)</td>
</tr>
<tr>
<td></td>
<td>[0.046]</td>
<td>[0.046]</td>
<td>[0.392]</td>
<td>[0.087]</td>
<td>[0.400]</td>
<td>[0.104]</td>
<td>[0.521]</td>
<td>[0.109]</td>
</tr>
<tr>
<td></td>
<td>0.042*</td>
<td>0.040*</td>
<td>0.046*</td>
<td></td>
<td></td>
<td>0.047*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *: 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 70%.
Table 5. "Naive" analysis of model $M_1$: $\beta = (-.8, 1)$

<table>
<thead>
<tr>
<th>Percentage of immunes in the sample</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{1,n}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{2,n}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{1,n}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{2,n}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{1,n}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{2,n}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100 -1.154</td>
<td>0.023</td>
<td>-1.632</td>
<td>0.017</td>
</tr>
<tr>
<td>(0.428)</td>
<td>(1.011)</td>
<td>(0.879)</td>
<td>(1.025)</td>
</tr>
<tr>
<td>[0.365]</td>
<td>[0.977]</td>
<td>[0.833]</td>
<td>[0.983]</td>
</tr>
<tr>
<td>0.049*</td>
<td>0.057*</td>
<td>0.052*</td>
<td></td>
</tr>
<tr>
<td>500 -1.128</td>
<td>0.087</td>
<td>-1.594</td>
<td>0.042</td>
</tr>
<tr>
<td>(0.344)</td>
<td>(0.915)</td>
<td>(0.803)</td>
<td>(0.963)</td>
</tr>
<tr>
<td>[0.328]</td>
<td>[0.913]</td>
<td>[0.794]</td>
<td>[0.958]</td>
</tr>
<tr>
<td>0.049*</td>
<td>0.051*</td>
<td>0.053*</td>
<td></td>
</tr>
<tr>
<td>1000 -1.131</td>
<td>0.059</td>
<td>-1.590</td>
<td>0.050</td>
</tr>
<tr>
<td>(0.338)</td>
<td>(0.941)</td>
<td>(0.795)</td>
<td>(0.952)</td>
</tr>
<tr>
<td>[0.330]</td>
<td>[0.940]</td>
<td>[0.790]</td>
<td>[0.950]</td>
</tr>
<tr>
<td>0.053*</td>
<td>0.051*</td>
<td>0.054*</td>
<td></td>
</tr>
<tr>
<td>1500 -1.127</td>
<td>0.050</td>
<td>-1.591</td>
<td>0.046</td>
</tr>
<tr>
<td>(0.332)</td>
<td>(0.953)</td>
<td>(0.794)</td>
<td>(0.955)</td>
</tr>
<tr>
<td>[0.327]</td>
<td>[0.952]</td>
<td>[0.791]</td>
<td>[0.954]</td>
</tr>
<tr>
<td>0.051*</td>
<td>0.050*</td>
<td>0.053*</td>
<td></td>
</tr>
</tbody>
</table>

Note: *: empirical power 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%. In the "naive" analysis, every uninfected individual (i.e. $Y_i = 0$) is considered as susceptible.
Figure 1. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $M_1$, with no immunes in the sample (the percentage of immunes is given in brackets). $n$ is the sample size. All results are based on 1500 simulated datasets.

Figure 2. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $M_1$, with 25% of immunes.
Maximum likelihood estimation in the logistic regression model with a cure fraction

Figure 3. Histograms and Q-Q plots for $\hat{\beta}_2, n_{100}$ in model $M_1$, with 50% of immunes.

Figure 4. Histograms and Q-Q plots for $\hat{\beta}_2, n_{100}$ in model $M_1$, with 75% of immunes.
Figure 5. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $\mathcal{M}_2$, with no immunes in the sample.

Figure 6. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $\mathcal{M}_2$, with 25% of immunes.
Maximum likelihood estimation in the logistic regression model with a cure fraction

Figure 7. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $M_2$, with 50% of immunes.

Figure 8. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $M_2$, with 75% of immunes.
Figure 9. Histograms and Q-Q plots for $\hat{\beta}_{\cdot n}$ in model $M_3$, with no immunes in the sample.

Figure 10. Histograms and Q-Q plots for $\hat{\beta}_{\cdot n}$ in model $M_3$, with 25% of immunes.
Maximum likelihood estimation in the logistic regression model with a cure fraction

Figure 11. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $M_3$, with 50% of immunes.

Figure 12. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $M_3$, with 75% of immunes.
Figure 13. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $\mathcal{M}_4$, with no immunes in the sample.

Figure 14. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $\mathcal{M}_4$, with 25% of immunes.
Maximum likelihood estimation in the logistic regression model with a cure fraction

Figure 15. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $M_4$, with 50% of immunes.

Figure 16. Histograms and Q-Q plots for $\hat{\beta}_{2,n}$ in model $M_4$, with 75% of immunes.