THÈSE

En cotutelle, présentée pour obtenir
LE GRADE DE DOCTEUR

DE L'UNIVERSITÉ GASTON BERGER
ET DE L'UNIVERSITÉ DE LA ROCHELLE

Spécialité : MATHEMATIQUES APPLIQUÉES

Par **Aba DIOP**

Inférence statistique dans le modèle de régression logistique avec fraction immune

Soutenue publiquement le 15 Novembre 2012 devant le jury :

Emmanuelle AUGERAUD-VERON	**Maître de conférences HDR, Université La Rochelle**	**Examinatrice**
Laurence BARIL	**Médecin PhD, GSK Belgique**	**Examinatrice**
Sophie DABO-NIANG	**Professeur, Université Charles de Gaulle Lille 3**	**Examinatrice**
Gane Samb LO	**Professeur, Université Gaston Berger**	**Examinateur**
Avner BAR-HEN	**Professeur, Université Paris Descartes**	**Rapporteur**
Philippe BESSE	**Professeur, INSA de Toulouse**	**Rapporteur**
Aliou DIOP	**Professeur, Université Gaston Berger**	**Directeur de Thèse**
Jean-François DUPUY	**Professeur, INSA de Rennes**	**Directeur de Thèse**

Thèse préparée au Laboratoire d'Études et de Recherches en Statistique et Développement et au Laboratoire Mathématiques, Image et Applications

A la mémoire de mon père Feu Moussa DIOP,

A ma mère Fatimata KEITA,

A ma famille, à mes amis et à tous ceux qui me sont chers.

Remerciements

J'adresse ces quelques mots de remerciements à l'endroit de toutes les personnes qui ont contribué à la réalisation de ce travail.

Tout d'abord, je tiens à remercier particulièrement les professeurs Aliou DIOP et Jean-François DUPUY pour avoir accepté de diriger ma thèse, pour avoir toujours été disponibles pendant tout mon travail et pour n'avoir ménagé aucun effort pour sa réalisation.

J'adresse mes plus sincères remerciements à mes rapporteurs, les professeurs Philippe BESSE et Avner BAR-HEN, pour avoir accepté de lire cette thèse et m'avoir fait bénéficier de leurs pertinentes remarques et suggestions. Je remercie aussi les professeurs Emmanuelle AUGERAUD-VERON, Sophie DABO-NIANG et Gane Samb LO d'avoir accepté de faire partie de mon jury de thèse. Mes sincères remerciements vont à l'endroit du Docteur Laurence BARIL pour non seulement avoir accepté de faire partie de mon jury mais aussi pour m'avoir fait bénéficier de sa large connaissance en épidémiologie et pour m'avoir également permis d'acquérir des données réelles sur lesquelles cette thèse a été appliquée.

Je remercie particulièrement le Service de Coopération et d'Actions Culturelles de l'Ambassade de France au Sénégal, d'AIRES-SUD et STAFAV pour avoir participé grandement au financement de cette thèse.

Je voudrais remercier tout le personnel enseignant et administratif de l'UFR Sciences Appliquées et Technologie de l'Université Gaston Berger. Mes remerciements vont également à l'endroit de tous les membres du Laboratoire Mathématiques Image et Applications de l'Université de La Rochelle pour leur accueil chaleureux au cours

des séjours passés au sein du laboratoire, ainsi que la secrétaire de l'école doctorale S2I Jennifer de la Corté GOMEZ. Je remercie aussi les membres de l'INSA du centre de mathématiques de Rennes.
Je remercie tous les membres de l'Unité d'Épidémiologie des Maladies Infectieuses de l'Institut Pasteur de Dakar et je témoigne ma reconnaissance particulière au Docteur Fatoumata Diene Sarr.

Un grand grand MERCI à l'endroit de mes collègues doctorants et docteurs du Laboratoire d'Études et de Recherches en Statistique et Développement pour leur soutien moral, pour les échanges fructueux lors des séminaires et discussions et avec qui j'ai partagé mes angoisses, mes rêves et des moments inoubliables.

Je dédie une mention spéciale et toute particulière à mes parents en particulier à ma très chère mère surtout pour sa patience, à ma famille et à mes amis les plus proches. Un grand merci à mon cher beau frère Mamadou THIAM, à mon cher tons Cheikh Makhtar Keïta, à mes adorables et aimables soeurs Touti DIOP, Diary DIOP, Magatte DIOP et Ndéye Meïssa DIOP , à mes chers frères Pape Amadou DIOP et Bayla DIOP et enfin à ma chère bien-aimé Awa Ly DIOUM.

Il m'est bien sûr impossible de terminer ces remerciements sans avoir une pensée à tous mes camarades de promotion 2008 de DEA STAFAV.

Table des matières

Résumé

Les modèles linéaires généralisés sont une généralisation des modèles de régression linéaire, et sont très utilisés dans le domaine du vivant. Le modèle de régression logistique, l'un des modèles de cette classe, très souvent utilisé dans les études biomédicales demeure le modèle de régression le plus approprié quand il s'agit de modéliser une variable discrète de nature binaire. Dans cette thèse, nous nous intéressons au problème de l'inférence statistique dans le modèle de régression logistique, en présence d'individus immunes dans la population d'étude.

Dans un premier temps, nous considérons le problème de l'estimation dans le modèle de régression logistique en présence d'individus immunes, qui entre dans le cadre des modèles de régression à excès de zéros (ou zéro-inflatés). Un individu est dit immune s'il n'est pas exposé à l'événement d'intérêt. Le statut d'immunité est inconnu sauf si l'événement d'intérêt a été observé. Nous développons une méthode d'estimation par maximum de vraisemblance en proposant une modélisation conjointe de l'immunité et des risques d'infection. Nous établissons d'abord l'identifiabilité du modèle proposé. Puis, nous montrons l'existence de l'estimateur du maximum de vraisemblance des paramètres de ce modèle. Nous montrons ensuite, la consistance de cet estimateur, et nous établissons sa normalité asymptotique. Enfin, nous étudions, au moyen de simulations, leur comportement sur des échantillons de taille finie.

Dans un deuxième temps, nous nous intéressons à la construction de bandes de confiance simultanées pour la probabilité d'infection, dans le modèle de régression logistique avec fraction immune. Nous proposons trois méthodes de constructions de

bandes de confiance pour la fonction de régression. La première méthode (méthode de Scheffé) utilise la propriété de normalité asymptotique de l'estimateur du maximum de vraisemblance, et une approximation par une loi du khi deux pour approcher le quantile nécessaire à la construction des bandes. La deuxième méthode utilise également la propriété de normalité asymptotique de l'estimateur du maximum de vraisemblance et est basée sur une égalité classique de (Landau & Sheep 1970). La troisième méthode (méthode bootstrap) repose sur des simulations, pour estimer le quantile approprié de la loi du supremum d'un processus gaussien. Enfin, nous évaluons, au moyen de simulations, leurs propriétés sur des échantillons de taille finie.

Enfin, nous appliquons les résultats de modélisation à des données réelles sur la dengue. Il s'agit d'une maladie vectorielle tropicale à transmission strictement inter-humaine. Les résultats montrent que les probabilités d'infection estimées à partir de notre approche de modélisation sont plus élevées que celles obtenues à partir d'un modèle de régression logistique standard qui ne tient pas compte d'une possible immunité. En particulier, les estimations fournies par notre approche suggèrent que le sous-poids constitue un facteur de risque majeur de l'infection par la dengue, indépendamment de l'âge.

Mots clés

Régression logistique, Fraction immune, Maximum de vraisemblance, Identifiabilité, Consistance, Normalité asymptotique, Bandes de confiance, Bootstrap, Simulations, Dengue.

Abstract

Generalized linear models are a generalization of linear regression models, and are widely used in the field of life. The logistic regression model, one of this class of models, widely used in biomedical studies remains the most appropriate regression model when it comes to model discrete variable, binary in nature. In this thesis, we investigate the problem of statistical inference in the logistic regression model, in the presence of immune individuals in the study population.

At first, we consider the problem of estimation in the logistic regression model in the presence of immune individuals that enters in the case of zero-inflated regression models. A subject is said to be immune if he cannot experience the outcome of interest. The immune status is unknown unless the event of interest has been observed. We develop a maximum likelihood estimation procedure for this problem, based on the joint modeling of the binary response of interest and the cure status. We investigate the identifiability of the resulting model. Then, we establish the existence, consistency and asymptotic normality of the proposed estimator, and we conduct a simulation study to investigate its finite-sample behavior.

In a second time, we focus on the construction of simultaneous confidence bands for the probability of infection in the logistic regression model with immune fraction. We propose three methods of construction of confidence bands for the regression function. The first method (Scheffe's method) uses the asymptotic normality of the maximum likelihood estimator, and an approximation by the chi-squared distribution to approximate the necessary quantile for the construction of bands. The second method uses also the asymptotic normality of the maximum likelihood es-

timator and is based on a classical equality by (Landau & Sheep 1970). The third method (bootstrap method) is based on simulations, to estimate the appropriate quantile of the law of a supremum of a Gaussian process. Finally, we conduct a simulation study to investigate its finite-sample properties.

Finally, we consider a study of dengue fever, which is a tropical mosquito-borne viral human disease, strictly inter-human. The results show that, the estimated probabilities of infection obtained from our approach are larger than the ones derived from a standard analysis that does not take account of the possible immunity. In particular, the estimates provided by our approach suggest that underweight constitutes a major risk factor for dengue infection, irrespectively of age.

Key words

Logistic regression, Immune fraction, Maximum likelihood, Identifiability, Consistancy, Asymptotic normality, Confidance Bands, Bootstrap, Simulations, Dengue.

Abréviations et Notations

$\mathbf{P}(A)$	: La probabilité de l'événement A.
$\mathbb{E}(X)$	: L'espérance mathématique de la variable aléatoire X.
$\mathrm{var}(X)$	: La variance de la variable aléatoire X.
$\mathrm{cov}(X, Y)$	: La covariance des variables aléatoires X et Y.
$X_n \Longrightarrow Y$	: La suite de variables aléatoires $(X_n)_{n\geq 0}$ converge faiblement vers Y.
$X_n \xrightarrow{p.s} Y$	: La suite de variables aléatoires $(X_n)_{n\geq 0}$ converge presque sûrement vers Y.
$\mathbb{N}$	: Ensemble des entiers naturels.
$\mathbb{N}^*$	: Ensemble des entiers naturels non nuls.
$\mathbb{R}$	: Ensemble des réels et $\mathbb{R}^d = \underbrace{\mathbb{R} \times \ldots \times \mathbb{R}}_{d \text{ fois}}$.
$X^\top$	: Transposée du vecteur X.
i.i.d.	: indépendantes et identiquement distribuées.
$\mathscr{M}(n \times p)$	: L'ensemble des matrices réelles de n lignes et p colonnes.
I_p	: matrice identité d'ordre p.
EMV	: Estimateur du maximum de vraisemblance.
ZIB	: Zero-Inflated Binomial.
ZIP	: Zero-Inflated Poisson.
ZINB	: Zero-Inflated Negative Binomial.
ZIPO	: Zero-Inflated Proportional Odds.

Chapitre 1

Introduction générale

L'analyse statistique des données constitue aujourd'hui un outil fondamental dans le domaine du vivant, et repose très souvent sur l'utilisation de modèles de régression. L'une de ces classes de modèles, parmi les plus utilisées, demeure la classe des modèles linéaires généralisés qui sont une généralisation bien connue des modèles de régression linéaire en termes de loi. Elle permet l'analyse de données discrètes mais aussi de données continues pour lesquelles la loi normale n'est pas très adaptée. (McCullagh & Nelder 1989) en font une présentation détaillée.

L'utilisation des modèles linéaires généralisés est très courante en santé publique pour la résolution de problèmes réels. En particulier, les épidémiologistes, lors de l'exploitation de données médicales, décrivent souvent un événement ou un phénomène lui-même influencé par la survenue d'autres événements ou phénomènes appelés facteurs d'exposition. La survenue d'une infection représentée par une variable dichotomique (*1 pour infecté et 0 pour non infecté*) peut en effet être ce phénomène étudié par les épidémiologistes en présence de certains facteurs d'exposition. Cependant il peut exister dans la population d'étude une proportion d'individus non susceptibles à l'infection qui peut être du à une immunité naturelle ou à une action préventive. En effet,l'immunité innée constitue la première ligne de défense et d'interaction entre l'hôte humain et les pathogènes. Ce phénomène entraîne ainsi la présence de beaucoup de zéros dans la variable dichotomique ; ce qui entraîne l'apparition d'un biais net sur l'estimation des paramètres du modèle et qui pourrait ainsi être à l'origine d'une mauvaise interprétation des résultats attendus. Ce phénomène connu sous le nom de "*zero-inflated*" a été à l'origine de l'extension des

modèles de régression de comptage de Poison et Binomiale Négative, du modèle de régression Binomiale et du modèle de régression à Odds Proportionnels.
Le modèle de régression logistique standard est naturellement utilisé en épidémiologie lorsqu'il s'agit de modéliser une variable dépendante binaire. L'objectif dans cette thèse est d'étendre ce modèle classique en prenant en compte la présence de cette fraction immune dans la population d'étude tout en proposant une modélisation conjointe de la probabilité d'infection et de la probabilité d'immunité.

Le premier chapitre de ce manuscrit est consacré à de brefs rappels sur le modèle linéaire, et à des rappels détaillés sur le modèle de régression logistique. Nous nous intéressons en particulier à l'identifiabilité de ce modèle, et à l'estimation de ses paramètres. Puis, nous rappelons les résultats d'existence d'estimateurs du maximum de vraisemblance dans ces modèles. Dans le chapitre 2, nous définissons les modèles zéro-inflatés existant : ZIP, ZINB, ZIB et ZIPO. Puis, nous présentons les résultats d'existence d'estimateur du maximum de vraisemblance et les propriétés asymptotiques de ces estimateurs.

Les parties 2 et 3 de ce manuscrit contiennent les contributions originales de cette thèse. Dans la deuxième partie de cette thèse, nous nous intéressons au problème de l'inférence statistique dans le modèle de régression logistique avec fraction immune. Dans le chapitre 3, nous commençons par décrire le modèle de régression logistique avec fraction immune, qui entre dans la famille des modèles zéro-inflatés, puis, nous proposons des estimateurs par maximum de vraisemblance pour ce modèle. Nous présentons enfin, les résultats asymptotiques obtenus pour ces estimateurs du maximum de vraisemblance. Dans le chapitre 4, nous utiliserons ces résultats asymptotiques pour construire des bandes de confiance simultanées pour la probabilité d'infection. Dans ce chapitre, nous présentons trois méthodes pour la construction des bandes de confiance pour la fonction réponse.

Dans la troisième partie de cette thèse, nous appliquons le modèle conjoint défini dans le chapitre 3 sur un jeu de données réelles. Nous commençons par une étude épidémiologique de la dengue dans le chapitre 5, où nous détaillons les caractéristiques de cette infection virale et présentons des résultats d'analyses statistiques univariées et multivariées. Dans le chapitre 6, nous appliquons le modèle conjoint à l'analyse statistique des données de dengue.

La dernière partie, est constituée des différentes annexes : des rappels mathématiques, des résultats complémentaires concernant certains lemmes énoncés dans la thèse, de tableaux des résultats de simulation.

Première partie

Rappels sur quelques modèles

Chapitre 2

Rappels sur le modèle de régression logistique

Sommaire

Notons y le vecteur des observations de taille n, réalisation du vecteur aléatoire Y, variable à expliquer. Un modèle linéaire généralisé se caractérise par les trois hypothèses suivantes :

– On suppose que les composantes Y_i $(i = 1, ..., n)$ de Y sont indépendantes et distribuées selon une loi appartenant à la famille exponentielle au sens de (Nelder & Wedderburn 1972), c'est-à-dire que la fonction de densité de la variable aléatoire Y_i s'écrit :

$$f_{Y_i}(y_i, \theta_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

où θ_i est un paramètre canonique et ϕ un paramètre de dispersion. Les fonctions b et c sont spécifiques à chaque distribution et la fonction a_i est donnée par : $a_i(\phi) = \frac{\phi}{w_i}$, où w_i est un poids connu associé à la réalisation y_i.

L'espérance et la variance de la variable Y_i s'exprime comme suit à l'aide des fonctions a_i et b :

$$\mathbb{E}(Y_i) = b'(\theta_i), \quad \text{var}(Y_i) = a_i(\phi)b''(\theta_i).$$

– Le prédicteur linéaire est définit comme suit :

$$\eta = \beta^\top \mathbf{X},$$

où β est un vecteur de paramètres inconnus de taille p et $\mathbf{X}$ le vecteur des p variables explicatives.

– Le lien entre l'espérance de Y_i et la i^{me} composante du prédicteur linéaire est réalisé par la fonction g (monotone et différentiable) appelée *fonction de lien* :

$$\eta_i = g(\mathbb{E}(Y_i)).$$

Une fonction de lien pour laquelle $\eta_i = \theta_i$ est appelée *fonction de lien canonique.*

Les modèles gaussiens (linéaires ou non) à variance constante ($a(\phi) = \sigma^2$), les modèles logistiques binaires ou polytomiques, les modèles de régression Poisson ou Gamma sont des exemples de modèles linéaires généralisés. Parmi ces modèles nous nous intéresserons particulièrement dans la suite au modèle de régression logistique.

2.1 Le modèle linéaire

Le modèle linéaire est souvent le premier outil de statistique inférentielle mis en oeuvre. Il permet d'expliquer une variable Y (appelée variable dépendante) par p variables explicatives $\mathbf{X} = (X_1, X_2, ..., X_p)$. Pour ce faire, nous disposons de n réalisations indépendantes $(X_1, Y_1), \ldots, (X_n, Y_n)$ du couple $(\mathbf{X}, Y)$. Le but est de modéliser la dépendance de la variable réponse Y par rapport aux variables explicatives $X_1, X_2, \ldots, X_p$. Des études diverses et détaillées ont été effectuées sur le modèle linéaire. Pour plus de détails on peut se référer à (Azaïs & Bardet 2006).
Le modèle linéaire standard traduit la dépendance linéaire de l'espérance en $\beta = (\beta_1, \ldots, \beta_p)^\top$, un paramètre inconnu non-contraint de $\mathbb{R}^p$ et il s'écrit :

$$Y_i = \beta^\top \mathbf{X}_i + \varepsilon_i, \quad i = 1, ..., n \tag{2.1}$$

Le modèle linéaire est caractérisé par les hypothèses suivantes :

- $\mathbb{E}(\varepsilon_i | \mathbf{X}_i) = 0$, pour $i = 1, ..., n$ (les erreurs sont centrées) ;

- $\mathrm{var}(\varepsilon_i) = \sigma^2$, pour $i = 1, ..., n$ (la variance des erreurs est constante : hypothèse d'homoscédasticité) ;

- les variables $\varepsilon_i, i = 1, ..., n$ sont indépendantes et de loi gaussienne.

Plaçons nous maintenant dans le cas où la variable à expliquer Y est qualitative ou de type facteur (sexe, couleur, présence ou absence d'une maladie...). Cette variable possède un nombre fini de modalités $g_1, \ldots, g_m$. Le problème consiste à expliquer l'appartenance d'un individu à un groupe à partir des p variables explicatives, on parlera alors de discrimination au lieu de régression.

Il est bien entendu impossible de modéliser directement la variable Y par une relation linéaire (imaginons que Y soit le sexe d'une personne ou son état de santé).

Afin de pallier à cette difficulté, on va s'intéresser aux probabilités $\mathbf{P}(Y = g_k|\mathbf{X} = x)$. Supposons pour simplifier que la variable Y prenne uniquement deux valeurs : 0 (groupe 0) ou 1 (groupe 1). La connaissance de $\mathbf{P}(Y = 1|\mathbf{X} = x)$ implique celle de $\mathbf{P}(Y = 0|\mathbf{X} = x)$: il suffit par conséquent de modéliser la probabilité $p(x) = \mathbf{P}(Y = 1|\mathbf{X} = x)$. On peut par exemple envisager une relation de la forme

$$p(x) = \beta_0 + \beta_1 x_1 +, \ldots, +\beta_p x_p = \beta^\top x. \tag{2.2}$$

Cette approche possède plusieurs caractéristiques :

- Remarquons tout d'abord que la variance de $Y|\mathbf{X} = x$ vaut $p(x)(1 - p(x))$. Contrairement au modèle linéaire classique, cette variance n'est pas constante et par conséquent l'hypothèse classique d'homoscédasticité des résidus ne sera pas vérifiée.

- Le fait qu'aucune restriction ne soit effectuée sur les β implique que $\beta^\top x$ peut prendre n'importe quelle valeur sur $\mathbb{R}$. Ce qui peut être gênant pour l'estimation d'une probabilité.

Pour ces raisons, nous devons étendre le modèle linéaire classique aux cas où :
- Y peut être une variable qualitative (par exemple, présence ou absence d'une maladie, appartenance à une catégorie...) ;

- les erreurs peuvent ne pas avoir la même variance (s'affranchir de l'hypothèse d'homoscédasticité).

2.2 Le modèle de régression logistique

2.2.1 Introduction

Le modèle de régression logistique est généralement utilisé pour modéliser une réponse binaire dans le cadre de données médicales. Un exemple de réponse binaire est le statut d'infection (infecté *vs* non infecté) à l'égard de certaines maladies. Un modèle de régression logistique peut être utilisé pour étudier la relation entre le statut d'infection et les différentes covariables qui peuvent être par contre soit qualitatives, soit quantitatives (voir (Hosmer & Lemeshow 2000) et (Aminot & Damon 2002)). Si Y_i représente le statut d'infection pour le $i^{\text{ème}}$ individu dans l'échantillon de taille n ($Y_i = 1$ si l'individu est infecté, et $Y_i = 0$ sinon), et $\mathbf{X}_i$ représente le vecteur de prédicteur linéaire de dimension p correspondant, le modèle de régression logistique traduit la relation entre Y_i et $\mathbf{X}_i = (X_1, X_2, ..., X_p)$ en terme de probabilité conditionnelle $\mathbf{P}(Y_i = 1|\mathbf{X}_i)$ pour le statut d'infection, comme suit :

$$\log\left(\frac{\mathbb{P}(Y_i = 1|\mathbf{X}_i)}{1 - \mathbb{P}(Y_i = 1|\mathbf{X}_i)}\right) = \beta^\top \mathbf{X}_i, \tag{2.3}$$

où $\beta \in \mathbb{R}^p$ est un paramètre inconnu (à estimer). Une littérature détaillée a été consacrée jusqu'ici à l'inférence statistique des modèles de régression logistique. Les procédures d'estimation et de tests pour cette catégorie de modèles sont maintenant bien établies et sont disponibles dans les logiciels standards de statistique. En particulier, l'estimateur du maximum de vraisemblance de β est obtenu en résolvant l'équation du score. Les résultats asymptotiques pour cet estimateur sont donnés dans (Gouriéroux & Monfort 1981) et (Fahrmeir & Kaufmann 1985), entre autres. Le lecteur pourra se référer à (Hosmer & Lemeshow 2000) et (Hilbe 2009) pour des études détaillées et de nombreux exemples.

La fonction de lien *logit* est la plus généralement utilisée. L'intérêt de cette fonction réside dans la simplicité de passage à l'estimation d'un odds-ratio (OR) ou rapport des cotes qui mesure la force de l'association entre la variable endogène et une variable exogène. En particulier, en épidémiologie, les résultats peuvent être aisément interprétés. Les coefficients estimés par le modèle sont en effet liés mathématiquement à l'odds-ratio, bien qu'il ne soit qu'une approximation du risque relatif. La méthode de régression logistique est donc la méthode multivariable de choix pour rechercher des facteurs de risque ou des facteurs protecteurs de maladie.

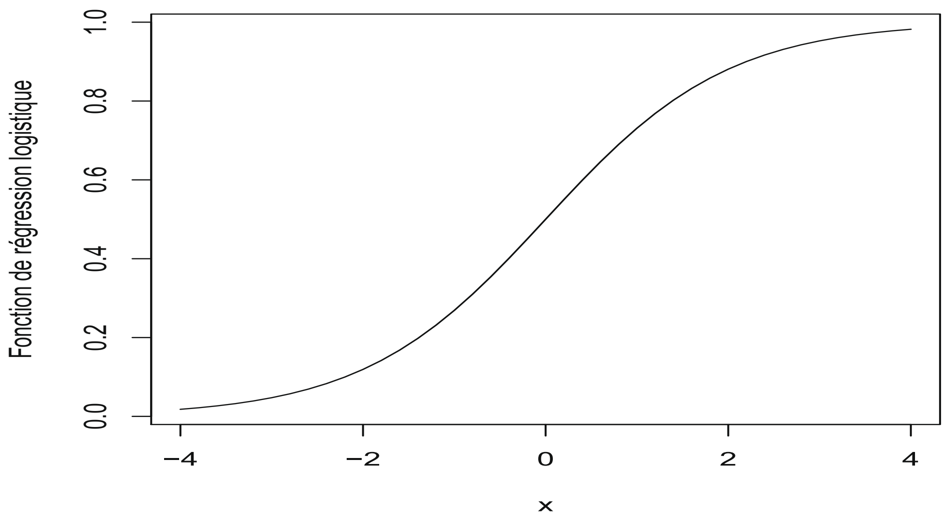

FIG. 2.1 – Fonction de régression logistique

2.2.2 Identifiabilité du modèle

On rappelle que le modèle est identifiable si pour $\beta \neq \tilde{\beta}$, les lois de $\{(Y_i|x_i), i = 1, \ldots, n\}$ associées à β et $\tilde{\beta}$ sont différentes. Tout comme pour le modèle linéaire (Guyon 2001), une condition nécessaire pour pouvoir estimer les paramètres est que l'échantillon rende le modèle identifiable : les lois de $\{(Y_i|\mathbf{X}_i = x_i), i = 1, \ldots, n\}$ associées à β et $\tilde{\beta}$ sont différentes. Comme $Y_i|\mathbf{X}_i = x_i$ suit une loi de Bernoulli de paramètre $p(x_i)$ et que la fonction `logit` est strictement croissante, cette condition équivaut à l'existence d'un x_i tel que $\beta^\top x_i \neq \tilde{\beta}^\top x_i$. Ce qui, comme pour le modèle linéaire, équivaut à $\texttt{rang}(\mathbf{X}) = p$. On supposera par la suite que cette condition est vérifiée.

2.2.3 Estimation du modèle

Nous allons utiliser l'échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ pour l'estimation des paramètres β par la méthode du maximum de vraisemblance. Rappelons que $Y_i|\mathbf{X}_i = x_i$ suit une loi de Bernoulli de paramètre

$$\begin{aligned} p(x_i) &= \mathbf{P}(Y_i|\mathbf{X}_i = x_i) \\ &= \frac{e^{\beta^\top \mathbf{x}_i}}{1 + e^{\beta^\top x_i}}. \end{aligned}$$

Pour trouver l'estimateur du maximum de vraisemblance de β, nous définissons la vraisemblance comme suit

$$\begin{aligned} L_n(\beta) &= \prod_{i=1}^{n} p(x_i)^{Y_i}(1 - p(x_i))^{1-Y_i} \\ &= \prod_{i=1}^{n} \frac{e^{Y_i \beta^\top x_i}}{1 + e^{\beta^\top x_i}}. \end{aligned}$$

Maintenant, nous définissons l'estimateur du maximum de vraisemblance, $\hat{\beta}$, de β en maximisant la fonction de log-vraisemblance des valeurs observées Y_i et x_i, $i = 1, \ldots, n$. Elle est donnée par

$$l_n(\beta) = \log L_n(\beta) = \sum_{i=1}^{n} [Y_i \beta^\top x_i - \log(1 + \exp(\beta^\top x_i)].$$

En dérivant une fois par rapport au paramètre β on obtient

$$\dot{l}_n(\beta) = \frac{\partial l_n(\beta)}{\partial \beta} = \sum_{i=1}^{n} [x_i(Y_i - p(x_i))]. \tag{2.4}$$

Une condition nécessaire d'existence de solution sur $\mathbb{R}^p$ est l'annulation de (2.4). Nous obtenons alors l'équation suivante (appelée équation du score)

$$S(\beta) = \frac{\partial l_n(\beta)}{\partial \beta} = 0. \tag{2.5}$$

En dérivant une seconde fois par rapport au paramètre β, on obtient

$$\ddot{l}_n(\beta) = \Big(\frac{\partial^2 l_n(\beta)}{\partial \beta_j \partial \beta_k}\Big)_{j,k} = -\sum_{i=1}^{n} p(x_i)(1-p(x_i))x_{ij}x_{ik}, \quad 1 \leq j,k \leq p, \tag{2.6}$$

la valeur x_{jk} représente l'observation k de l'individu j.

Nous pouvons ré-écrire de manière vectorielle et matricielle les équations (2.5) et (2.6) comme suit

$$\dot{l}_n(\beta) = [X^\top(Y-p)] \text{ et } \ddot{l}_n(\beta) = -X^\top W X,$$

où $p = (p(x_1), \ldots, p(x_n))^\top$ et

$$W = \begin{pmatrix} p(x_1)(1-p(x_1)) & 0 & \ldots \\ 0 & p(x_2)(1-p(x_2)) & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

Nous allons maintenant montrer que $\ddot{l}_n(\beta)$ est semi-définie négative pour tout $\beta \in \mathbb{R}^p$. Nous avons :

$$u^\top \ddot{l}_n(\beta) u = u^\top X^\top W X u = -\sum_{i=1}^{n} (x_i^\top u)^2 (p(x_i)(1-p(x_i))),$$

Nous remarquons que $p(x_i)(1-p(x_i))$ est toujours positif, nous avons donc

$$u^\top \ddot{l}_n(\beta) u \leq 0, \quad \forall\ u \in \mathbb{R}^p \text{ et } \forall\ \beta \in \mathbb{R}^p.$$

Comme $\ddot{l}_n(\beta)$ est semi-définie négative, la fonction de log-vraisemblance, l_n, est donc concave. La recherche d'une solution explicite pour l'équation (2.5) est une tâche délicate. Plusieurs techniques d'optimisation (méthodes itératives) sont disponible pour résoudre ce problème (voir par exemple (Mak 1993) et (Givens & Hoeting 2005)). Nous utilisons l'algorithme de Newton Raphson pour maximiser l_n.

2.2.4 Rappel sur l'algorithme de Newton-Raphson

La méthode de Newton-Raphson permet une résolution numérique des équations du score du type (2.5). On part tout d'abord d'une valeur initiale arbitraire de β, notée β^0. On note $\beta^1 = \beta^0 + h$ une valeur candidate pour être solution de $S(\beta) = 0$, c'est-à-dire $S(\beta^0 + h) = 0$. Par un développement limité à l'ordre un de la fonction S, on obtient l'approximation suivante

$$S(\beta^0 + h) = S(\beta^0) + hS'(\beta^0).$$

Comme $S(\beta^0 + h) = 0$, on obtient pour h la valeur suivante

$$h = -[S'(\beta^0)]^{-1} S(\beta^0).$$

Il vient

$$\beta^1 = \beta^0 - [\ddot{l}_n(\beta^0)]^{-1} \dot{l}_n(\beta^0).$$

On itère le processus. La procédure se résume de la manière suivante

1. Choix d'un point de départ β^0 ;
2. Calculer $\beta^{k+1} = \beta^k - [\ddot{l}_n(\beta^k)]^{-1} \dot{l}_n(\beta^k)$.

Algorithme 2.2.1 Choisir β^0

$k = 1$

Répéter

$\beta^{k+1} = \beta^k - [\ddot{l}_n(\beta^k)]^{-1} \dot{l}_n(\beta^k)$

$k = k + 1$

Jusqu'à

$\beta^{k+1} \approx \beta^k$ et/ou $L_n(\beta^{k+1}) \approx L_n(\beta^k)$.

2.2.5 Propriétés asymptotiques du modèle

Dans cette partie nous présentons les résultats d'existence, de consistance et de normalité asymptotique de l'estimateur du maximum de vraisemblance du paramètre β dans le modèle de régression logistique. Nous supposons que les hypothèses suivantes définies dans (Gouriéroux & Monfort 1981) et (Fahrmeir & Kaufmann 1985) sont vérifiées :

1. **H1** : Les variables explicatives sont uniformément bornées, i.e., $\exists\ C < \infty$: $\|X\| \leq C$.
2. **H2** : Soit λ_{1n} et λ_{pn} les valeurs propres respectivement minimale et maximale de la matrice $X^\top D(\beta_0)X$. Alors il existe une constante $K < \infty$ telle que $\frac{\lambda_{pn}}{\lambda_{1n}} \leq K$, pour tout n.

Théorème 2.2.2 *(Existence et consistance)*
*Sous les hypothèses **H1** et **H2**, l'EMV noté $\widehat{\beta}_n$ de β existe presque sûrement quand n tend vers $+\infty$, et $\widehat{\beta}_n$ converge presque sûrement quand n tend vers $+\infty$ vers la vraie valeur β_0 si et seulement si $\lim_{n\longrightarrow+\infty} \lambda_{1n} = +\infty$.*

Théorème 2.2.3 *(Normalité asymptotique)*
*Sous les hypothèses **H1** et **H2**, et si l'EMV est consistant, alors*

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) \longrightarrow \mathcal{N}(0, \mathcal{I}(\beta_0)^{-1}), \quad \textit{quand } n \rightarrow +\infty$$

où

$$\mathcal{I}(\beta) = -\mathbb{E}[\frac{\partial^2 l_n(\beta)}{\partial\beta\partial\beta^\top}]$$

est la matrice d'information de Fisher.

Le Théorème 2.2.3 nous permet facilement de déduire un estimateur de la variance de $\widehat{\beta}_n$. Ce qui nous permet de donner des intervalles de confiance de niveau $1-\alpha$ pour $\beta_j,\ j=1,\ldots,p$. Il est également possible de tester l'impact des variables explicatives par

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0, \quad \text{contre} \quad H_1 : \exists\, j \in \{1,\ldots,p\} : \beta_j \neq 0.$$

Pour cela trois tests sont généralement utilisés :
- Le test de Wald ;
- Le test du score ;
- Le test du rapport de vraisemblance ou de la déviance.

2.2.6 Interprétation des coefficients de régression et *odds ratio*

En général l'interprétation des coefficients β s'effectue en terme d'*odds ratio*. Les odds ratio sont des outils souvent appréciés dans le domaine de l'épidémiologie. Ils servent à mesurer l'effet d'une variable continue ou le contraste entre les effets d'une variable qualitative. L'idée générale est de raisonner en terme de probabilités ou de rapport de cotes (odds).

Définition 2.2.4 *L'odds (chance) pour un individu x d'obtenir la réponse $Y=1$ est définie par*

$$\texttt{odds}(x) = \frac{p(x)}{1-p(x)}, \quad \textit{où} \quad p(x) = \mathbf{P}(Y=1|X=x).$$

L'odds ratio (rapport des chances) entre deux individus x et x' est

$$OR(x,x') = \frac{\texttt{odds}(x)}{odds(x')}.$$

Les odds ratio peuvent être utilisés de différentes manières :

$OR(x,x') > 1$	$\Longleftrightarrow$	$p(x) > p(x')$
$OR(x,x') = 1$	$\Longleftrightarrow$	$p(x) = p(x')$
$OR(x,x') < 1$	$\Longleftrightarrow$	$p(x) < p(x')$

Tab. 2.1 – Règles d'interprétation des odds ratio

1. **Comparaison de probabilités de succès entre deux individus** (voir tableau 2.1)
2. **Interprétation en terme de risque relatif** : dans le cas où $p(x)$ et $p(x')$ sont très petits par rapport à 1, comme dans le cas d'une maladie très rare, alors on peut approximer l'odds ratio comme $OR(x,x') \approx p(x)/p(x')$ et interpréter simplement.
3. **Mesure de l'impact d'une variable** : pour le modèle de régression logistique

$$\texttt{logit}(p(x)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

il est facile de vérifier que :

$$OR(x,x') = \exp(\beta_1(x_1 - x_1')) \ldots \exp(\beta_p(x_p - x_p')).$$

Pour étudier l'influence d'une variable sur l'odds ratio, il suffit de considérer deux observations x et x' qui diffèrent uniquement par la j-ème composante. On obtient alors :

$$OR(x,x') = \exp(\beta_1(x_j - x_j')).$$

Ainsi une variation de la j-ème variable d'une unité (sur l'échelle de cette variable) correspond à un odds ratio $\exp(\beta_j)$ qui est uniquement fonction du coefficient β_j. Le coefficient permet de mesurer l'influence de la j-ème variable sur le rapport $p(x)/(1-p(x))$ lorsque x_j varie d'une unité, et ceux indépendamment de la valeur de x_j. Une telle analyse peut se révéler intéressante pour étudier l'influence d'un changement d'état d'une variable qualitative.

Remarque 2.2.5

La variable dépendante Y peut prendre plusieurs valeurs (ordinales ou non) : nous parlerons dans ce cas de régression polytomique. (Fahrmeir & Tutz 2001) ont fait une étude détaillée de ce modèle.

Chapitre 3

Modèles de régression avec inflation de zéros

Sommaire

La modélisation de données de comptage est une problématique très répandue dans divers domaines comme la banque, les assurances, l'économétrie, la médecine ou encore le marketing. Aussi, les méthodes de modélisation adaptées à ce type de données ont été largement explorées dans la littérature. La régression de Poisson est et reste le premier modèle auquel les utilisateurs font référence dans ce genre de situation. Cependant, des applications à des données réelles ont amené les gens à réfléchir

à des solutions alternatives au problème de sur-dispersion ou au problème d'excès de zéros induits par les mécanismes du phénomène étudié. Parmi celles-ci nous avons les modèles avec inflation de zéros introduits par (Lambert 1992) qui répondent de manière très claire au problème d'excès de zéros. On pourra également consulter à ce propos (Yip 1988), (Yip 1991), (Fong & Yip 1993), et (Fong & Yip 1995). La présence d'excès de zéros dans les données de comptage constitue un phénomène commun dans beaucoup d'applications parmi lesquelles nous pouvons citer la médecine (Bohning *et al.* 1999), la santé publique (Zhou & Tu 2000), les sciences environnementales (Agarwal *et al.* 2002), l'agriculture (Hall 2000) et le secteur des applications industrielles (Lambert 1992). Les travaux de recherche sur la généralisation de ces modèles ainsi que leurs mises en application sont nombreux.
(Consul & Famoye 1992) proposent une régression de Poisson généralisée avec l'introduction d'un nouveau paramètre dans le modèle standard pour modéliser la dispersion.
Récemment, (Famoye & Singh 2006) développent une régression de Poisson généralisée avec inflation de zéros pour modéliser les violences domestiques.
(Kelley & Anderson 2008) ont également utilisé les modèles avec inflation de zéros pour modéliser des données ordinales avec une présence excessive de zéros. Plus généralement, dans un modèle avec inflation de zéros, d'une part, un modèle logistique est utilisé pour déterminer si le statut de l'individu est dans le groupe des zéros ou non, et d'autre part un modèle Binomial, à Odds proportionnels ou de comptage (Poisson ou Binomial Négatif) est utilisé pour modéliser la survenue de l'événement d'intérêt dans le groupe des non zéros.

3.1 Modèles de régression ZIP et ZINB

3.1.1 Modèles de régression de Poisson et Binomial Négatif

Le modèle de régression de Poisson (régression log-linéaire) (Hilbe 2007) est le modèle de base qui prend explicitement en compte l'aspect entier positif des valeurs de la variable dépendante de comptage Y. Dans ce modèle, la probabilité d'un événement de comptage y_i, étant donné le vecteur de covariables $\mathbf{X}_i$, est donnée par la distribution de Poisson :

$$\mathbf{P}(Y_i = y_i|\mathbf{X}_i = x_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

avec

$$\mathbb{E}(y_i|\mathbf{X}_i = x_i) = \mu_i = \exp(\beta^\top x_i),$$

où $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ est le vecteur inconnu des paramètres.

Prendre l'exponentiel de $\beta^\top x_i$ assure que la moyenne conditionnelle μ_i est positive. Le nom de modèle de régression log-linéaire est également utilisé pour le modèle de régression de Poisson, car le logarithme de la moyenne conditionnelle est une fonction linéaire des paramètres :

$$\log[\mathbb{E}(y_i|\mathbf{X}_i = x_i)] = \log(\mu_i) = \beta^\top x_i.$$

Le modèle de régression de Poisson suppose que les données sont dispersées de manière égale, c'est-à-dire, que la variance conditionnelle est égale à la moyenne conditionnelle. Ce qui n'est pas toujours le cas. Les données réelles sont souvent caractérisées par une sur-dispersion c'est-à-dire que la variance dépasse la moyenne. Dans ce cas un modèle de mélange Gamma-Poisson est proposé. Ce qui conduit au modèle de régression Binomiale Négative (Hilbe 2007) qui permet de généraliser le modèle de régression de Poisson en prenant en compte cette sur-dispersion des données par l'introduction d'un terme d'hétérogénéité non observée chez l'observation

i. On a

$$\mathbb{E}(y_i|\mathbf{X}_i = x_i, \tau_i) = \mu_i\tau_i = \exp(\beta^\top x_i)\tau_i,$$

où τ_i suit une loi Gamma de moyenne 1 et de variance α. Conditionnellement à $\mathbf{X}_i$ et τ_i, la variable dépendante de comptage Y_i est toujours distribuée selon une loi de Poisson :

$$\mathbf{P}(Y_i = y_i|\mathbf{X}_i = x_i, \tau_i) = \frac{\exp(-\mu_i\tau_i)(\mu_i\tau_i)^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

Conditionnellement à $\mathbf{X}_i$, Y_i est distribuée selon une loi binomiale négative :

$$\mathbf{P}(Y_i = y_i|\mathbf{X}_i = x_i) = \frac{\Gamma(y_i + 1/\alpha)}{y_i!\Gamma(1/\alpha)}\Big(\frac{1}{1+\alpha\mu_i}\Big)^{1/\alpha}\Big(\frac{\mu_i}{1/\alpha + \mu_i}\Big)^{y_i}$$

où α est un paramètre auxiliaire mesurant le degré de sur-dispersion. Cette loi a une moyenne conditionnelle μ_i et une variance conditionnelle $\mu_i(1 + \alpha\mu_i)$. La loi Binomiale Négative tend vers la loi de Poisson lorsque α tend vers zéro.

3.1.2 Modèles de régression ZIP et ZINB

Les premièrs types de données sur lesquelles une inflation de zéros a été observée sont les données de comptage. D'où l'utilisation des modèles de régression Poisson et Binomial Négatif et la naissance des modèles de régression ZIP et ZINB ((Lambert 1992), (Greene 1994) et (Aldo *et al.* 2011)).

Soit $Y_i, i = 1, \ldots, n$ une variable dépendante de comptage positive. La probabilité pour qu'un individu i soit dans le groupe des zéros est notée π_i.

La variable Y_i est modélisée par un ZIP si :

$$\mathbf{P}(Y_i = y_i|\mathbf{X}_i, \mathbf{Z}_i) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\mu_i) & \text{si } y_i = 0 \\ \\ (1-\pi_i)\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} & \text{si } y_i > 0 \end{cases} \tag{3.1}$$

avec

$$\mathbb{E}(Y_i|\mathbf{X}_i,\mathbf{Z}_i) = (1-\pi_i)\lambda_i \quad \text{et} \quad \text{var}(Y_i|\mathbf{X}_i,\mathbf{Z}_i) = (1-\pi_i)\mu_i(1+\pi_i\mu_i).$$

La variable aléatoire Y_i est modélisée par un ZINB si

$$\mathbf{P}(Y_i = y_i|\mathbf{X}_i,\mathbf{Z}_i) = \begin{cases} \pi_i + (1-\pi_i)(\frac{1}{1+\alpha\mu_i})^{\alpha} & \text{si } y_i = 0 \\ \\ (1-\pi_i)\frac{\Gamma(y_i+1/\alpha)}{\Gamma(1/\alpha)y_i!}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} & \text{si } y_i > 0 \end{cases} \tag{3.2}$$

avec

$$\mathbb{E}(Y_i|\mathbf{X}_i,\mathbf{Z}_i) = (1-\pi_i)\mu_i \quad \text{et} \quad \text{var}(Y_i|\mathbf{X}_i,\mathbf{Z}_i) = (1-\pi_i)\mu_i(1+(\alpha+\pi_i)\mu_i),$$

où α est un paramètre de sur-dispersion. Comme pour les modèles de Poisson et Binomial Négatif, le modèle ZINB tend vers le modèle ZIP lorsque α tend vers zéro. Pour ces deux modèles (3.1)-(3.2), on suppose que la probabilité π_i et la moyenne conditionnelle μ_i sont respectivement modélisées par $logit(\pi_i) = \gamma^\top \mathbf{Z}_i$ et par $\log(\mu_i) = \beta^\top \mathbf{X}_i$. $\mathbf{X}_i \in \mathbb{R}^p$ et $\mathbf{Z}_i \in \mathbb{R}^q$ représentent les covariables. $\beta \in \mathbb{R}^p$ et $\theta \in \mathbb{R}^q$ représentent les vecteurs des paramètres inconnus. Les covariables $\mathbf{X}_i$ et $\mathbf{Z}_i$ peuvent ou non avoir des composantes communes (Pradhan & Leung 2006).

3.1.3 Propriétés asymptotiques du modèle ZIP

Nous nous intéressons dans cette partie aux propriétés asymptotiques de l'estimateur du maximum de vraisemblance de β et γ du modèle (3.1). Nous considérons également que tous les individus ont la même probabilité π d'appartenir au groupe des zéros. Nous nous plaçons dans les mêmes conditions que (Czado *et al.* 2007) avec les hypothèses suivantes :

- H1 : $\frac{n}{\lambda_{min}(\mathbf{F}_n)} \leq C_1 \quad \forall\, n \geq 1$, où C_1 est une constante positive, $\mathbf{F}_n$ la matrice d'information de Fisher et λ_{min} sa plus petite valeur propre.
- H2 : Les variables explicatives sont uniformément bornées, i.e., $\exists\ C_2 < \infty$: $\|X\| \leq C_2$.

– H3 : soit $\mathbb{B}$ un ensemble ouvert de $\mathbb{R}^p$ et $\theta_0 = (\beta_0^\top, \pi_0)^\top$ la vraie valeur de $\theta = (\beta^\top, \pi)^\top$ un point intérieur de $\mathbb{B} \times [0, 1]$.

Sous ces conditions, (Czado *et al.* 2007) ont montré le résultat suivant :

Théorème 3.1.1 *Il existe une suite de variables aléatoires $\widehat{\theta}_n$ telles que*

(i) $\mathbf{P}(s_n(\widehat{\theta}_n) = 0) \to 1$ *quand* $n \to \infty$ *(existence asymptotique),*

(ii) $\widehat{\theta}_n \xrightarrow{\mathbf{P}} \theta_0$ *quand* $n \to \infty$ *(consistance),*

(iii) $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathscr{L}} \mathscr{N}(0, \mathbf{F}^{-1}(\theta_0))$ *quand* $n \to \infty$ *(normalité asymptotique).*

où

$$s_n(\theta) = \frac{\partial \log L_n(\theta)}{\partial \theta} \quad \text{et} \quad \mathbf{F} = -\mathbb{E}\left[\frac{\partial^2 \log L_n(\theta)}{\partial \theta \partial \theta^\top}\right].$$

où L_n représente ici la vraisemblance du modèle ZIP considéré.

(Hall & Shen 2010) se sont appuyés sur le fait que l'EMV est sensible aux valeurs aberrantes pour utiliser une nouvelle procédure d'estimation du modèle 3.1 dite "robust expectation-solution (RES) estimation" reliée à la méthode de M-estimation qu'ils ont précisément appelée *expectation-solution* ou *algorithme ES*. Cet algorithme est une modification de l'algorithme *expectation-maximization* (EM) avec la propriété de robustesse. Dans les modèles ZIP, comme dans bien d'autres modèles de mélange, l'algorithme EM constitue une approche pratique pour le calcul de l'estimateur du maximum de vraisemblance (voir (Lambert 1992)). Cet algorithme tient compte de la présence de données manquantes dans le problème. En particulier supposons observer la variable v telle que $v_i = 1$ si y_i provient de l'ensemble des zéros (distribution dégénérée) et $v_i = 0$ si y_i provient de l'ensembles des non zéros (distribution non dégénérée). Ainsi la log-vraisemblance pour les données complètes (y,z) est donnée par

$$\begin{aligned} l^c(y,v,\beta,\gamma) &= \sum_{i=1}^{n}\left\{v_i\gamma^\top Z_i - \log(1+e^{\gamma^\top Z_i})\right\} \\ &+ \sum_{i=1}^{n}(v_i)\left\{y_i\beta^\top X_i - e^{\beta^\top X_i} - \log(y!)\right\} \\ &= l^c_\gamma(\gamma,y,v) + l^c_\beta(\beta,y,v), \end{aligned} \tag{3.3}$$

où $v = (v_1, ..., v_n)^\top$.

Avec l'algorithme EM (Dempster *et al.* 1977), la log-vraisemblance est maximisée de manière itérative en commençant par une valeur initiale $(\beta^{(0)\top}, \gamma^{(0)\top})^\top$ et en alternant les étapes 1 et 2 suivantes :

1. **Étape E** : estimer la variable v_i par son espérance conditionnelle $v_i^{(r)}$ sous les estimations courantes des paramètre $\beta^{(r)}$ et $\gamma^{(r)}$.
2. **Étape M** : trouver $\beta^{(r+1)}$ et $\gamma^{(r+1)}$ en maximisant respectivement les fonctions $l^c_\gamma(\gamma, y, v^{(r)})$ et $l^c_\beta(\beta, y, v^{(r)})$. (Hall & Shen 2010) ont montré que maximiser ces deux fonctions revient à résoudre respectivement les deux équations suivantes

$$\frac{1}{n}\sum_{i=1}^{n}\left\{v_i^{(r)} - \pi_i\right\} Z_i = 0. \tag{3.4}$$

$$\frac{1}{n}\sum_{i=1}^{n}(1-v_i^{(r)})\left\{y_i - e^{\beta^\top X_i}\right\} X_i = 0. \tag{3.5}$$

Dans l'approche RES, (Hall & Shen 2010) proposent de remplacer les équations (3.4) et (3.5) par des estimations de fonctions robustes. Essentiellement, ils proposent de pondérer les observations qui se situent dans la queue extrême supérieure et inférieure de la distribution de Poisson dans la fonction d'estimation.

Sous des conditions de régularité de (Rosen *et al.* 2000) liées à l'algorithme ES et de (Carroll *et al.* 1995), (Hall & Shen 2010) ont montré le résultat suivant plus général (dans le sens où $\theta = (\beta^\top, \gamma^\top)^\top \in \mathbb{R}^{p+q}$) que le théorème précédent :

Théorème 3.1.2 *Si l'algorithme RES converge, alors il existe une suite de variables aléatoires $\widehat{\theta}_n$ telles que*

(i) $\widehat{\theta}_n \xrightarrow{\mathbf{P}} \theta_0$ *quand* $n \to \infty$ *(consistance),*

(ii) $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathscr{L}} \mathscr{N}(0, \boldsymbol{V}(\theta_0))$ *quand* $n \to \infty$ *(normalité asymptotique).*

où l'expression de la variance asymptotique est donnée dans (Hall & Shen 2010).
Des extensions du modèle ZIP ont été étudiées dans le cadre semi-paramétrique et doublement semi-paramétrique et des résultats de convergence ont été également démontrés respectivement pour ces deux types de modèles par (Lam *et al.* 2006) et (He *et al.* 2010).

3.2 Modèle de régression ZIB

3.2.1 Modèle binomial

On considère, pour $i = 1 \ldots, I$, différentes valeurs fixées $x_{i1}, \ldots, x_{ip}$ des variables explicatives $X_1, \ldots, X_p$. Ces dernières pouvant être des variables quantitatives ou encore des variables qualitatives.
Pour chaque groupe, c'est-à-dire pour chacune des combinaisons de valeurs ou facteurs, on réalise n_i observations ($n = \sum_{i=1}^{I} n_i$) de la variable réponse binaire Y qui se mettent sous la forme $y_i/n_i, \ldots, y_I/n_I$ où y_i désigne le nombre de "*succès*" observés lors des n_i essais. On suppose que toutes les observations sont indépendantes et qu'à l'intérieur d'un même groupe, tous les individus ont la même probabilité de succès. Alors, la variable Y_i suit une loi binomiale $\mathscr{B}(n_i, \pi_i)$ dont la fonction de densité s'écrit

$$\mathbf{P}(Y = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

où la probabilité π_i est modélisée par une fonction de lien `logit` :

$$logit(\pi_i) = \beta^\top x_i, \quad i = 1, \ldots, I,$$

ce qui s'écrit encore

$$\pi_i = \frac{e^{\beta^\top x_i}}{1 + e^{\beta^\top x_i}}.$$

Le vecteur des paramètres est estimé par maximisation de la log-vraisemblance. Il n'y a pas de solution analytique, celle-ci est obtenue par des méthodes numériques itératives (par exemple Newton Raphson) dont certaines reviennent à itérer des estimations de modèles de régression par moindres carrés généralisés avec des poids et des métriques adaptés à chaque itération.
L'optimisation fournit une estimation $\widehat{\beta}_n$ de β, il est alors facile d'en déduire les estimations ou prévisions $\widehat{\pi}_i$ de π_i et celles des effectifs $\widehat{y}_i$ de y_i :

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_n^\top x_i}}{1 + e^{\widehat{\beta}_n^\top x_i}} \quad \text{et} \quad \widehat{y}_i = n_i \widehat{\pi}_i.$$

3.2.2 Définition du modèle ZIB

Le modèle de régression Binomial zéro-inflaté (ZIB) a été introduit en premier par (Kemp & Kemp 1988), mais ils l'ont seulement utilisé pour mettre en valeur quelques aspects importants de l'estimation de la fonction génératrice de probabilité empirique. (Hall 2000) a étudié et étendu le modèle ZIB à un modèle avec et sans effet aléatoire et a donné quelques applications détaillées dans le cadre de données réelles. En considérant les mêmes notations que (Hall 2000), nous définissons le modèle ZIB comme suit :

$$Y_i \sim \begin{cases} 0 & \text{avec une probabilité } p_i \\ \text{Binomiale}(n_i, \pi_i) & \text{avec une probabilité } 1 - p_i. \end{cases} \tag{3.6}$$

Ce qui implique que

$$Y_i = \begin{cases} 0 & \text{avec une probabilité } p_i + (1-p_i)(1-\pi_i)^{n_i} \\ k & \text{avec une probabilité } (1-p_i)\binom{n_i}{k}\pi_i^{n_i}(1-\pi_i)^{n_i-k}, k = 1, 2, ..., n_i, \end{cases} \tag{3.7}$$

avec

$$\mathbb{E}(Y_i) = (1-p_i)n_i\pi_i, \quad \text{et} \quad \mathrm{var}(Y_i) = (1-p_i)n_i\pi_i\Big(1-\pi_i(1-p_in_i)\Big).$$

Les deux probabilités peuvent également être exprimées conjointement comme une distribution de Bernoulli généralisée donnant la vraisemblance complète suivante :

$$L_n(\beta,\gamma) = \prod_{i=1}^{n}(p_i + (1-p_i)(1-\pi_i)^{n_i})^{u_i}((1-p_i)\binom{n_i}{k}\pi_i^{n_i}(1-\pi_i)^{n_i-k})^{1-u_i}. \quad (3.8)$$

Les paramètres $p = (p_1,\dots,p_n)$ et $\pi = (\pi_1,\dots,\pi_n)$ sont respectivement modélisés via une fonction de lien `logit`, $logit(p) = \gamma^\top \mathbf{Z}$ et $logit(\pi) = \beta^\top \mathbf{X}$ où $\mathbf{Z} \in \mathbb{R}^q$ et $\mathbf{X} \in \mathbb{R}^p$ sont les vecteurs de covariables, n étant le nombre d'individu, p et q sont respectivement le nombre de covariables dans le modèle de régression binomial et le nombre de covariables dans la partie inflation de zéros, $\gamma \in \mathbb{R}^q$ et $\beta \in \mathbb{R}^p$ sont les paramètres de régression. La log-vraisemblance de ce modèle ZIB est alors donnée par

$$\begin{aligned} l_n(\beta,\gamma) &= \sum_{i=1}^{n}\{u_i\log(e^\gamma + (1+e^{\beta^\top \mathbf{X_i}})^{-n_i}) - \log(1+e^\gamma) + (1-u_i) \\ &\times \quad (y_i\beta^\top \mathbf{X_i} - \mathbf{n_i}\log(\mathbf{1}+\mathbf{e}^{\beta^\top \mathbf{X_i}}) + \log\binom{\mathbf{n_i}}{\mathbf{k}})\}. \end{aligned} \quad (3.9)$$

3.2.3 Identifiabilité de modèle de régression ZIB

(Teicher 1960), (Teicher 1963), (Blischke 1978) et (Margolin *et al.* 1989) ont donné des conditions nécessaires et suffisantes pour l'identifiabilité d'un mélange d'un nombre fini k de modèles binomiaux.

$$f(y_j,\psi) = \sum_{i=1}^{k}\pi_i\binom{N}{y_j}\theta_i^{y_j}(1-\theta_i)\mathbb{1}_{A_N}(y_j), \quad (3.10)$$

où $A_N = \{0,1,...,N\}$, θ_i et π_i sont modélisés par des modèles de régression logistiques. Leurs résultats peuvent être résumés comme suit. Le modèle de mélange (3.10) est identifiable si et seulement si

$$k \leq \frac{1}{2}(N+1). \quad (3.11)$$

(Wang 1994) a étudié l'identifiabilité d'une collection de modèles de régression logistique. (Follmann & Lambert 1991a) ont établi des conditions suffisantes pour montrer l'identifiabilité de (3.10) dans le cas où les proportions de mélange ne sont pas fonction des covariables et que les paramètres de régression des composantes logistique différent seulement de par leurs intercepts. Ils ont montré que pour une réponse binaire, le nombre de composantes k dans le mélange doit être borné par une fonction du nombre de vecteurs de covariables qui convient ; et pour une réponse binomiale, k doit satisfaire la même condition de bornitude.

Des exemples d'application des modèles de mélange de modèles de régression logistiques standard à des jeux de données réelles biologiques peuvent être trouvés dans (Follmann & Lambert 1991a) et (Wang 1994). (Farewell & Sprott 1988) ont également donné un exemple d'application de modèle de mélange binomial.

3.3 Modèle de régression ZIPO

Lorsque la variable dépendante discrète a $K(K > 2)$ catégories, le modèle de régression est dit polytomique. Si les catégories sont ordonnées (par exemple la tension artérielle : faible, moyenne et élevée) on parle de régression ordinale. (Agresti 2002) a fait une étude détaillée de ce modèle. (Kelley & Anderson 2008) ont développé le modèle ZIPO qui fournit une méthode permettant de modéliser des données en présence d'une inflation de zéros. La spécification du modèle est similaire à celles étudiées en haut (ZIP, ZINB et ZIB). Soit $Y_i, 1 \leq i \leq n$ une variable ordinale avec J niveaux. Les probabilités cumulatives sont données par $\gamma_j = \mathbf{P}(Y_i \leq j), j = 0, 1, \ldots, J$. On a

$$Y_i \sim \begin{cases} 0 & \text{avec } p_i \\ \text{Multinomiale}(1, \gamma_{0,i}, \ldots, \gamma_{J,i}) & \text{avec } 1 - p_i. \end{cases} \tag{3.12}$$

Ce qui donne

$$Y_i = \begin{cases} 0 & \text{avec } p_i + (1 - p_i)\gamma_{0,i} \\ j & \text{avec } (1 - p_i)(\gamma_{j,i} - \gamma_{j-1,i}), \end{cases} \tag{3.13}$$

où $p = (p_1, \ldots, p_n)$ et $\gamma = (\gamma_0, \ldots, \gamma_J)$ sont modélisés respectivement par

$$logit(p) = \theta^\top \mathbf{Z} \quad \text{et} \quad logit(\pi) = \beta^\top \mathbf{X}.$$

Deuxième partie

Modèle de régression logistique avec fraction immune

Chapitre 4

Estimation par maximum de vraisemblance dans le modèle de régression logistique avec fraction immune

Sommaire

Ce chapitre a fait l'objet d'une publication :

Diop A., Diop A., and Dupuy J.-F.,

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

Electronic Journal of Statistics, Vol. 5, 460-483, 2011.

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, based on the joint modeling of the binary response of interest and the cure status. We investigate the identifiability of the resulting model. Then, we establish the consistency and asymptotic normality of the proposed estimator, and we conduct a simulation study to investigate its finite-sample behavior.

keywords **: Zero-inflation, Maximum likelihood estimation, Consistency,**

Asymptotic normality, Simulations

4.1 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 $\mathbf{X}_i$ denotes the corresponding (p-dimensional, say) predictor, the logistic regression model expresses the relationship between Y_i and $\mathbf{X}_i$ in term of the conditional probability $\mathbb{P}(Y_i = 1|\mathbf{X}_i)$ of infection, as :

$$\log\left(\frac{\mathbb{P}(Y_i = 1|\mathbf{X}_i)}{1 - \mathbb{P}(Y_i = 1|\mathbf{X}_i)}\right) = \beta^\top \mathbf{X}_i,$$

where $\beta \in \mathbb{R}^p$ is an unknown parameter to be estimated. An extensive literature has been devoted so far to 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 β is obtained by solving the following score equation :

$$\sum_{i=1}^{n} \mathbf{X}_i \left(Y_i - \frac{e^{\beta^\top \mathbf{X}_i}}{1 + e^{\beta^\top \mathbf{X}_i}}\right) = 0.$$

Asymptotic results (consistency and asymptotic normality) for this estimator were given by (Gouriéroux & Monfort 1981) and (Fahrmeir & Kaufmann 1985), among others. We refer the reader to (Hosmer & 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 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 usually does not know whether this subject is immune to the infection, or susceptible albeit still uninfected.

Estimating a regression model with a cure fraction can be viewed as a zero-inflated regression problem. Zero-inflation occurs in the analysis of count data when the observations contain more zeros than expected. Failure to account for these extra zeros is known to result in biased parameter estimates and inferences. The regression analysis of count data with excess zeros has attracted much attention so far. For example, (Lambert 1992) proposed the zero-inflated Poisson (ZIP) regression model for count data with many zeros. This was further extended to a semiparametric ZIP regression model by (Lam *et al.* 2006). We refer to (Dietz & Bohning 2000) and (Xiang *et al.* 2007) for a review of various other extensions of the ZIP model. Other popular models are the zero-inflated binomial (ZIB) regression model (see, for example, (Hall 2000)), and the zero-inflated negative binomial (ZINB) regression model (see, for example, (Ridout *et al.* 2001)). Recently, (Kelley & Anderson 2008) proposed a zero-inflated proportional odds model (ZIPO) for ordinal outcomes, when some individuals are not susceptible to the phenomenon being measured. Various other models and numerous references can be found in (Famoye & Singh 2006) and (Lee *et al.* 2006).

In our paper, we consider the problem of estimating a logistic regression model from binary response data with a cure fraction, when the cure probability is modeled by a logistic regression. This can be viewed as a zero-inflated Bernoulli regression problem, where logistic link functions are used for both the binary response of interest (the probability of infection, say) and the zero-inflation probability (the probability of being cured). The literature on zero-inflated models is extensive but to the best of our knowledge, the theoretical and numerical issues related to the statistical inference in this model have not been yet investigated. In this paper, we intend to fill this gap. We first investigate the identifiability question in this model. Then, we turn to the problem of estimation. The estimator we propose is obtained by maximizing the joint likelihood for the binary response of interest and the cure indicator. 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 4.2, we describe the problem of logistic regression with a cure fraction, and we propose an estimation method adapted to this setting. The proposed procedure is based on a joint regression model for the binary response of interest and the cure indicator. In Section 4.3, we investigate the identifiability of this model, and we state some regularity conditions. In Section 4.4, we derive the asymptotic properties of the resulting estimator. Section 4.5 describes a simulation study, where we numerically investigate the small to large sample properties of this estimator. A real data example illustrates the methodology. A discussion and some perspectives are given in Section 4.6.

4.2 Logistic regression with a cure fraction

4.2.1 Notations and the model set-up

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 $(\Omega, \mathscr{A}, \mathbb{P})$. 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), and S_i is a binary variable indicating whether individual i is susceptible to the infection ($S_i = 1$) or immune ($S_i = 0$). If $Y_i = 0$, then the value of S_i is unknown. Let $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ and $\mathbf{Z}_i = (1, Z_{i2}, \ldots, Z_{iq})'$ be random vectors of predictors or covariates (both categorical and continuous predictors are allowed). We shall assume in the following that the $\mathbf{X}_i$'s are related to the infection status, while the $\mathbf{Z}_i$'s are related to immunity. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are allowed to share some components.

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

$$\log\left(\frac{\mathbb{P}(Y = 1|\mathbf{X}_i, S_i)}{1 - \mathbb{P}(Y = 1|\mathbf{X}_i, S_i)}\right) = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} := \beta^\top \mathbf{X}_i \tag{4.1}$$

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

$$\mathbb{P}(Y = 1|\mathbf{X}_i, S_i) = 0 \tag{4.2}$$

if $\{S_i = 0\}$, where $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ 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 (4.1) includes estimation and testing for β. Without immunity (that is, if $S_i = 1$ for every $i = 1, \ldots, n$), inference on β from the sample $(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \ldots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)$ can be based on the maximum likelihood principle. When immunity is present, deriving the maximum likelihood estimator of β is no longer straightforward : if $Y_i = 0$, we do not know whether $\{S_i = 1\}$, so that (4.1) applies, or whether $\{S_i = 0\}$, so that (4.2) 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 4.5). Therefore in this paper, we aim at providing an alternative estimation procedure for β. This can be achieved if a model for immunity is available, as is explained in the next section.

4.2.2 The proposed estimation procedure

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, (Fang *et al.* 2005) and (Lu 2008; Lu 2010) who considered estimation in various survival regression models with a cure fraction) :

$$\log\left(\frac{\mathbb{P}(S = 1|\mathbf{Z}_i)}{1 - \mathbb{P}(S = 1|\mathbf{Z}_i)}\right) = \theta_1 + \theta_2 Z_{i2} + \cdots + \theta_q Z_{iq} := \theta^\top \mathbf{Z}_i \tag{4.3}$$

where $\theta = (\theta_1, \ldots, \theta_q)^\top \in \mathbb{R}^q$ is an unknown regression parameter.

Remark 4.2.1 We note that the model defined by (4.1)-(4.2)-(4.3) can be viewed as a zero-inflated Bernoulli regression model, with logit links for both the binary response of interest and the zero-inflation component. As far as we know, no theoretical investigation of this model has been undertaken yet. Such a work is carried out in the following.

From (4.1), (4.2), and (4.3), a straightforward calculation yields that

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

Let $\psi := (\beta^\top, \theta^\top)^\top$ denote the unknown k-dimensional ($k = p + q$) parameter in the conditional distribution of Y given $\mathbf{X}_i$ and $\mathbf{Z}_i$. ψ includes both β (considered as the parameter of interest) and θ (considered as a nuisance parameter). Now, the likelihood for ψ 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 :

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

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

$$\dot{l}_n(\psi) = \frac{\partial l_n(\psi)}{\partial \psi} = 0, \tag{4.4}$$

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 β, considered as a sub-component of $\widehat{\psi}_n$. We will however obtain consistency and asymptotic normality results for the whole $\widehat{\psi}_n$. Before proceeding, we need to set some further notations.

4.2.3 Some further notations

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

$$\mathbb{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{pmatrix} \quad \text{and} \quad \mathbb{Z} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ Z_{12} & Z_{22} & \cdots & Z_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1q} & Z_{2q} & \cdots & Z_{nq} \end{pmatrix},$$

and let $\mathbb{W}$ be the $(k \times 2n)$ block-matrix defined as

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

where 0_{ab} 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

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

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

$$A_i^{\beta}(\psi) = \frac{1 + e^{\theta^{\top}\mathbf{Z}_i}}{1 + e^{\beta^{\top}\mathbf{X}_i} + e^{\theta^{\top}\mathbf{Z}_i}} Y_i, \quad B_i^{\beta}(\psi) = \frac{e^{\beta^{\top}\mathbf{X}_i + \theta^{\top}\mathbf{Z}_i}}{(1 + e^{\beta^{\top}\mathbf{X}_i})(1 + e^{\beta^{\top}\mathbf{X}_i} + e^{\theta^{\top}\mathbf{Z}_i})},$$
$$A_i^{\theta}(\psi) = \frac{1 + e^{\beta^{\top}\mathbf{X}_i}}{1 + e^{\beta^{\top}\mathbf{X}_i} + e^{\theta^{\top}\mathbf{Z}_i}} Y_i, \quad B_i^{\theta}(\psi) = \frac{e^{\beta^{\top}\mathbf{X}_i + \theta^{\top}\mathbf{Z}_i}}{(1 + e^{\theta^{\top}\mathbf{Z}_i})(1 + e^{\beta^{\top}\mathbf{X}_i} + e^{\theta^{\top}\mathbf{Z}_i})}.$$

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

$$\dot{l}_n(\psi) = \mathbb{W}C(\psi) = 0.$$

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

$$\dot{l}_n(\psi) = \sum_{j=1}^{2n} \mathbb{W}_{\bullet j} C_j(\psi).$$

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

$$\mathbb{D}(\psi) = \begin{bmatrix} \mathbb{D}_1(\psi) & \mathbb{D}_3(\psi) \\ \mathbb{D}_3(\psi) & \mathbb{D}_2(\psi) \end{bmatrix},$$

where $\mathbb{D}_1(\psi), \mathbb{D}_2(\psi)$, and $\mathbb{D}_3(\psi)$ are $(n \times n)$ diagonal matrices, with i-th diagonal elements ($i = 1, \ldots, n$) respectively given by

$$\mathbb{D}_{1,ii}(\psi) = \frac{e^{\beta^\top \mathbf{X}_i + \theta^\top \mathbf{Z}_i}}{(1 + e^{\beta^\top \mathbf{X}_i})^2 (1 + e^{\beta^\top \mathbf{X}_i} + e^{\theta^\top \mathbf{Z}_i})},$$

$$\mathbb{D}_{2,ii}(\psi) = \frac{e^{\beta^\top \mathbf{X}_i + \theta^\top \mathbf{Z}_i}}{(1 + e^{\theta^\top \mathbf{Z}_i})^2 (1 + e^{\beta^\top \mathbf{X}_i} + e^{\theta^\top \mathbf{Z}_i})},$$

$$\mathbb{D}_{3,ii}(\psi) = \frac{e^{\beta^\top \mathbf{X}_i + \theta^\top \mathbf{Z}_i}}{(1 + e^{\beta^\top \mathbf{X}_i})(1 + e^{\theta^\top \mathbf{Z}_i})(1 + e^{\beta^\top \mathbf{X}_i} + e^{\theta^\top \mathbf{Z}_i})}.$$

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

$$\ddot{l}_n(\psi) = -\mathbb{W}\mathbb{D}(\psi)\mathbb{W}'.$$

Note that the size of $C(\psi), \mathbb{W}$, and $\mathbb{D}(\psi)$ depends on n. However, in order to simplify notations, n will not be used as a lower index for these vector and matrices. In the next section, we investigate the question of parameter identifiability in model (4.1)-(4.2)-(4.3).

4.3 Identifiability and regularity conditions

We first state some regularity conditions that will be needed to ensure identifiability and the asymptotic results in Section 4.4 :

C1 The covariates are bounded that is, there exist compact sets $F \subset \mathbb{R}^p$ and $G \subset \mathbb{R}^q$ such that $\mathbf{X}_i \in F$ and $\mathbf{Z}_i \in G$ for every $i = 1, 2, \ldots$ For every $i = 1, 2, \ldots$, $j = 2, \ldots, p$, $k = 2, \ldots, q$, $\text{var}[X_{ij}] > 0$ and $\text{var}[Z_{ik}] > 0$. For every $i = 1, 2, \ldots$, the X_{ij} ($j = 1, \ldots, p$) are linearly independent, and the Z_{ik} ($k = 1, \ldots, q$) are linearly independent.

C2 Let $\psi_0 = (\beta_0^\top, \theta_0^\top)^\top$ denote the true parameter value. β_0 and θ_0 lie in the interior of known compact sets $\mathscr{B} \subset \mathbb{R}^p$ and $\mathscr{G} \subset \mathbb{R}^q$ respectively.

C3 The Hessian matrix $\ddot{l}_n(\psi)$ is negative definite and of full rank, for every $n = 1, 2, \ldots$ Let λ_n and Λ_n be respectively the smallest and largest eigenvalues of $\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top$. There exists a finite positive constant c_2 such that $\Lambda_n/\lambda_n < c_2$ for every $n = 1, 2, \ldots$

C4 There exists a continuous covariate V which is in $\mathbf{X}$ but not in $\mathbf{Z}$ that is, if β_V and θ_V denote the coefficients of V in the linear predictors (4.1) and (4.3) respectively, then $\beta_V \neq 0$ and $\theta_V = 0$. At a model-building stage, it is known that V is in $\mathbf{X}$.

The conditions C1, C2, C3 are classical conditions for identifiability and asymptotic results in standard logistic regression (see, for example, (Gouriéroux & Monfort 1981) and (Guyon 2001)). The condition C4, which imposes some restrictions on the covariates, is required for identifiability of ψ in the joint model (4.1)-(4.2)-(4.3) (we may alternatively assume that the continuous covariate V is in $\mathbf{Z}$ but not in $\mathbf{X}$). In the following, we will assume that V is in $\mathbf{X}$ but not in $\mathbf{Z}$, with $\beta_V := \beta_l$ for some $l \in \{2, \ldots, p\}$, and for the i-th individual, we will denote V_i by X_{il}. The condition C4 is discussed in greater details in the following two remarks.

Remark 4.3.1 We may relate the identifiability issue in model (4.1)-(4.2)-(4.3) to the problem of identifiability of mixtures of logistic regression models, which was investigated by (Follmann & Lambert 1991b). (Follmann & Lambert 1991b) considered the case where there is a finite number c of components in the mixture (we consider here the case where $c = 2$, with one degenerate component) and the mixing probabilities are constant (here, the mixing probabilities given by (4.3) are allowed to depend on covariates). The authors have shown that finite mixtures of logistic regressions are identifiable provided that the number of unique covariate combinations values is sufficiently large. C4 can be viewed as a sufficient condition for achieving the same kind of requirement. A similar condition appears in (Kelley & Anderson 2008).

To understand C4, note that if $\mathbf{X}_i = \mathbf{Z}_i$, then exchanging the parameters β and θ in (4.1) and (4.3) yields the same likelihood value $L_n(\psi)$, which is a cause of model non-identifiability. A similar remark holds if we invert the linear predictors $\beta^\top \mathbf{X}_i$ and $\theta^\top \mathbf{Z}_i$. The condition C4 evacuates these problems.

First, by asking one of the covariates to be significant in one and only one linear predictor, C4 prevents $\beta^\top \mathbf{X}$ and $\theta^\top \mathbf{Z}$ from being of the same form, and the parameters are thus not exchangeable. Secondly, by assuming that we know, prior to model fitting, that there exists a covariate V which is in $\mathbf{X}$ but not in $\mathbf{Z}$, C4 will force each linear predictor to be attached to the correct corresponding model (4.1) or (4.3).

These facts are illustrated in a supplementary document available on annex B. There, we provide the results of a simulation study which investigates numerically the identifiability of model (4.1)-(4.2)-(4.3). For each of the models considered in this study, we assume that C4 is satisfied : the linear predictors $\beta^\top \mathbf{X}_i$ and $\theta^\top \mathbf{Z}_i$ share three covariates (one is continuous, two are discrete), and an additional continuous covariate is included in $\mathbf{X}_i$. Using the procedure described in Section 4.2, maximum likelihood estimates are obtained for β and θ, and are averaged over $N = 1000$ samples (we considered several combinations of sample size, proportion of immunes, proportion of infected among the susceptibles). Both parameters β and θ appear to be identifiable (the averaged estimates appear to be close to the true parameters, including those corresponding to the three shared covariates).

Remark 4.3.2 The condition C4 does not appear to be too restrictive in practice. Consider the example of the transmission of some disease by breastfeeding. If every child in the sample is breastfeeded, it can be expected that the length (in days, say) of the breastfeeding period (a continuous covariate) will influence the probability of infection, while the susceptibility probability will rather depend on risk factors such as say, the mother's infection status. It is also worth noting that the consequences of C4, in terms of model-building, are rather mild. At a model-building stage, we may be tempted to incorporate all available covariates in both linear predictors (4.1) and (4.3), and to remove irrelevant factors by using backward elimination. The condition C4 slightly restricts this fitting strategy, by imposing that one relevant continuous covariate is incorporated in one (and only one) linear predictor. This should often be doable in practice, since the statistician often gets some prior knowledge (from the clinicians, epidemiologists, ...) about the dataset to be analyzed.

We are now in position to prove the following result :

Theorem 4.3.3 (Identifiability) *Under the conditions C1-C4, the model (4.1)-(4.2)-(4.3) is identifiable ; that is, $L_1(\psi) = L_1(\psi^*)$ almost surely implies $\psi = \psi^*$.*

Proof of Theorem 4.3.3

Suppose that $L_1(\psi) = L_1(\psi^*)$ almost surely. Under C1 and C2, there exists a positive constant c_1 such that for every $\mathbf{x} \in F$, $\mathbf{z} \in G$, and $\psi \in \mathscr{B} \times \mathscr{G}$, $c_1 < \mathbb{P}(Y = 1|\mathbf{x}, \mathbf{z}) < 1 - c_1$. Thus we can find a $\omega \in \Omega$, outside the negligible set where $L_1(\psi) \neq L_1(\psi^*)$, and such that $Y(\omega) = 1$ when $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$. For this ω, $L_1(\psi) = L_1(\psi^*)$

becomes

$$\frac{e^{\beta^\top \mathbf{x}+\theta^\top \mathbf{z}}}{(1+e^{\beta^\top \mathbf{x}})(1+e^{\theta^\top \mathbf{z}})} = \frac{e^{\beta^{*\top} \mathbf{x}+\theta^{*\top} \mathbf{z}}}{(1+e^{\beta^{*\top} \mathbf{x}})(1+e^{\theta^{*\top} \mathbf{z}})}.$$

This can be rewritten as

$$\frac{1+e^{-\beta^\top \mathbf{x}}}{1+e^{-\beta^{*\top} \mathbf{x}}} = \frac{1+e^{-\theta^{*\top} \mathbf{z}}}{1+e^{-\theta^\top \mathbf{z}}}. \tag{4.5}$$

Now, under the condition C4, taking the partial derivative of both sides of (4.5) with respect to the l-th component of $\mathbf{x}$ (X_{il} is a continuous covariate) yields

$$\frac{-\beta_l e^{-\beta^\top \mathbf{x}}(1+e^{-\beta^{*\top} \mathbf{x}}) + \beta_l^* e^{-\beta^{*\top} \mathbf{x}}(1+e^{-\beta^\top \mathbf{x}})}{(1+e^{-\beta^{*\top} \mathbf{x}})^2} = 0$$

since the right-hand-side of (4.5) does not depend on $\mathbf{x}$. Thus, it follows that

$$\frac{\beta_l}{\beta_l^*} = \frac{1+e^{\beta^\top \mathbf{x}}}{1+e^{\beta^{*\top} \mathbf{x}}}.$$

Differentiating both sides of this equality with respect to the l-th component of $\mathbf{x}$ further yields $(\beta-\beta^*)^\top \mathbf{x} = 0$, which implies that $\beta = \beta^*$ under C1. It remains to show that $\theta = \theta^*$, which reduces to the identifiability problem in the standard logistic regression model. We have that $\theta = \theta^*$ under C1 (see (Guyon 2001) for example), which concludes the proof.

We now turn to the asymptotic theory for the proposed estimator.

4.4 Asymptotic theory

In this section, we establish rigorously the existence, consistency and asymptotic normality of the maximum likelihood estimator $\widehat{\beta}_n$ of β in model (4.1), obtained from a sample of binary response data with a cure fraction. 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 4.4.1 (Existence and consistency) *Under the conditions C1-C3, the maximum likelihood estimator $\widehat{\psi}_n$ exists almost surely as $n \to \infty$, and converges almost surely to ψ_0, if and only if λ_n tends to infinity as $n \to \infty$.*

Proof of Theorem 4.4.1

The principle of the proof is similar to (Gouriéroux & 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 4.4.2 *Let $\phi_n : \mathbb{R}^k \longrightarrow \mathbb{R}^k$ be defined as*

$$\phi_n(\psi) = \psi + (\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top)^{-1}\dot{l}_n(\psi).$$

Then there exists an open ball $B(\psi_0, r)$ (with $r > 0$) such that ϕ_n satisfies the Lipschitz condition on $B(\psi_0, r)$ that is,

$$\|\phi_n(\psi_1) - \phi_n(\psi_2)\| \leq c\,\|\psi_1 - \psi_2\| \text{ for all } \psi_1, \psi_2 \in B(\psi_0, r), \tag{4.6}$$

and $0 < c < 1$.

Lemma 4.4.3 *The maximum likelihood estimator $\widehat{\psi}_n$ exists almost surely as $n \to \infty$, and converges almost surely to ψ_0, if and only if $(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top)^{-1}\dot{l}_n(\psi_0)$ converges almost surely to 0.*

Proof of Lemma 4.4.3

We first prove that the condition is sufficient. Thus, we assume that $(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top)^{-1}\dot{l}_n(\psi_0)$ converges almost surely to 0. Define $\eta_n(\psi) = \psi - \phi_n(\psi) = -(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top)^{-1}\dot{l}_n(\psi)$ and let ε be an arbitrary positive value. Then for almost every $\omega \in \Omega$, there exists an integer value $n(\varepsilon, \omega)$ such that for any $n \geq n(\varepsilon, \omega)$, $\|\eta_n(\psi_0)\| \leq \varepsilon$ or equivalently, $0 \in B(\eta_n(\psi_0), \varepsilon)$. In particular, let $\varepsilon = (1-c)s$ with $0 < c < 1$ such as in Lemma 4.4.2. Since ϕ_n satisfies the Lipschitz condition (4.6) (by Lemma 4.4.2), the lemma 2 of (Gouriéroux & Monfort 1981) ensures that there exists an element of $B(\psi_0, s)$ (let denote this element by $\widehat{\psi}_n$) such that $\eta_n(\widehat{\psi}_n) = 0$ that is,

$$(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top)^{-1}\dot{l}_n(\widehat{\psi}_n) = 0.$$

The condition C3 implies that $\dot{l}_n(\widehat{\psi}_n) = 0$ and that $\widehat{\psi}_n$ 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 $\widehat{\psi}_n$ exists, and $\|\widehat{\psi}_n - \psi_0\| \leq s$ (that is, $\widehat{\psi}_n$ converges almost surely to ψ_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 $n \to \infty$, $\widehat{\psi}_n$ exists and converges almost surely to ψ_0, but $\eta_n(\psi_0)$ does not converge almost surely to 0. Then there exists a set $\widetilde{\Omega} \subset \Omega$ with $\mathbb{P}(\widetilde{\Omega}) > 0$, such that if $\omega \in \widetilde{\Omega}$, there exists $\varepsilon > 0$ such that for every $m \in \mathbb{N}$, there exists $n \geq m$ with $\|\eta_n(\psi_0)\| > \varepsilon$. Now, let $t = \frac{\varepsilon}{d(1+c)}$, with $d > 1$ sufficiently large so that $t \leq r$, where r is such as in Lemma 4.4.2. Then for every $\psi \in B(\psi_0, t)$, the following holds :

$$\begin{aligned}
\|\eta_n(\psi_0) - \eta_n(\psi)\| &= \|\psi_0 - \phi_n(\psi_0) - \psi + \phi_n(\psi)\| \\
&\leq \|\psi_0 - \psi\| + \|\phi_n(\psi) - \phi_n(\psi_0)\| \\
&\leq t(1+c) = \frac{\varepsilon}{d},
\end{aligned}$$

where the second to third line follows by Lemma 4.4.2. Therefore, for every $\psi \in B(\psi_0, t)$,

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

and we conclude that for every $\psi \in B(\psi_0, t)$, $\|\eta_n(\psi)\| > \varepsilon(1 - \frac{1}{d}) > 0$. Since $\eta_n(\widehat{\psi}_n) = 0$, $\widehat{\psi}_n$ cannot belong to $B(\psi_0, t)$ for large n, which implies that $\widehat{\psi}_n$ does not converge almost surely to ψ_0. This is the desired contradiction.

Lemma 4.4.4 $(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top)^{-1}\dot{l}_n(\psi_0)$ *converges almost surely to* 0 *if and only if* λ_n *tends to infinity as* $n \to \infty$.

Proof of Lemma 4.4.4

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

$$\mathbb{E}[U] = 0 \text{ and } \operatorname{var}[U] = I_{2n}, \tag{4.7}$$

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

$$\begin{aligned}
\mathbb{E}[U] &= \mathbb{E}[\mathbb{E}[(\mathbb{D}(\psi_0))^{-\frac{1}{2}}C(\psi_0)|\mathbb{X}, \mathbb{Z}]] \\
&= \mathbb{E}[(\mathbb{D}(\psi_0))^{-\frac{1}{2}}\mathbb{E}[C(\psi_0)|\mathbb{X}, \mathbb{Z}]] \\
&= \mathbb{E}[(\mathbb{D}(\psi_0))^{-\frac{1}{2}}\mathbb{E}[\left((A^\beta(\psi_0) - B^\beta(\psi_0))^\top, (A^\theta(\psi_0) - B^\theta(\psi_0))^\top\right)^\top |\mathbb{X}, \mathbb{Z}]].
\end{aligned}$$

For every $i = 1, \ldots, n$, $\mathbb{E}[A_i^\beta(\psi_0) - B_i^\beta(\psi_0)|\mathbb{X}, \mathbb{Z}]] = \mathbb{E}[A_i^\beta(\psi_0) - B_i^\beta(\psi_0)|\mathbf{X}_i, \mathbf{Z}_i]]$ by independence between the individuals, and

$$\begin{aligned}
\mathbb{E}[A_i^\beta(\psi_0) - B_i^\beta(\psi_0)|\mathbf{X}_i, \mathbf{Z}_i]] &= \frac{1 + e^{\theta_0^\top \mathbf{Z}_i}}{1 + e^{\beta_0^\top \mathbf{X}_i} + e^{\theta_0^\top \mathbf{Z}_i}}\mathbb{P}(Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i) - B_i^\beta(\psi_0) \\
&= B_i^\beta(\psi_0) - B_i^\beta(\psi_0) \\
&= 0.
\end{aligned}$$

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

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

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

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

$$\begin{bmatrix} \mathbb{V}_1 & \mathbb{V}_3 \\ \mathbb{V}_3 & \mathbb{V}_2 \end{bmatrix}$$

where $\mathbb{V}_1, \mathbb{V}_2$, and $\mathbb{V}_3$ are $(n \times n)$ matrices. The i-th diagonal elements ($i = 1, \ldots, n$) of $\mathbb{V}_1, \mathbb{V}_2$, and $\mathbb{V}_3$ are $\mathrm{var}[A_i^\beta(\psi_0)|\mathbb{X}, \mathbb{Z}]$, $\mathrm{var}[A_i^\theta(\psi_0)|\mathbb{X}, \mathbb{Z}]$, and $\mathrm{cov}[A_i^\beta(\psi_0), A_i^\theta(\psi_0)|\mathbb{X}, \mathbb{Z}]$ respectively. Similar calculations as above yield : $\mathrm{var}[A_i^\beta(\psi_0)|\mathbb{X}, \mathbb{Z}] = \mathbb{D}_{1,ii}(\psi_0)$, $\mathrm{var}[A_i^\theta(\psi_0)|\mathbb{X}, \mathbb{Z}] = \mathbb{D}_{2,ii}(\psi_0)$, and $\mathrm{cov}[A_i^\beta(\psi_0), A_i^\theta(\psi_0)|\mathbb{X}, \mathbb{Z}] = \mathbb{D}_{3,ii}(\psi_0)$. Note also that $\mathbb{V}_1, \mathbb{V}_2$, and $\mathbb{V}_3$ are diagonal matrices, by independence between the individuals. It follows that $\mathrm{var}[C(\psi_0)|\mathbb{X}, \mathbb{Z}] = \mathbb{D}(\psi_0)$ and thus, $\mathrm{var}[U|\mathbb{X}, \mathbb{Z}] = I_{2n}$ and $\mathrm{var}[U] = I_{2n}$.

By (Gouriéroux & Monfort 1981) (proof of Lemma 4), if (4.7) holds, $\Lambda_n/\lambda_n < c_2$ for every $n = 1, 2, \ldots$, and λ_n tends to infinity as $n \to \infty$, then

$$(\mathbb{V}^\top \mathbb{V})^{-1}\mathbb{V}^\top U \xrightarrow{a.s.} 0 \text{ as } n \to \infty$$

that is, $(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top)^{-1}\dot{l}_n(\psi_0)$ converges almost surely to 0.

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

Finally, Theorem 4.4.1 follows by Lemma 4.4.3 and Lemma 4.4.4.

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

Theorem 4.4.5 (Asymptotic normality) *Assume that the conditions C1-C3 hold and that $\widehat{\psi}_n$ converges almost surely to ψ_0. Let $\widehat{\Sigma}_n = \mathbb{W}\mathbb{D}(\widehat{\psi}_n)\mathbb{W}^\top$ 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 $\mathscr{N}(0, I_k)$.*

Proof of Theorem 4.4.5

A Taylor expansion of the score function is as

$$0 = \dot{l}_n(\widehat{\psi}_n) = \dot{l}_n(\psi_0) + \ddot{l}_n(\widetilde{\psi}_n)(\widehat{\psi}_n - \psi_0)$$

where $\widetilde{\psi}_n$ lies between $\widehat{\psi}_n$ and ψ_0, and thus $\dot{l}_n(\psi_0) = -\ddot{l}_n(\widetilde{\psi}_n)(\widehat{\psi}_n - \psi_0)$. Let $\widetilde{\Sigma}_n := -\ddot{l}_n(\widetilde{\psi}_n) = \mathbb{W}\mathbb{D}(\widetilde{\psi}_n)\mathbb{W}^\top$ and $\Sigma_{n,0} := \mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top$. Now,

$$\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0) = \left[\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}}\right]\left[\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}}\right]\Sigma_{n,0}^{-\frac{1}{2}}\left(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0)\right). \tag{4.8}$$

The two terms in brackets in (4.8) converge almost surely to I_k. To see this, we show for example that $\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\right\| \xrightarrow{a.s.} 0$ as $n \to \infty$. First, note that

$$\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\right\| \leq \Lambda_n^{\frac{1}{2}}\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\right\|\left\|\Lambda_n^{-\frac{1}{2}}\left(\Sigma_{n,0}^{\frac{1}{2}} - \widetilde{\Sigma}_n^{\frac{1}{2}}\right)\right\|, \tag{4.9}$$

and

$$\Lambda_n^{-1}\left\|\Sigma_{n,0} - \widetilde{\Sigma}_n\right\| = \Lambda_n^{-1}\left\|\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{W}^\top\right\|.$$

Note also that $\widetilde{\psi}_n$ converges almost surely to ψ_0 (that is, for every $\omega \in \breve{\Omega}$, where $\breve{\Omega} \subset \Omega$ and $\mathbb{P}(\breve{\Omega}) = 1$). Let $\omega \in \breve{\Omega}$. By the same arguments as in the proof of Lemma 4.4.2, for every $\varepsilon > 0$, there exists a positive $n(\varepsilon, \omega) \in \mathbb{N}$ such that if $n \geq n(\varepsilon, \omega)$, then $\Lambda_n^{-1}\left\|\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{W}^\top\right\| \leq \varepsilon$. Hence $\Lambda_n^{-1}\left\|\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{W}^\top\right\|$ converges almost surely to 0. By continuity of the map $x \mapsto x^{\frac{1}{2}}$, $\left\|\Lambda_n^{-\frac{1}{2}}\left(\Sigma_{n,0}^{\frac{1}{2}} - \widetilde{\Sigma}_n^{\frac{1}{2}}\right)\right\|$ converges also almost surely to 0. Moreover, for n sufficiently large, there exists a positive constant $c_4 < \infty$ such that almost surely, $\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\right\| \leq c_4 \lambda_n^{-\frac{1}{2}}$. It follows from (4.9) and the condition C3 that $\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\right\|$ converges almost surely to 0. The almost sure convergence to 0 of $\left\|\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}} - I_k\right\|$ follows by similar arguments.

It remains for us to show that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0))$ converges in distribution to $\mathcal{N}(0, I_k)$, or equivalently, that $(\mathbb{V}^\top\mathbb{V})^{-\frac{1}{2}}\mathbb{V}^\top U$ converges in distribution to $\mathcal{N}(0, I_k)$. Following (Eicker 1966), this convergence holds if we can check the following conditions : i) $\max_{1\leq i\leq 2n} \mathbb{V}_{i\bullet}\Sigma_{n,0}^{-1}\mathbb{V}_{i\bullet}^\top \longrightarrow 0$ as $n \to \infty$, ii) $\sup_{1\leq i\leq 2n} \mathbb{E}[U_i^2 1_{\{|U_i|>\alpha\}}] \longrightarrow 0$ as $\alpha \to \infty$, iii) $\inf_{1\leq i\leq 2n} \mathbb{E}[U_i^2] > 0$, where $\mathbb{V}_{i\bullet}$ 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

$$0 \leq \max_{1\leq i\leq 2n} \mathbb{V}_{i\bullet}\Sigma_{n,0}^{-1}\mathbb{V}_{i\bullet}^\top \leq \max_{1\leq i\leq 2n} \|\mathbb{V}_{i\bullet}\|^2 \left\|\Sigma_{n,0}^{-1}\right\| = \max_{1\leq i\leq 2n} \frac{1}{\lambda_n} \|\mathbb{V}_{i\bullet}\|^2,$$

and that $\|\mathbb{V}_{i\bullet}\|$ is bounded above, by C1 and C2. Moreover, $\frac{1}{\lambda_n}$ tends to 0 as $n \to \infty$, since $\widehat{\psi}_n$ converges almost surely to ψ_0. Condition ii) follows by noting that the components U_i of U are bounded under C1 and C2. Finally, for every $i = 1, \ldots, 2n$, $\mathbb{E}[U_i^2] = \mathrm{var}[U_i]$ since U is centered. We have proved (see Lemma 4.4.4) that $\mathrm{var}[U] = I_{2n}$, thus for every $i = 1, \ldots, 2n$, $\mathrm{var}[U_i] = 1$, and finally, $\inf_{1\leq i\leq 2n} \mathbb{E}[U_i^2] = 1 > 0$.

To summarize, we have proved that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0))$ converges in distribution to $\mathcal{N}(0, I_k)$. This result, combined with Slutsky's theorem and equation (4.8), implies that $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0)$ converges in distribution to $\mathcal{N}(0, I_k)$.

The asymptotic distribution of the maximum likelihood estimator $\widehat{\beta}_n$ of the parameter of interest β in the model is given by the following corollary (the proof is done in Appendix B) :

Corollary 4.4.6 *Let M be the $(p \times (p+q))$ block-matrix $[I_p, 0_{p,q}]$, where $0_{p,q}$ is the $(p \times q)$ matrix whose components are all equal to 0. Then $\sqrt{n}(\widehat{\beta}_n - \beta)$ converges in distribution to a zero-mean Gaussian vector with covariance matrix $M\mathscr{I}_\psi^{-1}M^\top$, which is the upper-left $(p \times p)$ block of $\mathscr{I}^{-1}(\psi)$.*

The convergence in distribution of $\widehat{\beta}_n$ can be used to make statistical inference about β. For example, if one wishes to test the null hypothesis $H_0 : \beta_l = 0$ against the alternative $H_1 : \beta_l \neq 0$ (for some $1 \leq l \leq p$), one can use a Wald-type test, which rejects H_0 at the asymptotic level α $(0 < \alpha < 1)$ if

$$\left| \frac{\widehat{\beta}_{n,l}}{\sqrt{\frac{(M\widehat{\mathscr{I}}^{-1}(\widehat{\psi}_n)M^\top)_{ll}}{n}}} \right| > u_{1-\frac{\alpha}{2}},$$

where $u_{1-\frac{\alpha}{2}}$ is the quantile of order $(1 - \frac{\alpha}{2})$ of the standard normal distribution, $\widehat{\beta}_{n,l}$ is the l-th component of $\widehat{\beta}_n$, and $(M\widehat{\mathscr{I}^{-1}}(\widehat{\psi}_n)M^\top)_{ll}$ denotes the l-th diagonal component of $M\widehat{\mathscr{I}^{-1}}(\widehat{\psi}_n)M^\top$.

In logistic regression, it is also of interest to estimate the probability $p(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}, S = 1)$ of infection for a given value $\mathbf{x}$ of the covariate. An obvious estimator of $p(\mathbf{x})$ is $\widehat{p}_n(\mathbf{x}) := \exp(\widehat{\beta}_n^\top \mathbf{x})/(1 + \exp(\widehat{\beta}_n^\top \mathbf{x}))$. Its asymptotic properties are summarized in the theorem below. It's proof is postponed in Appendix A

Theorem 4.4.7 *Let $\mathbf{x}$ be a given value of the covariate $\mathbf{X}$. As n tends to infinity, $\sqrt{n}(\widehat{p}_n(\mathbf{x}) - p(\mathbf{x}))$ converges in distribution to a zero-mean Gaussian random variable with variance* $\exp(2\beta^\top \mathbf{x}).\mathbf{x}^\top M\mathscr{I}_\psi^{-1}M^\top \mathbf{x}/(1 + \exp(\beta^\top \mathbf{x}))^4$.

4.5 A simulation study

4.5.1 Study design

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

$$\begin{cases} \log\left(\frac{\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}{1-\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}\right) = \beta_1 + \beta_2 X_{i2} & \text{if } S_i = 1 \\ \mathbb{P}(Y=1|\mathbf{X}_i,S_i) = 0 & \text{if } S_i = 0 \end{cases}$$

and the immunity status :

$$\log\left(\frac{\mathbb{P}(S=1|\mathbf{Z}_i)}{1-\mathbb{P}(S=1|\mathbf{Z}_i)}\right) = \theta_1 + \theta_2 Z_{i2},$$

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 $(Y, S, \mathbf{X}, \mathbf{Z})$ is generated from this model, and for each individual i, we get a realization $(y_i, s_i, \mathbf{x}_i, \mathbf{z}_i)$, where s_i is considered as unknown if $y_i = 0$. A maximum likelihood estimator $\widehat{\beta}_n$ of $\beta = (\beta_1, \beta_2)$ is obtained from this incomplete dataset by solving the score equation (4.4), using the `optim` function of the software `R`. An estimate is also obtained for $\theta = (\theta_1, \theta_2)$, but θ is not the primary parameter of interest hence we only focus on the simulation results for $\widehat{\beta}_n$.

The finite-sample behavior of the maximum likelihood estimator $\widehat{\beta}_n$ 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 β (the parameter of interest) and θ (the nuisance parameter). The following values were considered for β : i) model $\mathscr{M}_1 : \beta = (-.8, 1)$ (using these values, approximately 30% of the susceptibles are infected), ii) model $\mathscr{M}_2 : \beta = (1, .7)$ (approximately 70% of the susceptibles are infected), iii) model $\mathscr{M}_3 : \beta = (-.8, 0)$ (approximately 30% of the susceptibles are infected), iv) model $\mathscr{M}_4 : \beta = (1, 0)$ (approximately 70% of the susceptibles are infected).

4.5.2 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 β_1 and β_2, which are calculated as $N^{-1} \sum_{j=1}^{N} \widehat{\beta}_{1,n}^{(j)}$ and $N^{-1} \sum_{j=1}^{N} \widehat{\beta}_{2,n}^{(j)}$, where $\widehat{\beta}_n^{(j)} = (\widehat{\beta}_{1,n}^{(j)}, \widehat{\beta}_{2,n}^{(j)})$ is the estimate obtained from the j-th simulated sample. For each of the parameters β_1 and β_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 $\mathscr{M}_1$ and $\mathscr{M}_2$, see Tables 4.1 and 4.2) (respectively models $\mathscr{M}_3$ and $\mathscr{M}_4$, see Tables 4.1 and 4.2). 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 4.2 and 4.3.

From these tables, it appears that the proposed maximum likelihood estimator

$\widehat{\beta}_n$ provides a reasonable approximation of the true parameter value, even when the percentage of immunes is high. While the bias of $\widehat{\beta}_n$ 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 β_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%. The level of the Wald test for nullity of β_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. The results of such "naive" analysis for model $\mathscr{M}_1$ are given in Table 4.3 (the results for models $\mathscr{M}_2, \mathscr{M}_3, \mathscr{M}_4$ yield similar observations and thus, they are not given here. However, the complete simulation study is available from the web-based supplementary document mentioned above).

From this table, it appears that ignoring the immunity present in the sample results in strongly biased estimates of β. The bias of the intercept estimate increases with the immune proportion. At the same time, the estimate of the regression coefficient β_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 β_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 $\widehat{\beta}_{2,n}$ was also investigated. For each configuration of the design parameters, histograms of the $\widehat{\beta}_{2,n}^{(j)}$ ($j = 1, \ldots, N$) are obtained, along with the corresponding QQ-plots. The plots for the model $\mathscr{M}_1$ are pictured on Figures 4.1 to 4.4 (the plots for the models $\mathscr{M}_2$, $\mathscr{M}_3$, $\mathscr{M}_4$ are given in the web-based file).

From these figures, it appears that the normal approximation stated in Theorem 4.4.5 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

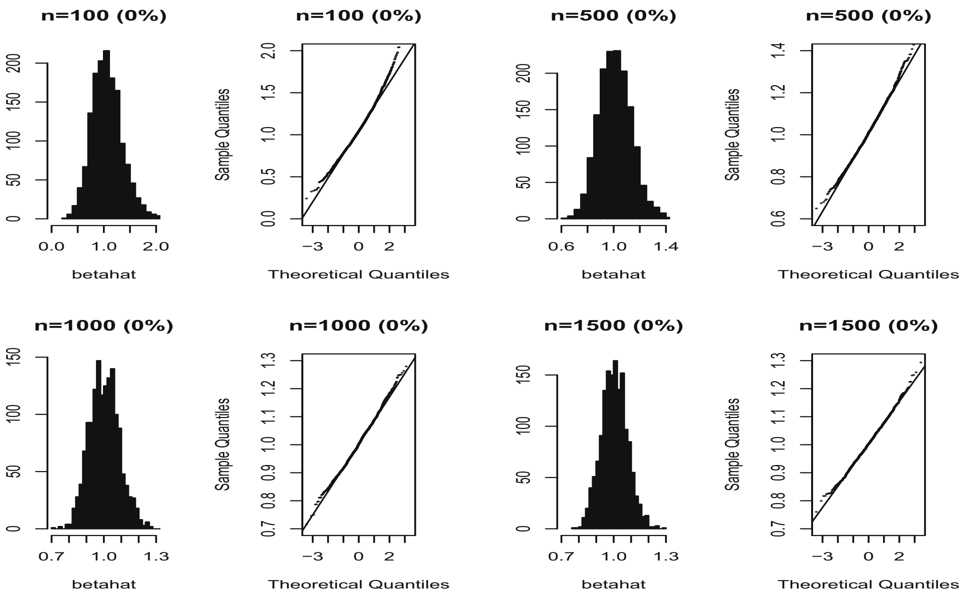

FIG. 4.1 – Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathscr{M}_1$, with no immunes in the sample (the percentage of immunes is given in brackets). n is the sample size. All results are based on 1500 simulated datasets.

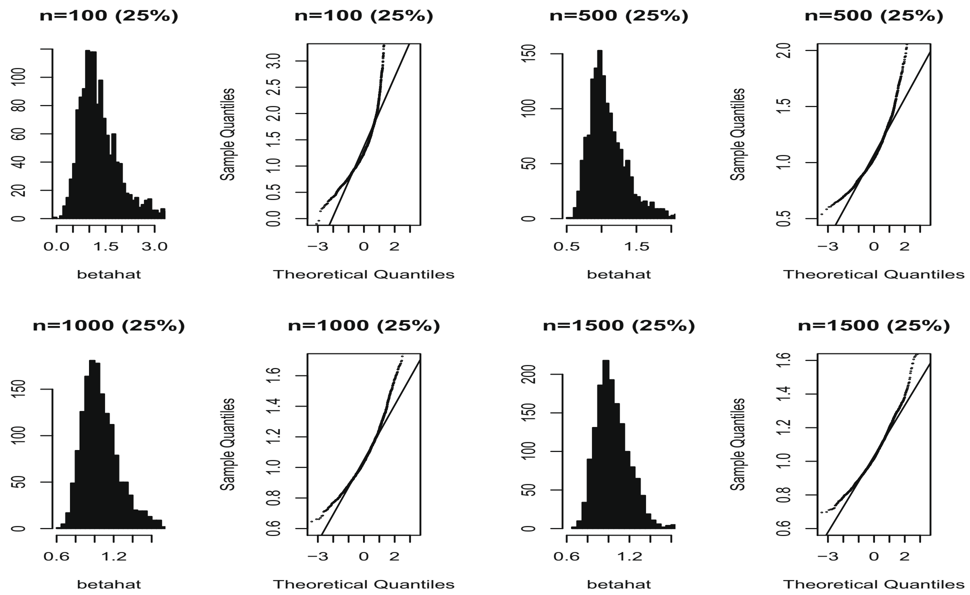

FIG. 4.2 – Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathscr{M}_1$, with 25% of immunes.

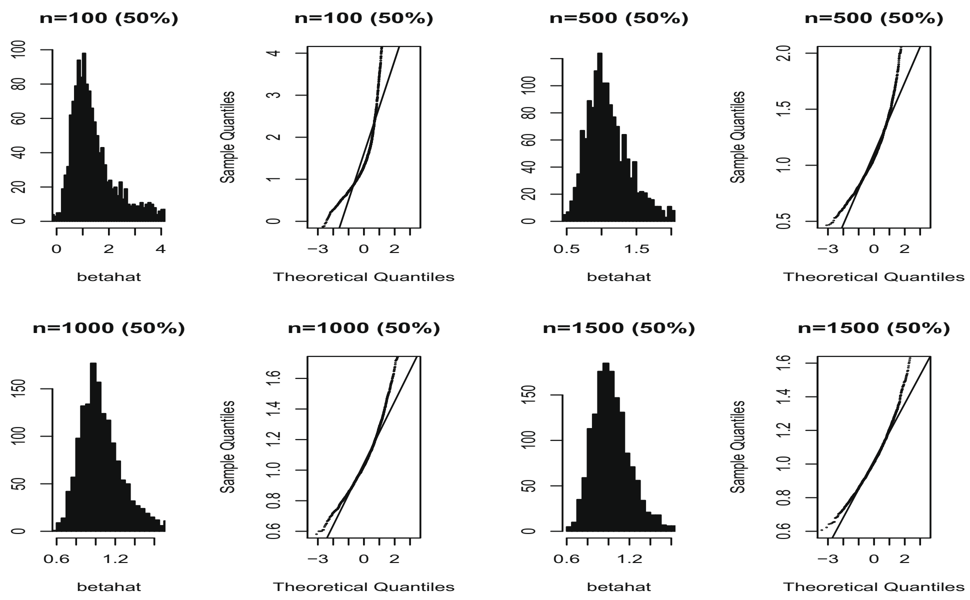

FIG. 4.3 – Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathscr{M}_1$, with 50% of immunes.

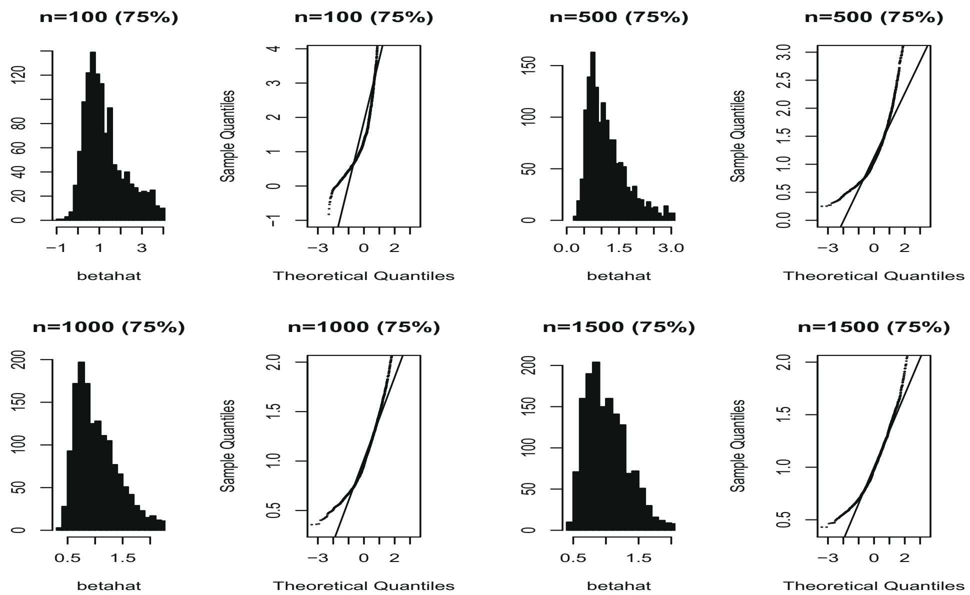

FIG. 4.4 – Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathscr{M}_1$, with 75% of immunes.

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 $\widehat{\beta}_{2,n}$ 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 $\widehat{\beta}_{2,n}$ 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 $\widehat{\beta}_{2,n}$.

Overall, these results indicate that a reliable statistical inference on the regression effect in the model (5.1) with a cure fraction should be based on a sample having, at least, a moderately large size ($n \geq 500$, say) when the immune fraction is moderate (25%), or a large size ($n \geq 1000$, say) when the immune proportion is large (50%).

4.6 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 (the proposed joint model can thus be viewed as a zero-inflated Bernoulli regression model, with logit links for both the binary response of interest and the zero-inflation component). 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 & 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 4.4.2. Recall that I_k denotes the identity matrix of order k. Then we write :

$$\begin{aligned}
\left\| \frac{\partial \phi_n(\psi)}{\partial \psi^\top} \right\| &= \left\| I_k - (\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top)^{-1}\mathbb{W}\mathbb{D}(\psi)\mathbb{W}^\top \right\| \\
&= \left\| (\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top)^{-1}\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{W}^\top \right\| \\
&\leq \left\| (\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top)^{-1} \right\| \left\| \mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{W}^\top \right\| \\
&= \lambda_n^{-1} \left\| \mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{W}^\top \right\|.
\end{aligned}$$

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

$$\begin{aligned}
\left\| \mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{W}^\top \right\| &= \left\| \sum_{i=1}^{2n} \sum_{j=1}^{2n} \mathbb{W}_{\bullet i}\mathbb{W}_{\bullet j}^\top (\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)) \right\| \\
&\leq \sum_{(i,j)\in\mathscr{S}} \left\| \mathbb{W}_{\bullet i}\mathbb{W}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0) \right\| \left| \frac{\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)}{\mathbb{D}_{ij}(\psi_0)} \right|.
\end{aligned}$$

From C1 and C2, there exists a real constant c_3 ($c_3 > 0$) such that $\mathbb{D}_{ij}(\psi_0) > c_3$ for every $(i,j) \in \mathscr{S}$. Moreover, $\mathbb{D}_{ij}(\cdot)$ is uniformly continuous on $\mathscr{B} \times \mathscr{G}$, thus for every $\varepsilon > 0$, there exists a positive r such that for all $\psi \in B(\psi_0, r)$, $|\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)| < \varepsilon$. It follows that

$$\begin{aligned}
\left\| \mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{W}^\top \right\| &\leq \frac{\varepsilon}{c_3} \sum_{(i,j)\in\mathscr{S}} \left\| \mathbb{W}_{\bullet i}\mathbb{W}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0) \right\| \\
&\leq \frac{\varepsilon}{c_3} \mathrm{tr} \left(\sum_{(i,j)\in\mathscr{S}} \mathbb{W}_{\bullet i}\mathbb{W}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0) \right) \\
&= \frac{\varepsilon}{c_3} \mathrm{tr} \left(\sum_{i=1}^{2n} \sum_{j=1}^{2n} \mathbb{W}_{\bullet i}\mathbb{W}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0) \right) \\
&= \frac{\varepsilon}{c_3} \mathrm{tr} \left(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}^\top \right) \\
&\leq \frac{\varepsilon}{c_3} \Lambda_n k.
\end{aligned}$$

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

$\square$

TAB. 4.1 – Simulation results for models $\mathscr{M}_1 : \beta = (-.8, 1)$ and $\mathscr{M}_3 : \beta = (-.8, 0)$

	percentage of immunes in the sample							
	0%		25%		50%		75%	
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$
				Model $\mathscr{M}_1$				
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*
				Model $\mathscr{M}_3$				
100	-0.815	-0.001	-0.721	-0.007	-0.734	0.000	-0.746	-0.004
	(0.224)	(0.229)	(0.465)	(1.341)	(0.800)	(2.109)	(1.966)	(3.258)
	[0.177]	[0.179]	[0.377]	[0.762]	[0.636]	[1.111]	[1.516]	[1.715]
		0.052†		0.077†		0.069†		0.087†
500	-0.801	-0.001	-0.748	0.007	-0.750	0.001	-0.775	-0.006
	(0.097)	(0.099)	(0.280)	(0.415)	(0.520)	(0.469)	(1.209)	(0.711)
	[0.078]	[0.080]	[0.241]	[0.231]	[0.422]	[0.241]	[1.007]	[0.363]
		0.041†		0.058†		0.052†		0.057†
1000	-0.803	-0.001	-0.759	0.008	-0.763	0.005	-0.793	0.005
	(0.067)	(0.066)	(0.221)	(0.237)	(0.367)	(0.266)	(1.154)	(0.312)
	[0.053]	[0.053]	[0.182]	[0.137]	[0.299]	[0.140]	[0.911]	[0.175]
		0.042†		0.045†		0.037†		0.048†
1500	-0.801	0.000	-0.782	0.009	-0.784	0.003	-0.783	0.009
	(0.053)	(0.054)	(0.208)	(0.168)	(0.328)	(0.212)	(1.149)	(0.258)
	[0.042]	[0.043]	[0.178]	[0.099]	[0.267]	[0.102]	[0.901]	[0.144]
		0.051†		0.048†		0.027†		0.039†

Note : n : sample size. $(\cdot)$: root mean square error. $[\cdot]$: mean absolute error. * : empirical power († : 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.

TAB. 4.2 – Simulation results for models $\mathscr{M}_2 : \beta = (1, .7)$ and $\mathscr{M}_4 : \beta = (1, 0)$

	percentage of immunes in the sample							
	0%		25%		50%		75%	
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$
				Model $\mathscr{M}_2$				
100	1.026	0.720	0.945	0.723	0.949	0.740	0.834	0.647
	(0.246)	(0.273)	(0.780)	(0.534)	(0.988)	(0.788)	(1.549)	(1.455)
	[0.196]	[0.215]	[0.655]	[0.376]	[0.829]	[0.555]	[1.326]	[0.933]
		0.746*		0.132*		0.118*		0.088*
500	1.003	0.712	1.098	0.717	1.112	0.721	0.840	0.652
	(0.107)	(0.115)	(0.651)	(0.247)	(0.672)	(0.279)	(0.969)	(0.534)
	[0.086]	[0.091]	[0.518]	[0.202]	[0.534]	[0.230]	[0.802]	[0.421]
		1*		0.503*		0.418*		0.168*
1000	1.003	0.707	1.078	0.711	1.096	0.719	0.842	0.657
	(0.071)	(0.082)	(0.590)	(0.215)	(0.571)	(0.224)	(0.796)	(0.439)
	[0.057]	[0.065]	[0.428]	[0.168]	[0.441]	[0.181]	[0.670]	[0.352]
		1*		0.779*		0.675*		0.205*
1500	1.001	0.701	1.035	0.705	1.069	0.709	0.887	0.655
	(0.064)	(0.065)	(0.450)	(0.163)	(0.466)	(0.177)	(0.604)	(0.312)
	[0.050]	[0.052]	[0.344]	[0.135]	[0.358]	[0.144]	[0.502]	[0.257]
		1*		0.986*		0.926*		0.300*
				Model $\mathscr{M}_4$				
100	1.030	0.001	1.110	0.007	1.154	0.017	0.913	-0.003
	(0.233)	(0.234)	(0.852)	(0.969)	(1.211)	(1.347)	(1.775)	(1.640)
	[0.182]	[0.187]	[0.684]	[0.587]	[0.995]	[0.792]	[1.450]	[0.865]
		$0.058^\dagger$		$0.072^\dagger$		$0.083^\dagger$		$0.066^\dagger$
500	1.007	-0.005	1.105	0.020	1.123	0.054	0.915	-0.009
	(0.103)	(0.103)	(0.609)	(0.293)	(0.690)	(0.318)	(0.817)	(0.370)
	[0.081]	[0.082]	[0.492]	[0.180]	[0.562]	[0.208]	[0.614]	[0.215]
		$0.046^\dagger$		$0.050^\dagger$		$0.063^\dagger$		$0.051^\dagger$
1000	1.003	0.000	1.091	-0.003	1.101	0.033	0.934	-0.003
	(0.071)	(0.070)	(0.521)	(0.198)	(0.578)	(0.210)	(0.757)	(0.256)
	[0.057]	[0.055]	[0.437]	[0.125]	[0.455]	[0.135]	[0.600]	[0.142]
		$0.051^\dagger$		$0.045^\dagger$		$0.042^\dagger$		$0.039^\dagger$
1500	1.003	0.001	1.073	0.009	1.115	0.015	0.934	0.002
	(0.057)	(0.057)	(0.480)	(0.132)	(0.501)	(0.139)	(0.633)	(0.175)
	[0.046]	[0.046]	[0.392]	[0.087]	[0.400]	[0.104]	[0.521]	[0.109]
		$0.042^\dagger$		$0.040^\dagger$		$0.046^\dagger$		$0.047^\dagger$

Note : * : 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 70%.

TAB. 4.3 – "Naive" analysis of model $\mathscr{M}_1 : \beta = (-.8, 1)$

	percentage of immunes in the sample					
	25%		50%		75%	
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$
100	-1.154	0.023	-1.632	0.017	-2.410	0.001
	(0.428)	(1.011)	(0.879)	(1.025)	(1.776)	(1.071)
	[0.365]	[0.977]	[0.833]	[0.983]	[1.610]	[1.003]
		0.049*		0.057*		0.052*
500	-1.128	0.087	-1.594	0.042	-2.305	0.002
	(0.344)	(0.915)	(0.803)	(0.963)	(1.513)	(1.010)
	[0.328]	[0.913]	[0.794]	[0.958]	[1.505]	[0.997]
		0.049*		0.051*		0.053*
1000	-1.131	0.059	-1.590	0.050	-2.297	0.033
	(0.338)	(0.941)	(0.795)	(0.952)	(1.501)	(0.970)
	[0.330]	[0.940]	[0.790]	[0.950]	[1.497]	[0.966]
		0.053*		0.051*		0.054*
1500	-1.127	0.050	-1.591	0.046	-2.302	0.039
	(0.332)	(0.953)	(0.794)	(0.955)	(1.504)	(0.962)
	[0.327]	[0.952]	[0.791]	[0.954]	[1.502]	[0.960]
		0.051*		0.050*		0.053*

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.

Chapitre 5

Bandes de confiance simultanées dans le modèle de régression logistique avec fraction immune

Sommaire

5.1 Introduction

Dans ce chapitre nous nous proposons d'étudier des bandes de confiance simultanées pour la fonction réponse $\mathbf{P}(Y = 1|\mathbf{X} = x, S)$ définie dans le modèle ZIB (4.1)-(4.2) dans le chapitre 4. Les bandes de confiance simultanées fournissent des informations utiles sur le domaine dans lequel devrait se situer la vraie fonction de régression qui est inconnue, et leur construction constitue un problème difficile lorsque le nombre de variables explicatives est supérieur à 1. La construction de bandes de confiance remonte à (Working & Hotelling 1929).

Il existe plusieurs études récentes considérant les applications des bandes de confiance. Par exemple, (Sun *et al.* 1999) ont utilisé les bandes de confiance simultanées pour faire de l'inférence sur la croissance et des courbes de réponse, (Al-Saidy *et al.* 2003) et (Piegorsch *et al.* 2005) ont utilisé les bandes de confiance dans l'analyse quantitative des risques, (Spurrier 1999), (Bhargava & Spurrier 2004), (Liu *et al.* 2004) et (Liu *et al.* 2007) pour des comparaisons simultanées de plusieurs modèles de régression linéaires dans certains problèmes médicaux, (Zhang & Peng 2010) ont utilisé les bandes de confiance pour analyser des données sur l'utilisation des contraceptifs, tandis que (Azaïs *et al.* 2010) ont utilisé les bandes de confiance pour la prédiction de courbes de charge annuelles en électricité.
Les bandes de confiance pour des modèles de régression linéaires multiples (il y'a au moins deux prédicteurs linéaires et l'espace $\mathscr{X}$ peut prendre diverses formes) sont beaucoup plus difficiles à construire. Nous pouvons cependant noter certains travaux incluant ceux de (Liu 2011), de (Hauck 1983) dans le cas où il n'existe aucune contrainte sur l'espace $\mathscr{X}$ contenant les covariables ($\mathscr{X} \subseteq \mathbb{R}^{p-1}$), ceux de (Bohrer 1973), (Casella & Strawderman 1980) et (Seppanen & Uusipaikka 1992)

dans le cas où les covariables sont restreintes à une hyper-ellipsoïde. Dans le cas où les covariables sont restreintes à un hyper-rectangle, le lecteur pourra se référer aux travaux de (Naiman 1987), (Naiman 1990), (Sun & Loader 1994), (Sun *et al.* 2000) et (Liu *et al.* 2005) . En particulier, (Sun & Loader 1994) supposent que les $p-1$ variables prédictives sont des fonctions de $q \geq 1$ variables indépendantes (par exemple dans les modèles de régression polynomiale) et ont fourni des bandes de confiance hyperboliques pour le modèle de régression lorsqu'il y a contraintes sur les q variables indépendantes $q = 1$ et $q = 2$.
Nous commençons par décrire d'abord dans la section 5.2 le modèle ZIB défini par (Diop *et al.* 2011). Ensuite nous utilisons les résultats asymptotiques de l'estimateur du maximum de vraisemblance $\widehat{\psi}_n$ de ψ pour montrer un résultat de convergence faible de processus gaussien. Dans la section 5.3, nous présentons trois méthodes pour la construction de bandes de confiance simultanées pour la fonction réponse $\{p(x),\ x \in \mathscr{X}\}$. Enfin dans la section 5.4, nous présentons une étude de simulations pour étudier la précision des bandes de confiance.

5.2 Modèle

Soit $\mathscr{O}_i = (Y_i, S_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$ des copies i.i.d. du vecteur aléatoire $\mathscr{O} = (Y, S, \mathbf{X}, \mathbf{Z})$. A partir de cet échantillon, nous considérons le modèle ZIB suivant, défini dans le chapitre 3 par :

$$\begin{cases} \log\left(\frac{\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}{1-\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}\right) = \beta^\top \mathbf{X}_i & \text{if } \{S_i = 1\} \\ \mathbb{P}(Y=0|\mathbf{X}_i, S_i) = 1 & \text{if } \{S_i = 0\} \end{cases} \tag{5.1}$$

et par le modèle suivant pour le statut d'immunité :

$$\log\left(\frac{\mathbb{P}(S=1|\mathbf{Z}_i)}{1-\mathbb{P}(S=1|\mathbf{Z}_i)}\right)=\theta^\top \mathbf{Z}_i. \tag{5.2}$$

(Diop *et al.* 2011) ont établi l'existence et la consistance de l'estimateur du maximum de vraisemblance $\widehat{\psi}_n$ du vecteur de paramètre $\psi = (\beta^\top, \theta^\top)^\top$ (voir théorème 4.4.1), et la distribution asymptotique de cet estimateur (voir Théorème 4.4.5). Nous pouvons ainsi en déduire la distribution asymptotique de l'estimateur $\widehat{\beta}_n$ du paramètre d'intérêt β (voir corollaire 4.4.6). Dans la suite nous considérons les notations suivantes.

5.2.1 Notations

Soit les notations suivantes : $M \in \mathscr{M}(p \times (p+q))$ désigne la matrice par blocs $[I_p, 0_{p,q}]$ et $0_{p,q} \in \mathscr{M}(p \times q)$ représente une matrice où toutes les composantes sont égales à 0. Nous notons $\widehat{\Sigma}_{\beta,n} = M\widehat{I}^{-1}(\widehat{\psi}_n)M^\top$ la partie de la matrice $\widehat{I}^{-1}(\widehat{\psi}_n)$ restreinte au paramètre β. Nous posons ensuite $\widehat{\sigma}_n^2(x) = x^\top \widehat{\Sigma}_{\beta,n} x$ et $\sigma^2(x) = x^\top \Sigma_\beta x$. Notons $< .,. >$ le produit scalaire usuel sur $\mathbb{R}^p$, $\|.\|$ la norme associée et $|||.|||$ la norme matricielle subordonnée à $\|.\|$.

5.2.2 Supremum de processus Gaussiens

La construction de bande de confiance peut se faire de manière classique en utilisant certaines méthodes existant déjà dans la littérature telles que la méthode de

Scheffé par exemple (voir (Govindarajulu 2001) et (Liu 2011)) que nous présenterons dans la suite. Nous utiliserons ensuite la théorie des processus Gaussiens pour construire d'autres types de bandes de confiance pour $\{p(x),\ x \in \mathscr{X}\}$. Nous commençons par construire des bandes de confiance pour $\{\eta(x) = \beta^\top x,\ x \in \mathscr{X}\}$ de la forme

$$\left\{\widehat{\eta}_n(x) \pm \lambda \frac{\widehat{\sigma}_n(x)}{\sqrt{n}}, x \in \mathscr{X}\right\}. \tag{5.3}$$

Ensuite nous en déduirons les bandes pour la probabilité $\{p(x),\ x \in \mathscr{X}\}$, où λ est un nombre réel. Plus précisément, pour un niveau de confiance $1-\alpha \in (0,1)$ donné, nous cherchons la valeur $\lambda = \lambda_\alpha$ qui satisfait approximativement :

$$\mathbf{P}\Big(\sup_{x\in\mathscr{X}} |W(x)| \leq \lambda_\alpha\Big) = 1-\alpha, \tag{5.4}$$

où W est un processus Gaussien de moyenne zéro et de fonction de covariance ρ, et où $\sigma^2(x) = \rho(x,x)$. Des études de majoration de suprémum de processus Gaussien ont été faites pour seulement un nombre faible de cas particuliers (voir (Adler & Taylor 2007)).

En utilisant le théorème 4.4.5 et le théorème de Slutsky (voir Annexe A), les bandes de confiance définies dans (5.3) avec une valeur λ choisie comme dans (5.4) aura approximativement une probabilité de couverture égale à $1-\alpha$. Le résultat suivant fournit une méthode pour calculer λ.

Theorem 5.2.1 *Sous les conditions (**A1**)-(**A4**) définies dans le chapitre 4, le processus* $\{W_n(x), x \in \mathscr{X}\}$ *défini par :*

$$W_n(x) = \sqrt{n}\frac{x^\top(\widehat{\beta}_n - \beta)}{\widehat{\sigma}_n(x)} \tag{5.5}$$

converge faiblement quand $n \to +\infty$ dans l'espace $C(\mathscr{X})$ muni de la norme uniforme vers un processus Gaussien centré W de fonction de covariance donnée par

$$\rho(x,y) = \frac{x^\top \Sigma_\beta y}{\sigma(x)\sigma(y)}, \quad x,y \in \mathscr{X} \tag{5.6}$$

Grâce au Théorème 5.2.1, nous avons, pour tout $\lambda > 0$, quand $n \to +\infty$, la convergence suivante

$$\mathbf{P}\Big(\sup_{x\in\mathscr{X}} |W_n(x)| \leq \lambda\Big) \to \mathbf{P}\Big(\sup_{x\in\mathscr{X}} |W(x)| \leq \lambda\Big). \tag{5.7}$$

L'importance pratique de l'expression 5.7 est qu'elle permet de quantifier la valeur λ définie dans 5.4, et donc s'avérera être très utile dans les études de simulations. Pour ainsi construire les bandes de confiance en utilisant la théorie des processus Gaussiens, nous disposons principalement de deux méthodes : la première utilise l'inégalité classique de (Landau & Sheep 1970) et la seconde consiste à faire du Bootstrap. Nous présenterons ces deux méthodes dans la suite. Nous commençons d'abord par prouver le Théorème 5.2.1.
Le lemme suivant établit la convergence uniforme de la varaince $\widehat{\sigma}_n^2(x)$.

Lemme 5.2.2

$$\sup_{x\in\mathscr{X}} |\widehat{\sigma}_n^2(x) - \widehat{\sigma}^2(x)| \xrightarrow{P} 0, \ \textit{quand } n \to \infty.$$

Preuve du lemme 5.2.2 :

$$\begin{aligned}
\sup_{x\in\mathscr{X}} |\widehat{\sigma}_n^2(x)-\sigma^2(x)| &= \sup_{x\in\mathscr{X}} |x'M(\widehat{I}^{-1}(\widehat{\psi}_n)-I^{-1}(\psi))M'x| \\
&= \sup_{x\in\mathscr{X}} |<x, M(\widehat{I}^{-1}(\widehat{\psi}_n)-I^{-1}(\psi))M'x>| \\
&\leq \sup_{x\in\mathscr{X}} \|x\|\|M(\widehat{I}^{-1}(\widehat{\psi}_n)-I^{-1}(\psi))M'x\| \quad \text{par Cauchy-Schwartz} \\
&= \sup_{x\in\mathscr{X},\|x\|\neq 0} \|x\|^2 \frac{\|M(\widehat{I}^{-1}(\widehat{\psi}_n)-I^{-1}(\psi))M'x\|}{\|x\|} \\
&\leq \sup_{x\in\mathscr{X},\|x\|\neq 0} \|x\|^2 \sup_{x\in\mathscr{X},\|x\|\neq 0} \frac{\|M(\widehat{I}^{-1}(\widehat{\psi}_n)-I^{-1}(\psi))M'x\|}{\|x\|} \\
&\leq \sup_{x\in\mathscr{X},\|x\|\neq 0} \|x\|^2 |||M(\widehat{I}^{-1}(\widehat{\psi}_n)-I^{-1}(\psi))M'|||,
\end{aligned}$$

Puisque $\widehat{I}^{-1}(\widehat{\psi}_n)$ est un estimateur consistant de $I^{-1}(\psi)$, donc par continuité de la norme $|||M(\widehat{I}^{-1}(\widehat{\psi}_n)-I^{-1}(\psi))M'|||$ tend en probabilité vers 0 quand $n\to\infty$. De plus, x est borné. D'où le résultat.

Preuve du Théorème 5.2.1 :

Considérons le processus $(G_n(x); x\in\mathscr{X}) = (\sqrt{n}(\widehat{\beta}_n-\beta)^\top x,\ x\in\mathscr{X})$ et l'application φ définie par :

$$\begin{aligned}
\mathbb{R}^p &\xrightarrow{\varphi} C(\mathscr{X}) \\
a &\longrightarrow \varphi(a) : x \longrightarrow \varphi(a)(x) = a^\top x.
\end{aligned}$$

φ est linéaire. En effet soit a, b et $x\in\mathbb{R}^p$ et $\alpha\in\mathbb{R}$ on a

$$\varphi(a+\alpha b)(x) = (a+\alpha b)^\top x = \varphi(a)(x) + \alpha\varphi(b)(x).$$

L'application φ est également continue car $p<\infty$.

Nous avons ainsi

$$\varphi(\sqrt{n}(\widehat{\beta}_n-\beta)) \Rightarrow \varphi(Z).$$

Par suite, $G_n \Rightarrow G$ dans $C(\mathscr{X})$ quand $n \to \infty$, où $G(x) = Z'x$ ($\forall x \in \mathscr{X}$) désigne un processus. Par la convergence uniforme en probabilité de $\widehat{\sigma}_n^2(x)$, et le Théorème 4.4.7 de Slutsky,

$$W_n = \frac{G_n}{\widehat{\sigma}_n} \Rightarrow W = \frac{G}{\sigma} \quad \text{dans } C(\mathscr{X}).$$

De plus, la fonction de covariance de W est donnée par

$$\begin{aligned}
\rho(x,y) = \operatorname{cov}(W(x), W(y)) &= \mathbb{E}\left[W(x)W(y)\right] \\
&= \frac{1}{\sigma(x)\sigma(y)}\mathbb{E}\left[G(x)G(y)\right] \\
&= \frac{1}{\sigma(x)\sigma(y)}\mathbb{E}\left[Z'x.Z'y\right] \\
&= \frac{1}{\sigma(x)\sigma(y)}x'\mathbb{E}\left[ZZ'\right]y \\
&= \frac{x'MI^{-1}(\psi)M'y}{\sigma(x)\sigma(y)} \\
&= \frac{x'\Sigma_\beta y}{\sigma(x)\sigma(y)}.
\end{aligned}$$

5.3 Bandes de confiance

L'un des problèmes dans les modèles de régression présentant un intérêt est l'estimation de l'intervalle contenant la vraie fonction de régression. Nous nous proposons donc dans cette partie de construire des bandes de confiance de la probabilité d'infection $\{p(x),\ x \in \mathscr{X}\}$ en utilisant trois méthodes. L'idée principale consiste à utiliser la loi asymptotique de l'estimateur du maximum de vraisemblance $\widehat{\beta}_n$ des coefficients de régression β.

Lorsque l'on dispose d'une seule variable explicative, (Brand *et al.* 1973) donnent une méthode permettant de construire des bandes de confiance pour la probabilité d'infection dans le modèle de régression logistique. Ces bandes sont basées sur la loi de l'estimateur du maximum de vraisemblance des paramètres du modèle. (Hauck 1983) donne une méthode alternative, dite méthode de Scheffé, qui suit de près la méthode de (Brand *et al.* 1973) mais demeure cependant la plus facile à appliquer. Nous la présentons dans la suite, et nous l'adaptons au modèle de régression logistique avec une fraction immune.

5.3.1 Méthode 1 : Méthode de Scheffé

Cette méthode a été essentiellement développée par (Hauck 1983) pour le modèle de régression logistique. Elle n'impose aucune restriction aux variables explicatives $X_2, X_3, \ldots, X_p$. Par conséquent aucune restriction sur la nature de l'espace $\mathscr{X}$ contenant les covariables.
Les deux lemmes suivants sont donnés sans démonstration.

Lemme 5.3.1 *Soit $Z \in \mathbb{R}^m$ un vecteur aléatoire suivant une loi normale multivariée $\mathscr{N}(0, V)$. Si V est inversible, alors $Z^\top V^{-1} Z$ suit une loi du χ^2 à m degrés de liberté.*

Lemme 5.3.2 *Soit $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^n$ et $A \in \mathscr{M}(n \times n)$ une matrice inversible telle que $A = B^\top B$. Alors*

1.
$$(X^\top Y)^2 \leq (X^\top AX)(Y^\top AY).$$

2.
$$(X^\top Y)^2 \leq (X^\top AX)(Y^\top A^{-1} Y), \ \textit{si } A^{-1} \textit{existe}.$$

Ces deux lemmes nous permettent de construire des bandes de confiance de type Scheffé pour la probabilité d'infection dans le modèle de régression logistique avec fraction immune. D'après le Lemme 5.3.1 $n(\widehat{\beta}_n - \beta)^\top \widehat{\Sigma}_{\beta,n}^{-1} (\widehat{\beta}_n - \beta)$ converge en loi vers une χ^2_p. Soit $\alpha \in (0,1)$ et $\chi^2_{p,1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi χ^2_p. Alors quand $n \to \infty$,

$$\mathbf{P}\left[n(\widehat{\beta}_n - \beta)^\top \widehat{\Sigma}^{-1}_{\beta,n}(\widehat{\beta}_n - \beta) \leq \chi^2_{p,1-\alpha}\right] \to 1-\alpha. \tag{5.8}$$

D'après le point 2 du Lemme 5.3.2, on a, pour tout $x \in \mathscr{X}$,

$$\left[\sqrt{n}(\widehat{\beta}_n - \beta)^\top x\right]^2 \leq n(\widehat{\beta}_n - \beta)^\top \widehat{\Sigma}^{-1}_{\beta,n}(\widehat{\beta}_n - \beta)\left[x^\top \widehat{\Sigma}_{\beta,n} x\right]. \tag{5.9}$$

D'où

$$\frac{\left[\sqrt{n}(\widehat{\beta}_n - \beta)^\top x\right]^2}{x^\top \widehat{\Sigma}_{\beta,n} x} \leq n(\widehat{\beta}_n - \beta)^\top \widehat{\Sigma}^{-1}_{\beta,n}(\widehat{\beta}_n - \beta). \tag{5.10}$$

Ce qui implique que

$$\mathbf{P}\Big(n(\widehat{\beta}_n - \beta)^\top \widehat{\Sigma}^{-1}_{\beta,n}(\widehat{\beta}_n - \beta) \leq \chi^2_{p,1-\alpha}\Big) \leq \mathbf{P}\Big(\frac{\left[\sqrt{n}(\widehat{\beta}_n - \beta)^\top x\right]^2}{\widehat{\sigma}^2_n(x)} \leq \chi^2_{p,1-\alpha},\ \forall\, x \in \mathscr{X}\Big),$$

où

$$\widehat{\sigma}^2_n(x) = x^\top \widehat{\Sigma}_{\beta,n} x.$$

Il suit d'après (5.8)

$$\begin{aligned} 1-\alpha &\leq \lim_{n\to\infty} \mathbf{P}\Big(\frac{\left[\sqrt{n}(\widehat{\beta}_n - \beta)^\top x\right]^2}{\widehat{\sigma}^2_n(x)} \leq \chi^2_{p,1-\alpha},\ \forall\, x \in \mathscr{X}\Big) \\ &\leq \lim_{n\to\infty} \Big(|(\widehat{\beta}_n - \beta)^\top x| \leq \widehat{\sigma}_n(x)\sqrt{\frac{\chi^2_{p,1-\alpha}}{n}}, \forall\, x \in \mathscr{X}\Big) \\ &\leq \lim_{n\to\infty} \mathbf{P}\Big(\widehat{l}_n(x) \leq \beta^\top x \leq \widehat{u}_n(x), \forall\, x \in \mathscr{X}\Big), \end{aligned}$$

où

$$\widehat{l}_n(x) = \widehat{\beta}_n^\top x - \widehat{\sigma}_n(x)\sqrt{\frac{\chi^2_{p,1-\alpha}}{n}} \quad \text{et} \quad \widehat{u}_n(x) = \widehat{\beta}_n^\top x + \widehat{\sigma}_n(x)\sqrt{\frac{\chi^2_{p,1-\alpha}}{n}}, \quad x \in \mathscr{X}.$$

On en déduit une bande de confiance de niveau asymptotique supérieur ou égal à $1-\alpha$ pour $\{p(x), x \in \mathscr{X}\}$ donnée par

$$\left[\widehat{l}_{p(x),n}, \widehat{u}_{p(x),n}\right] \tag{5.11}$$

où

$$\widehat{l}_{p(x),n} = \frac{e^{\widehat{l}_n(x)}}{1+e^{\widehat{l}_n(x)}} \quad \text{et} \quad \widehat{u}_{p(x),n} = \frac{e^{\widehat{u}_n(x)}}{1+e^{\widehat{u}_n(x)}}, \quad x \in \mathscr{X}.$$

On obtient le corollaire suivant :

Corollaire 5.3.3 *Sous les conditions (**A1**)-(**A4**) définies dans le chapitre 4, et sous les conditions des Théorèmes 4.4.7 et 5.2.1, les bandes de confiance définies par (5.11) ont une probabilité de couverture asymptotique supérieure ou égale à $1-\alpha$, i.e.*

$$\lim_{n\to\infty} \mathbf{P}\left(\widehat{l}_{p(x),n} \le p(x) \le \widehat{u}_{p(x),n}, \forall\, x \in \mathscr{X}\right) \ge 1-\alpha.$$

5.3.2 Méthode 2 : Égalité de Landau et Sheep (1970)

Nous présentons tout d'abord l'inégalité classique de (Landau & Sheep 1970) qui ont travaillé sur la majoration de suprémum de processus Gaussiens à trajectoires continues. Des résultats similaires ont été également établis par (Marcus & Sheep 1971).

Théorème 5.3.1 *Soit $(X_t)_{t\in T}$, (T un intervalle), un processus Gaussien centré à trajectoires continues. Alors on a*

$$\lim_{\lambda\to\infty} \lambda^{-2} \log \mathbf{P}\left(\sup_{t\in T} X_t > \lambda\right) = -\frac{1}{2\sigma_T^2}, \tag{5.12}$$

où

$$\sigma_T^2 = \sup_{t\in T} \mathbb{E}(X_t^2).$$

En utilisant le Théorème 5.2.1, il est donc possible d'appliquer l'égalité classique 5.12 de (Landau & Sheep 1970) pour obtenir l'égalité

$$\lim_{\lambda\to\infty}\frac{1}{\lambda^2}\log \mathbf{P}\Big(\sup_{x\in\mathscr{X}} W(x)>\lambda\Big) = -\Big[2\sup_{x\in\mathscr{X}}\rho(x,x)\Big]^{-1} = -\frac{1}{2}. \tag{5.13}$$

En utilisant l'inégalité suivante

$$\mathbf{P}\Big(\sup_{x\in\mathscr{X}}|W(x)|>\lambda\Big) \leq 2\mathbf{P}\Big(\sup_{x\in\mathscr{X}} W(x)>\lambda\Big). \tag{5.14}$$

et l'égalité (5.13), on obtient

$$\lim_{\lambda\to\infty}\frac{1}{\lambda^2}\log \mathbf{P}\Big(\sup_{x\in\mathscr{X}}|W(x)|>\lambda\Big) \leq -\frac{1}{2}. \tag{5.15}$$

Maintenant il suffit de trouver pour quelle valeur de λ on a l'équation (5.4). Nous avons $\lim_{\lambda\to\infty}\frac{1}{\lambda^2}\log(\alpha) \leq -\frac{1}{2}$ et cette inégalité reste vraie pour $\lambda_p = \sqrt{-2\log(\alpha)}$. Pour cette valeur λ_p on a l'égalité (5.4) et grâce à la convergence établie dans (5.7), on obtient quand n tend vers l'infini :

$$\mathbf{P}\Big(\sup_{x\in\mathscr{X}}\Big|\sqrt{n}\frac{\widehat{\beta}_n^\top x-\beta^\top x}{\widehat{\sigma}_n(x)}\Big| \leq \lambda_p\Big) \to 1-\alpha.$$

Ce qui implique :

$$\mathbf{P}\Big(\Big|\sqrt{n}\frac{\widehat{\beta}_n^\top x-\beta^\top x}{\widehat{\sigma}_n(x)}\Big| \leq \lambda_p, \forall\, x\in\mathscr{X}\Big) \to 1-\alpha.$$

On obtient de manière équivalente :

$$\mathbf{P}\Big(\widehat{l}_n(x) \leq \beta^\top x \leq \widehat{u}_n(x), \forall\, x\in\mathscr{X}\Big) \to 1-\alpha,$$

où

$$\widehat{l}_n(x) = \widehat{\beta}_n^\top x - \frac{\widehat{\sigma}_n(x)}{\sqrt{n}}\sqrt{-2\log(\alpha)} \quad \text{et} \quad \widehat{u}_n(x) = \widehat{\beta}_n^\top x + \frac{\widehat{\sigma}_n(x)}{\sqrt{n}}\sqrt{-2\log(\alpha)},$$

pour tout $x \in \mathscr{X}$.

On en déduit une bande de confiance pour $\{p(x),\ x \in \mathscr{X}\}$ donnée par

$$\left[\widehat{l}_{p(x),n}, \widehat{u}_{p(x),n}\right], \tag{5.16}$$

où

$$\widehat{l}_{p(x),n} = \frac{e^{\widehat{l}_n(x)}}{1+e^{\widehat{l}_n(x)}} \quad \text{et} \quad \widehat{u}_{p(x),n} = \frac{e^{\widehat{u}_n(x)}}{1+e^{\widehat{u}_n(x)}}, \quad x \in \mathscr{X}.$$

On obtient le corollaire suivant :

Corollaire 5.3.4 *Sous les conditions (**A1**)-(**A4**) définies dans le chapitre 4, et sous les conditions des théorèmes 4.4.7 et 5.2.1, les bandes de confiance définies par (5.16) avec le quantile choisi $\lambda_p = \sqrt{-2\log(\alpha)}$ ont une probabilité de couverture asymptotique égale à $1-\alpha$.*
i.e.

$$\lim_{n\to\infty} \mathbf{P}\left(\widehat{l}_{p(x),n} \leq p(x) \leq \widehat{u}_{p(x),n}, \forall\ x \in \mathscr{X}\right) = 1-\alpha.$$

5.3.3 Méthode 3 : Bootstrap - Monte Carlo

Dans cette partie, nous proposons une approche *bootstrap* (*Monte Carlo*) pour construire les bandes de confiance pour $\{p(x), x \in \mathscr{X}\}$ dans le modèle de régression logistique avec fraction immune. Nous utilisons d'abord des échantillons bootstrappés pour calculer la matrice de covariance empirique des estimateurs du modèle ZIB (5.1-5.2). Ensuite, un nombre important de simulations est utilisé pour approcher la distribution de $\sup_{x\in\mathscr{X}} |W_n(x)|$. Enfin, nous déterminons le quantile d'ordre α de ce nouvel échantillon pour construire les bandes de confiance. Ce type de technique est utilisé par (Efron & DiCiccio 1996), (Mandel & Betensky 2008) et (Li *et al.* 2010) pour la construction d'intervalles de confiance. (Claeskens & Van Keilegom 2003) ont utilisé une technique similaire pour la construction de bandes de confiance pour des fonctions de régression et leurs dérivées. (Neumann & Polzehl 1998) l'ont utilisé en régression non-paramétrique et (Zhang & Peng 2010) pour la construction de bandes de confiance dans des modèles de régression avec paramètres de régression non constants.

Cette approche est également utilisée en analyse statistique des durées de vie pour construire des bandes de confiance simultanées pour la fonction de survie ((Li & Datta 2001)) et pour la fonction de risque instantané ((Dudek *et al.* 2008)), tandis que (Cowling *et al.* 1996) ont utilisé la méthode de ré-échantillonnage pour la construction de bandes de confiance pour la fonction d'intensité d'un processus de Poisson.

Nous décrivons dans la suite l'algorithme utilisé pour la construction de nos bandes de confiance simultanées. Posons $U_n = \sup_{x\in\mathscr{X}} |W_n(x)|$ et $\zeta = \sqrt{n}(\widehat{\beta}_n - \beta)$. L'idée consiste à obtenir $M > 0$ réplications $U_n^{(1)}, ..., U_n^{(M)}$ et à estimer le quantile d'ordre $(1-\alpha)$ de la loi de U_n par le quantile empirique d'ordre $(1-\alpha)$ des $U_n^{(1)}, ..., U_n^{(M)}$.

Algorithme :

1. Générer B échantillons bootstraps de taille n, $\{(Y_1^{(b)}, \mathbb{X}_1^{(b)}, \mathbb{Z}_1^{(b)}), ..., (Y_n^{(b)}, \mathbb{X}_n^{(b)}, \mathbb{Z}_n^{(b)})\}$, $b = 1, ..., B$ à partir des observations.

2. Calculer $\widehat{\psi}_n^{(b)} = (\widehat{\beta}_n^{\top(b)}, \widehat{\theta}_n^{\top(b)})^\top, b = 1, ..., B$, à partir des modèles (4.1)-(4.2).

3. Calculer l'estimateur bootstrappé de $\widehat{\Sigma}_{\beta,n}$: $\widehat{\Sigma}_{\text{boot}} = \text{var}(\widehat{\beta}_n^{(1)}, ..., \widehat{\beta}_n^{(B)})$.

4. Générer M vecteurs aléatoires indépendants ($\zeta_m \in \mathbb{R}^p, m = 1, ..., M$) suivant une distribution normale multivariée de moyenne 0 et de matrice de covariance $\widehat{\Sigma}_{\text{boot}}$.

5. Évaluer

$$U_n^{(m)} = \max_{x_i \in \mathscr{X}} \left| \frac{\zeta_m^\top x_i}{\widehat{\sigma}_{\text{boot},n}(x_i)} \right|, \quad x_i = (1, x_{i2}, ..., x_{ip})^\top, m = 1, ..., M,$$

où

$$\widehat{\sigma}_{\text{boot},n}(x_i) = x^\top \widehat{\Sigma}_{\text{boot}} x,$$

6. Déterminer le quantile empirique d'ordre $(1-\alpha)$ de l'échantillon $\{U_n^{(1)}, ..., U_n^{(M)}\}$. On le note $\widehat{c}_{1-\alpha}$.

Comme la distribution empirique de $\{U_n^{(1)}, ..., U_n^{(M)}\}$ est une approximation de celle de $\sup_{x \in \mathscr{X}} |W_n(x)|$, d'après le Théorème 5.2.1 on peut construire une bande de confiance $[\widehat{l}_n(x), \widehat{u}_n(x)]$ pour $\{\beta^\top x,\ x \in \mathscr{X}\}$ comme suit :

$$[\widehat{\beta}_n^\top x - \frac{\widehat{\sigma}_n(x)}{\sqrt{n}} \widehat{c}_{1-\alpha}, \widehat{\beta}_n^\top x + \frac{\widehat{\sigma}_n(x)}{\sqrt{n}} \widehat{c}_{1-\alpha}], \quad x \in \mathscr{X}.$$

On en déduit une bande de confiance pour $\{p(x); x \in \mathscr{X}\}$ donnée par

$$\left[\widehat{l}_{p(x),n}, \widehat{u}_{p(x),n}\right] \tag{5.17}$$

où

$$\widehat{l}_{p(x),n} = \frac{e^{\widehat{l}_n(x)}}{1 + e^{\widehat{l}_n(x)}} \quad \text{et} \quad \widehat{u}_{p(x),n} = \frac{e^{\widehat{u}_n(x)}}{1 + e^{\widehat{u}_n(x)}}, \quad x \in \mathscr{X}.$$

On obtient le corollaire suivant :

Corollaire 5.3.5 *Sous les conditions (**A1**)-(**A4**) définies dans le chapitre 4, et sous les conditions des théorèmes 4.4.7 et 5.2.1, les bandes de confiance définies par (5.17) ont une probabilité de couverture asymptotique égale à $1-\alpha$, c'est-à-dire*

$$\lim_{n\to\infty} \mathbf{P}\Big(\widehat{l}_{p(x),n} \leq p(x) \leq \widehat{u}_{p(x),n}, \forall\ x \in \mathscr{X}\Big) = 1-\alpha.$$

5.4 Étude de simulation

Les objectifs visés dans cette partie d'étude de simulation sont de comprendre et de montrer les performances des bandes de confiance en taille d'échantillon finie. Précisément, nous évaluons l'influence de différents paramètres de simulation (valeur du niveau de confiance, taille d'échantillon, pourcentage d'immunes et proportion d'individus infectés parmi les susceptibles) sur la probabilité de couverture et la longueur moyenne des bandes. Nous décrivons tout d'abord la procédure de simulation.

5.4.1 Plan de simulation

Nous considérons les modèles suivants pour le statut infection :

$$\begin{cases} \log\left(\frac{\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}{1-\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}\right) = \beta_1 + \beta_2 X_{i2} + \beta_3 Z_{i2} & \text{if } S_i = 1 \\ \mathbb{P}(Y=1|\mathbf{X}_i,S_i) = 0 & \text{if } S_i = 0 \end{cases} \tag{5.18}$$

et le statut d'immunité :

$$\log\left(\frac{\mathbb{P}(S=1|\mathbf{Z}_i)}{1-\mathbb{P}(S=1|\mathbf{Z}_i)}\right) = \theta_1 + \theta_2 Z_{i2}, \tag{5.19}$$

où X_{i2} est une variable de loi normale de moyenne 0 et de variance 1 et Z_{i2} est une variable aléatoire de loi uniforme dans $(0,1]$. Notons que la variable $X_{i2}, 1 \leq i \leq n$

joue le rôle de la covariable continue V dans la condition A2 (voir chapitre 4). Nous pouvons également noter que les modèles d'infection et d'immunité partagent une covariable. Un échantillon i.i.d. de taille n du vecteur $\mathscr{O}$ est généré à partir de ce modèle, et pour chaque individu i, nous obtenons une réalisation $\mathscr{O}_i = (y_i, s_i, \mathbf{x}_i, \mathbf{z}_i)$, où s_i est considéré comme inconnu si $y_i = 0$. L'estimateur du maximum de vraisemblance $\widehat{\beta}_n$ de $\beta = (\beta_1, \beta_2, \beta_3)^\top$ est obtenu à partir de cette base de données incomplète en résolvant l'équation du score donnée dans la Section 4.2 du chapitre 3, en utilisant la fonction `optim` du logiciel `R` (version 2.14). Un estimateur du paramètre de nuisance $\theta = (\theta_1, \theta_2)^\top$ est également obtenu. Nous nous sommes particulièrement intéressés à la probabilité de survenue de l'infection $p(\mathbf{x}) = \mathbf{P}(Y = 1|\mathbf{X} = \mathbf{x}, S = 1)$. Pour chaque configuration des différents paramètres de simulation, nous calculons la probabilité de couverture empirique et la longueur moyenne des bandes de confiance, pour chacune des 3 méthodes proposées dans la section 5.3.

Le comportement sur des tailles d'échantillon finies des bandes de confiances est évalué pour plusieurs tailles d'échantillon ($n = 500, 1000, 1500$) et différents pourcentages d'individus immunes dans l'échantillon, à savoir 25%, 50%. Nous considérons également différentes valeurs de la proportion d'individus infectés parmi les susceptibles (30% and 70%) et différentes valeurs du niveau de confiance des bandes construites : 90%, 95% et 99%. Les proportions souhaitées d'individus immunes et infectés sont obtenues en choisissant des valeurs appropriées pour β et θ. Les valeurs suivantes sont considérées pour β :

- modèle $\mathscr{M}_1$: $\beta = (-1, 1.5, -.4)^\top$: approximativement 30% des susceptibles sont infectés.
- modèle $\mathscr{M}_2$: $\beta = (.5, -1, -1.1)^\top$: approximativement 70% des susceptibles sont infectés.

5.4.2 Résultats

Pour chaque configuration (niveau de confiance $\times$ taille d'échantillon $\times$ pourcentage d'individus immunes $\times$ pourcentage d'infectés parmi les individus susceptibles) de choix des paramètres, $N = 1000$ échantillons sont obtenus. Sur la base de ces 1000 réplications, nous obtenons $\widehat{p}_n^{(1)}(x), \ldots, \widehat{p}_n^{(1000)}(x)$, pour tout $x \in \mathscr{X}$. Pour chacune des trois méthodes proposées, nous calculons la probabilité de couverture empirique comme le pourcentage de bandes contenant exactement la vraie fonction réponse parmi les 1000 bandes obtenues. Nous avons également étudié la précision des bandes en regardant leurs largeurs. Des indicateurs de précision sont calculés en prenant :

1. pour chaque réplication la moyenne et la médiane des largeurs des bandes sur les n valeurs $x_1, \ldots, x_n$ de x générées pour les n individus de l'échantillon,
2. puis en calculant la moyenne des deux séries de n valeurs (moyenne et médiane) ainsi obtenues.

Les résultats sont résumés dans les tableaux 5.1 et 5.2. Dans ces tableaux, les méthodes 1, 2 et 3 correspondent respectivement à la méthode de Scheffé, à la méthode de Landau et Sheep et à la méthode Bootstrap. Les largeurs présentées dans ces tableaux sont les largeurs moyennes des bandes.

A partir des Tableaux 5.1 et 5.2, nous notons que la méthode 2 fournit des bandes de confiance ayant une probabilité de couverture plus faible que les méthodes 1 et 3. Nous constatons également que la probabilité de couverture diminue quand la taille d'échantillon augmente. On peut faire la même constatation pour la largeur moyenne des bandes de confiance : les bandes se rétrécissent lorsque la taille de l'échantillon augmente, et ce, indépendamment des trois méthodes. La méthode 3 (méthode bootstrap) est associée à des probabilités de couverture et des largeurs de bande plus élevées que les deux autres méthodes. Le caractère conservatif des bandes de confiances constaté sur la méthode 3 est très similaire à ce qui est observé dans la plupart des travaux portant sur les bandes et intervalles de confiance simultanés dans le modèle de régression linéaire et logistique (voir (Zhang & Peng 2010) et (Li *et al.* 2010)). En pratique, nous recommandons l'utilisation des méthodes 1 ou 2 pour des raisons de temps de calculs (moins longs que pour la méthode 3) et également pour des raisons de meilleures précisions de largeurs de bandes. Nous remarquons également que si l'on augmente la valeur du niveau de confiance, la largeur des bandes de confiance augmente.

Les résultats (probabilités de couverture et largeurs moyennes des bandes) sont stables lorsque le pourcentage d'immunes est de l'ordre de 25% et se dégradent lorsqu'il augmente. Quand le pourcentage d'immune est de 50%, les probabilités de couverture sont faibles, par contre les largeurs des bandes restent néanmoins stables. Ensuite, nous comparons ces résultats du modèle $\mathscr{M}_1$ à ceux obtenus par la méthode d'analyse "naïve" où :

- nous considérons tout individu i tel que $\{Y_i = 0\}$ comme susceptible mais non infecté (c'est-à-dire, nous ignorons une éventuelle immunité de cet individu),
- nous calculons les bandes de confiance pour $p(x)$ à partir des données ainsi obtenues.

Les résultats de cette analyse "naïve" pour le modèle $\mathscr{M}_1$ sont donnés dans le Tableau 5.3. Les probabilités de couverture sont proches de 0, indépendamment de la taille d'échantillon lorsque le pourcentage d'immunes est égal à 25%. Elles sont pratiquement toutes égales à 0 lorsque le pourcentage d'immunes est de 50%.

Ceci nous permet d'affirmer qu'il est important de tenir compte de l'immunité lorsqu'elle est présente dans la population.
Le modèle $\mathscr{M}_2$ fournit des résultats similaires. Ils sont présentés dans la section B.2 dans l'annexe B.

TAB. 5.1 – Modèle $\mathscr{M}_1$: $\beta = (-1, 1.5, -.4)$ avec 25% d'immunes.

		Méthode 1		Méthode 2		Méthode 3	
$1-\alpha$	n	Couverture[a]	Largeur	Couverture[a]	Largeur	Couverture[a]	Largeur
			Pourcentage d'immunes $= 25\%, \theta = (1.6, -1)$				
0.99	500	0.946	0.217^{+}	0.938	0.205^{+}	0.999	0.734^{+}
			0.082†		0.074†		0.627†
			0.174*		0.161*		0.830*
			0.338$^{\mp}$		0.322$^{\mp}$		0.898$^{\mp}$
	1000	0.931	0.134^{+}	0.922	0.128^{+}	0.994	0.593^{+}
			0.031†		0.028†		0.282†
			0.087*		0.082*		0.775*
			0.220$^{\mp}$		0.211$^{\mp}$		0.877$^{\mp}$
	1500	0.837	0.102^{+}	0.818	0.098^{+}	0.996	0.588^{+}
			0.018†		0.017†		0.332†
			0.062*		0.059*		0.736*
			0.172$^{\mp}$		0.165$^{\mp}$		0.827$^{\mp}$
0.95	500	0.854	0.169^{+}	0.841	0.163^{+}	0.989	0.667^{+}
			0.045†		0.042†		0.526†
			0.117*		0.110*		0.761*
			0.274$^{\mp}$		0.264$^{\mp}$		0.852$^{\mp}$
	1000	0.791	0.108^{+}	0.773	0.105^{+}	0.989	0.530^{+}
			0.019†		0.018†		0.187†
			0.063*		0.061*		0.698*
			0.181$^{\mp}$		0.175$^{\mp}$		0.829$^{\mp}$
	1500	0.836	0.081^{+}	0.827	0.078^{+}	0.988	0.518^{+}
			0.012†		0.012†		0.216†
			0.046*		0.044*		0.684*
			0.137$^{\mp}$		0.133$^{\mp}$		0.784$^{\mp}$
0.90	500	0.846	0.142^{+}	0.842	0.138^{+}	0.989	0.607^{+}
			0.035†		0.033†		0.457†
			0.093*		0.089*		0.694*
			0.230$^{\mp}$		0.224$^{\mp}$		0.793$^{\mp}$
	1000	0.788	0.092^{+}	0.776	0.090^{+}	0.990	0.485^{+}
			0.015†		0.015†		0.160†
			0.053*		0.051*		0.624*
			0.154$^{\mp}$		0.150$^{\mp}$		0.775$^{\mp}$
	1500	0.8	0.073^{+}	0.792	0.071^{+}	0.983	0.462^{+}
			0.010†		0.010†		0.127†
			0.039*		0.039*		0.624*
			0.123$^{\mp}$		0.121$^{\mp}$		0.741$^{\mp}$

Note : ^{a} : probabilité de couverture, $^{+}$: moyenne, † : 1er quartile, * : médiane, $^{\mp}$: 3ème quartile.

TAB. 5.2 – Modèle $\mathscr{M}_1$: $\beta = (-1, 1.5, -.4)$ avec 50% d'immunes.

		Méthode 1		Méthode 2		Méthode 3	
$1-\alpha$	n	Couverture^a	Largeur	Couverture^a	Largeur	Couverture^a	Largeur
		Pourcentage d'immunes $= 50\%, \theta = (-1, 2)$					
0.99	500	0.879	0.325^{+}	0.862	0.307^{+}	0.998	0.885^{+}
			$0.138^{\dagger}$		$0.124^{\dagger}$		$0.827^{\dagger}$
			$0.279^{\star}$		$0.258^{\star}$		$0.954^{\star}$
			$0.495^{\mp}$		$0.472^{\mp}$		$0.983^{\mp}$
	1000	0.847	0.180^{+}	0.836	0.171^{+}	0.994	0.596^{+}
			$0.053^{\dagger}$		$0.048^{\dagger}$		$0.325^{\dagger}$
			$0.131^{\star}$		$0.123^{\star}$		$0.727^{\star}$
			$0.292^{\mp}$		$0.279^{\mp}$		$0.856^{\mp}$
	1500	0.865	0.143^{+}	0.850	0.137^{+}	0.991	0.533^{+}
			$0.031^{\dagger}$		$0.029^{\dagger}$		$0.196^{\dagger}$
			$0.092^{\star}$		$0.087^{\star}$		$0.690^{\star}$
			$0.239^{\mp}$		$0.229^{\mp}$		$0.823^{\mp}$
0.95	500	0.797	0.232^{+}	0.779	0.222^{+}	0.997	0.797^{+}
			$0.082^{\dagger}$		$0.075^{\dagger}$		$0.684^{\dagger}$
			$0.178^{\star}$		$0.167^{\star}$		$0.908^{\star}$
			$0.365^{\mp}$		$0.351^{\mp}$		$0.966^{\mp}$
	1000	0.636	0.139^{+}	0.627	0.134^{+}	0.985	0.538^{+}
			$0.031^{\dagger}$		$0.029^{\dagger}$		$0.272^{\dagger}$
			$0.089^{\star}$		$0.085^{\star}$		$0.622^{\star}$
			$0.230^{\mp}$		$0.222^{\mp}$		$0.784^{\mp}$
	1500	0.731	0.106^{+}	0.720	0.103^{+}	0.986	0.448^{+}
			$0.020^{\dagger}$		$0.019^{\dagger}$		$0.135^{\dagger}$
			$0.064^{\star}$		$0.061^{\star}$		$0.547^{\star}$
			$0.177^{\mp}$		$0.171^{\mp}$		$0.722^{\mp}$
0.90	500	0.716	0.197^{+}	0.705	0.191^{+}	0.984	0.757^{+}
			$0.059^{\dagger}$		$0.056^{\dagger}$		$0.614^{\dagger}$
			$0.141^{\star}$		$0.135^{\star}$		$0.875^{\star}$
			$0.316^{\mp}$		$0.307^{\mp}$		$0.946^{\mp}$
	1000	0.744	0.121^{+}	0.726	0.118^{+}	0.977	0.469^{+}
			$0.025^{\dagger}$		$0.024^{\dagger}$		$0.207^{\dagger}$
			$0.075^{\star}$		$0.072^{\star}$		$0.547^{\star}$
			$0.199^{\mp}$		$0.194^{\mp}$		$0.718^{\mp}$
	1500	0.607	0.093^{+}	0.596	0.091^{+}	0.966	0.410^{+}
			$0.016^{\dagger}$		$0.015^{\dagger}$		$0.111^{\dagger}$
			$0.054^{\star}$		$0.052^{\star}$		$0.479^{\star}$
			$0.156^{\mp}$		$0.153^{\mp}$		$0.670^{\mp}$

Note : ^{a} : probabilité de couverture, $^+$: moyenne, † : 1er quartile, * : médiane, $^\mp$: 3ème quartile.

TAB. 5.3 – Analyse "naïve" : $\mathscr{M}_1 : \beta = (-1, 1.5, -.4)$ avec 25% et 50% d'immunes.

		Méthode 1		Méthode 2		Méthode 3	
$1-\alpha$	n	Couverture[a]	Largeur	Couverture[a]	Largeur	Couverture[a]	Largeur
				Pourcentage d'immunes $= 25\%, \theta = (1.6, -1)$			
0.99	500	0.301	0.138	0.253	0.132	0.320	0.139
	1000	0.001	0.092	0.001	0.089	0.001	0.093
	1500	0.358	0.089	0.313	0.085	0.355	0.088
0.95	500	0.004	0.116	0.003	0.112	0.007	0.118
	1000	0	0.079	0	0.077	0	0.080
	1500	0.012	0.063	0.009	0.061	0.015	0.063
0.90	500	0.008	0.096	0.005	0.094	0.010	0.098
	1000	0	0.065	0	0.064	0	0.066
	1500	0	0.052	0	0.051	0	0.053
				Pourcentage d'immunes $= 50\%, \theta = (-1, 2)$			
0.99	500	0.241	0.236	0.214	0.226	0.261	0.233
	1000	0	0.141	0	0.136	0	0.137
	1500	0	0.121	0	0.117	0	0.119
0.95	500	0	0.093	0	0.091	0	0.091
	1000	0	0.130	0	0.126	0	0.127
	1500	0	0.093	0	0.091	0	0.091
0.90	500	0	0.146	0	0.142	0	0.145
	1000	0	0.099	0	0.098		0 0.098
	1500	0	0.090	0	0.088	0	0.088

Note : a : probabilité de couverture.

Troisième partie

Applications

Chapitre 6

Étude épidémiologique de la dengue : projet DENFRAME

Sommaire

Le contenu du présent chapitre a fait l'objet d'une publication :

Dussart, P., Baril, L. et *al.*

Study of dengue cases and the members of their households : a familial cluster analysis in the multinational DENFRAME project

PLoS. Negl. Trop. Dis., vol. 6(1) : e1482, 2012. pages 80-106.

Abstract

Background **: Dengue has emerged as the most important vector-borne viral disease in tropical areas. Evaluations of the burden and severity of dengue disease have been hindered by the frequent lack of laboratory confirmation and strong selection bias toward more severe cases.**

Methodology **: A multinational, prospective clinical study was carried out in South-East Asia (SEA) and Latin America (LA), to ascertain the proportion of inapparent dengue infections in households of febrile dengue cases, and to compare clinical data and biological markers from subjects with various dengue disease patterns. Dengue infection was laboratory-confirmed during the acute phase, by virus isolation and detection of the genome. The four participating reference laboratories used standardized methods.**

Principal Findings **: Among 215 febrile dengue subjects, 114 in SEA and 101 in LA, 28 (13.0%) were diagnosed with severe dengue (from SEA only) using the WHO definition. Household investigations were carried out for 177 febrile subjects. Among household members at the time of the first home visit, 39 acute dengue infections were detected of which 29 were inapparent. A further 62 dengue cases were classified at early convalescent phase. Therefore, 101 dengue infections were found among**

the 408 household members. Adding these together with the 177 Dengue Index Cases, the overall proportion of dengue infections among the study participants was estimated at 47.5% (278/585 ; 95% CI 43.5-51.6). Lymphocyte counts and detection of the NS1 antigen differed significantly between inapparent and symptomatic dengue subjects ; among inapparent cases lymphocyte counts were normal and only 20% were positive for NS1 antigen. Primary dengue infection and a specific dengue virus serotype were not associated with symptomatic dengue infection.

Conclusion : **Household investigation demonstrated a high proportion of household members positive for dengue infection, including a number of inapparent cases, the frequency of which was higher in SEA than in LA**.

6.1 Introduction

Dengue is the most important mosquito-borne viral disease of humans. The disease is now endemic in more than 100 countries and threatens more than 2.5 billion people. It currently affects about 50 to 100 million people each year (Guzman & Kouri 2002). Dengue viruses (DENV) are enveloped, single-stranded positive-sense RNA viruses (family Flaviviridae, genus Flavivirus). There are four types of DENV : DENV-1, DENV-2, DENV-3 and DENV-4. Dengue virus infection induces life-long protective immunity to the homologous serotype, but confers only partial and transient protection against subsequent infections with any of the other three serotypes (WHO 2009). The disease spectrum ranges from inapparent infection or mild dengue fever (Endy *et al.* 2002), probably the most common form, to a potentially severe form of dengue characterized by plasma leakage and hemorrhage, known as severe dengue. Uncommonly, severe dengue may manifest as hepatitis, encephalopathy or rhabdomyolysis ((WHO 2009), (Kalayanarooj *et al.* 1997), (Hommel *et al.* 1998), (Murgue *et al.* 1999) and (Thomas *et al.* 2008)). About 500000 people are estimated to have severe dengue and about 25000, mostly children, die from it each year (Mackenzie *et al.* 2004). The underlying causes determining the outcome of DENV infection remain unknown. Although previous exposure, viral strain and human host genetic polymorphisms also influence the clinical outcome of DENV infection, we still know little about the complex interplay between host and pathogen in the pathogenesis of dengue ((Watts *et al.* 1999), (Rico-Hesse 2007), (Sakuntabhai *et al.* 2005) and (Silva *et al.* 2010)). Inapparent infections have largely been detected retrospectively through serology. The uses of genome detection or virus isolation have enabled detection of inapparent infections in cluster studies designed to detect natural infections in the community ((Beckett *et al.* 2005) and (Mammen *et al.* 2008)). The present study was designed to identify symptomatic and inapparent dengue-infected subjects in genetically-related individuals living in the same household, in line with the main aim of the DENFRAME project which is to explore the influence of human genetic variants and their functional roles in the pathogenesis of dengue disease in humans. We based the identification of dengueinfected subjects upon virological techniques, namely virus isolation and detection of the genome. We also took this opportunity to evaluate prospectively a commercial NS1 capture assay ((Young *et al.* 2000) and (Alcon *et al.* 2002)) that could potentially be implemented in laboratories for the diagnosis of acute dengue ((Dussart *et al.* 2006), (Dussart *et al.* 2008) and (Blacksell *et al.* 2008)).

6.2 Methods

6.2.1 Objectives

A multinational, prospective study was conducted in South-East Asia (Cambodia and Vietnam) and Latin America (Brazil and French Guiana). We used virological techniques to identify dengue patients diagnosed at the acute phase of disease among the patients presenting with dengue-like illness. We then performed a household investigation, comparing clinical data and biological markers from subjects with a broad range of dengue disease patterns, including inapparent dengue cases that are rarely captured in clinical studies. This clinical study's aims were : (i) to estimate the proportion of inapparent dengue infections among members of the households of laboratory-confirmed symptomatic dengue cases, (ii) to calculate the proportion of dengue-infected subjects at the time of the household investigation, and (iii) to compare clinical and biological data from inapparent and symptomatic dengue-infected subjects.

6.2.2 Study sites

Five institutions were involved in this study during the recruitment period : Instituto Evandro Chagas (IEC) in Belém (Parástate, Brazil), *Institut Pasteur du Cambodge* (IPC) in Phnom Penh (Cambodia), *Institut Pasteur de la Guyane* (IPG) in Cayenne (French Guiana) and *Institut Pasteur de Ho Chi Minh Ville* (IPHCM) in Vietnam were responsible for the recruitment of patients and virological analyses ; the *Institut Pasteur* (IP) in Paris (France) designed the study and was responsible for central monitoring and data analysis.
As shown in the two maps (Figure 1), volunteers were recruited at four clinical sites :

Vinh Thuan District Hospital (Vietnam), Kampong Cham Referral Hospital (Cambodia), the IPG in Cayenne (French Guiana) and public outpatient and emergency rooms managed by the Belém Health Secretariat in the districts of Guamá, Marco, Marambaia and Sacramenta, and the outpatient unit of the IEC (Brazil). The virology laboratories of the four institutions responsible for recruitment are all National Reference Centers (NRC) for Arboviruses (IEC is also a WHO collaborative center). These laboratories carried out virological, NS1 antigen (Platelia Dengue NS1 Antigen, Bio-Rad, Marnes La Coquette, France), and serological techniques.

6.2.3 Study design

We recruited subjects with acute dengue-like illness at the study sites. These subjects were identified by the treating physicians and were included if they satisfied the following criteria : (i) aged over 24 months ; (ii) oral temperature .38uC and onset of symptoms within the last 72 h ; and (iii) presenting with at least one clinical manifestation suggestive of dengue-like illness : severe headache, retro-orbital pain, myalgia, joint pain, rash or any bleeding symptom. Furthermore, for inclusion in the second step of the study, the subject had to come from a familial household containing more than two people during the seven days preceding illness. We first identified the dengue-infected subjects (referred to in this study as Dengue Index Cases or DIC) and non-dengueinfected subjects (defined as Non-Dengue Cases - NDC) on the basis of virological results from an acute sample (see below). We then recruited individuals from the households of the DIC. We thus constituted three groups of participants : 1) DIC, 2) household members (HHM), and 3) NDC not related to the DIC. For all groups (DIC, HHM and NDC), we applied the same exclusion criteria : women who were pregnant or breastfeeding, individuals with a focal source of infection (e.g. otitis media, pneumonia, meningitis), patients presenting with a known chronic illness, and patients with malaria. Moreover, to ensure the feasibility of this study, each study site was asked to target a convenient sample of 50 households and to recruit subjects from July 2006 to June 2007 in line with the approval granted by the Institutional Review Board and the timing of the dengue season at each site.

6.2.4 Clinical data and blood sample collection

Participants were examined during sequential visits, as shown in the study design charts (Figure 2). At each visit, data were collected with a standardized questionnaire. Severe dengue cases were classified, according to WHO recommendations on the basis of the clinical data. Biological data were also recorded at the sequential visits [2]. Blood samples were collected during the visits and were rapidly processed by the laboratories of each of the recruiting sites, for dengue diagnosis and biological testing. Blood sample volume was adapted for children weighing less than 20 kg.
Paired blood samples were collected for subjects presenting dengue-like illness to allow classification as DIC or NDC : during the acute phase (Visit 1) and during the convalescence phase (Visit 4 : 15 to 21 days after the onset of fever). Blood samples were taken from hospitalized DIC within 24 hours of defervescence (Visit 3). HHM were visited at home for blood collection within 24 to 72 hours of DIC identification (Home Visit 1). For practical and logistical reasons this delay of up to 72 hours was unavoidable. HHM were supplied with a monitoring diary card and a thermometer, to enable them to follow their temperature over a 7-day period. For HHM with a positive diagnosis of dengue or with an onset of fever during the seven days of monitoring, a second visit with blood collection for dengue diagnosis

was organized (Home Visit 2). Blood analyses included virological and serological dengue diagnosis, complete blood count, transaminases and bilirubin levels. Finally, the data were coded and entered into the computer via a secure website specifically developed with the PHP/MySQL system.

6.2.5 Classification of dengue cases on the basis of acute dengue diagnosis

All serum samples collected at Visit 1 or at Home Visit 1 or Home Visit 2 were tested : (i) for acute dengue diagnosis, defined as positive virus isolation on mosquito cells ((Gubler *et al.* 1984)) and/or positive viral RNA detection by reverse transcriptase-polymerase chain reaction (RT-PCR) (Lanciotti *et al.* 1992), and (ii) for the diagnosis of early convalescent dengue cases based on a standardized DENV IgM capture enzyme-linked immunosorbent assay (MAC-ELISA) (Nunes *et al.* 2011), and DENV IgG detection by indirect ELISA (in-house protocol developed by each NRC for Arboviruses). NS1 antigen detection was also performed.
Only subjects with febrile dengue infection diagnosis were classified as DIC. Subjects in the early stage of dengue convalescence at Visit 1 (i.e. positive NS1 antigen detection with concomitant DENV IgM detection, or isolated DENV IgM detection with no positive viral tests) were not classified as DIC ; we did not perform a household investigation for them. For the classification of dengue-infected HHM at Home Visit 1, we included both HHM with an acute (febrile or inapparent) dengue infection diagnosis and HHM with isolated DENV IgM detection, presumably related to an infection preceding that of the DIC (i.e in the early convalescence phase). During the 7-day period of home monitoring, several new febrile cases of dengue-infectedHHMwere also confirmed through Home Visit 2.

We were unable to use the DENV IgM/IgG ratio to distinguish between primary and secondary dengue infections, due to a lack of standardization of DENV IgG tests among laboratories (Shu & Huang 2004). We therefore established two groups of dengue-infected participants, based on the presence or absence of DENV IgG during the acute phase of the disease. In this study, we considered the presence of DENV IgG in the acute phase of the study to be suggestive of previous dengue infection. All sera were also checked for DENV IgM and IgG at Visit 4. Finally, if all these dengue tests were negative, participants were classified as NDC.

6.2.6 Ethics

The study was approved by the Institutional Review Board of the Institut Pasteur and by the ethics committees of each of the countries concerned. It was conducted in accordance with the Declaration of Helsinki, and the participants or the parents of minors participating in the study gave written informed consent before inclusion. The clinical protocol, the questionnaires, the standard operating procedures and informed consent forms were adapted and translated for each clinical site. All the documentation was accessible through a dedicated website with a specific login access (www.denframe.org). The centralized electronic database was based at the *Institut Pasteur* in Paris and registered with the *Commission Nationale de l'Informatique et des Libertés* (CNIL) in France.

6.2.7 Statistical methods

We present here the data from all four study sites in Latin America and South-East Asia. DIC are described according to region, disease severity, DENV type, age group and IgG status. We estimated the proportion of inapparent dengue infections among HHM, and we calculated the proportions of dengueinfected subjects among household subjects, in total and according to the IgG status at the time of household investigation. We compared clinical data and biological markers between inapparent dengue-infected subjects, symptomatic dengue-infected subjects, and non-dengue-infected participants at the time of the household investigation. We created binary variables to evaluate the potential effect of DENV infection on biological markers (hematocrit, platelets, neutrophils, lymphocytes, monocytes, ASAT, ALAT, bilirubin). For lymphocytes and neutrophils, we used a threshold of 26109/l. We used chi-squared or Fisher's exact tests to compare categorical variables between symptomatic cases, inapparent dengue-infected cases and non-dengue-infected subjects among HHM. Univariate and multivariable logistic regression models were used to assess the effect of covariates on the odds ratios (OR) of symptomatic dengue-infected cases, inapparent dengue-infected cases, and non-dengue-infected subjects among HHM. For the multivariable logistic regression models including data from household members, we used two-stage hierarchical regression models taking into account the family household structure (Greenland 2000). Potential confounders with a P value of less than 0.20 in univariate analysis were retained for the final multivariable analyses. STATA version 10.0 (Stata Corp., College Station, TX, USA) and a significance level of 5% were used for all statistical analyses.

6.3 Results

Flowcharts for the recruitment of participants at each step are shown in Figure 6.3.

– Step 1 : identification of dengue index cases (DIC)

We screened 473 febrile subjects for dengue infection. Thirty (6.3%) had at least one criterion for non inclusion in the study at presentation ; the remaining 443 (93.7%) were included in the study. We identified 215 (48.5%) of these 443 subjects as DIC, 21 (4.7%) as dengue convalescent cases, 187 (42.2%) as NDC, and 20 (4.5%) could not be classified because some biological markers were lacking. Recruitment levels during the study period were very low in French Guiana (9 DIC and 24 NDC), whereas there had been a large number of dengue cases during the rainy season of the previous year [25]. For the 215 subjects classified as DIC, 149 (69.3%) were positive by genome detection and viral isolation, 43 (20.0%) were positive by genome detection only, 15 (7.0%) were positive by viral isolation only, and a very few subjects (n= 8, 3.7%) were ultimately classified as DIC by the virologists, based on positive NS1 detection, clinical data and serological results (negative IgM at Visit 1 followed by seroconversion IgM at convalescent phase).

The proportions of subjects classified as either NDC or DIC differed between Latin America and South-East Asia : 69.5% (130/187) of the total NDC in the study, and 47.0% (101/215) of the DIC, were recruited in Latin America whereas 30.5% (57/187) of the NDC and 53.0% (114/215) of the DIC were recruited in South-East Asia (P,1024) (Figure 3A). In other words, in Latin America, in two thirds of subjects presenting with dengue-like illness, the cause was not related to dengue infection. Given the inclusion criteria, the dengue-like illness symptoms were not different between NDC and DIC (data not shown). However, all biological variables, including counts of platelets, lymphocytes and neutrophils, were significantly lower, whereas hematocrit and liver enzyme levels were higher in the DIC group than in the NDC group (data not shown).

Table 1 shows the distribution of DIC by region and according to IgG status at Visit 1 as a function of DENV type and age group. The proportions of severe dengue and dengue fever cases with DENV IgG (suggestive of previous DENV infection) and without DENV IgG in the acute phase were similar (Table 1) : 15 (55.6%) severe dengue cases tested negative for DENV IgG and 12 (44.4tested positive for DENV IgG, versus 49 (31.8%) and 105 (68.2%) of the subjects with non severe disease, respectively (P =0.017). DENV-1, -2 and -3 were found with similar frequencies in South- East Asia, whereas DENV-3 predominated in Latin America. Fifteen of the severe dengue cases reported in South-East Asia were infected with DENV-2 (53.6% ; 15/28). Interestingly, seven severe dengue cases positive for DENV-2 virus and negative for DENV IgG in the acute phase but with subsequent DENV IgM and IgG seroconversion were identified. This serological pattern suggests that these patients had primary DENV infection. Two DIC in Vietnam were reported with co-detection of multiple DENV strains by RT-PCR : DENV-2/DENV-1 and DENV-4/DENV-2 respectively ; the viral cultures were negative for both subjects. Only the first virus detected was considered for further statistical analysis (DENV-2 and DENV-4, respectively).

According to the WHO criteria, twenty-eight (13.0%) subjects were classified as severe dengue (based on severe plasma leakage and/or severe hemorrhages and/or severe organ impairment). All these cases were from clinical sites in South-East Asia (25 in Vietnam and 3 in Cambodia, as presented in Table S1). At visit 1, presentation with the following combination of features was significantly associated with the occurrence of severe dengue in this population : being male, over the age of seven years, with no retro-orbital pain but with bleeding, low monocyte count, normal liver enzyme levels and DENV-2 type infection.

For 163 (75.8%) DIC, data were available for all the biological markers at visits 1 and 4 (Figure 6.3). All these markers had returned to normal levels by visit 4, and all participants, including the 28 severe dengue cases displayed clinical recovery from dengue disease (data not shown).

- Step 2 : identification of household members (HHM)

  Agreement for household investigations was obtained from 177 (82.3%) DIC, corresponding to a total of 651 household members. We compared the distribution of the covariates (as listed in Table S1) between the 38 DIC with no familial investigation and the 177 DIC who underwent familial investigation ; no significant differences were found in the distribution of the covariates between these two groups (data not shown). All 28 patients with severe dengue infection underwent household investigation. In total, 141 (21.7%) of the 651 household members refused to participate in the study. We therefore screened 510 participants, 497 (97.5%) of whom were eligible for the study. All but one of these 497 household members were genetically related to the DIC. Eightyfour were not classifiable due to the lack of some biological results. Full assessment of DENV infection was carried out according to the study protocol for the remaining 413 of these subjects (Figure 6.3) during Home Visit 1.

  At the time of the household investigation (Home Visit 1), 39 subjects were identified as being in the acute phase of dengue infection : 29 (74.4%) cases were inapparent and 10 (25.6had symptomatic dengue infection. An additional 62 subjects were classified as being in the early phase of convalescence from dengue infection. The remaining 312 subjects were considered as non-dengue-infected at the time of Home Visit 1 (Figure 3B) ; however, five of them developed some clinical symptoms of dengue fever and were laboratory-confirmed as having acute dengue infection during the 7-day home monitoring. We excluded them (n=5) from the remaining analysis (n=312 subjects with 7-day home monitoring) that thus included 307 subjects (Figure 6.3). It should be noted that a second home visit and blood sampling was not possible, for ethical and logistical reasons, forHHM without any clinical symptoms after the 7-day home monitoring. Hence, among the 307 remaining subjects, some may have had an inapparent dengue infection after Home Visit 1. Therefore, we considered that at least 101 (39 acute or 62 early convalescent) dengue infections were found amongst 408 HHM (24.8% ; 95% confidence interval (CI) : 20.6-28.9) at the time of Home Visit 1 (Figure 6.3). Thus, adding together the 177 DIC and the 101 DENV-infected HHM, the overall proportion for dengue among the study participants was estimated at 47.5% (278/585 ; 95% CI : 43.5-51.6) (Figure 6.3). We have also estimated these proportions according to the IgG status (Table 6.2) at the time of Home Visit 1 (excluding the 5 subjects with known symptomatic infection, 3 were IgG positive and 2 were IgG negative). Among the 585 subjects, 6 had missing IgG data. Among 425 subjects with positive IgG, the estimated proportion of dengueinfected subjects was 43.8% (186/425 ; 95% CI : 39.0-48.5) and, among the 154 with negative IgG, this estimated proportion was 57.1% (88/154 ; 95% CI : 49.3-65.0).

In 101 (57.1%) households, there was only one dengue-infected case. For the 76 (42.9%) households with at least two dengueinfected cases, DENV type had been determined for all subjects in 29 households. Nine (31.0%) households were found to have two different DENV types circulating during the same time period : DENV-1 & DENV-3 (n=2 in Brazil, n=4 in Cambodia), DENV 1 & DENV-2 (n=1 in Vietnam), and DENV-2 & DENV-3 (n=2 in Vietnam).

Hematologic and hepatic biological markers observed among non-dengue-infected cases (n=307), inapparent dengue-infected cases (n=29), and symptomatic dengue-infected subjects (n= 192) are described in Table S2. Tables 6.3 & 6.4 show comparisons between non-dengue-infected and inapparent dengue-infected cases, and symptomatic and inapparent dengue-infected subjects, respectively, among the household subjects. Table S3 presents the main characteristics of subjects with acute dengue infection compared to non-dengue-infected subjects among the household subjects. In the comparisons between non-dengue-infected and inapparent dengue-infected subjects, taking into account potential confounders, only neutrophil and monocyte levels differed significantly whereas presence of IgG at Visit 1 was almost significant with the non-dengue-infected group. The comparison between symptomatic and inapparent dengue-infected subjects (Table 6.4) showed significant difference between groups for lymphocyte counts and positive NS1 antigen detection. In this analysis, no significant difference was found for DENV types identified or IgG detection during the acute phase.

6.4 Discussion

Several previous epidemiological studies have focused on school-based surveillance aiming at improving dengue-vector control measures ((Endy *et al.* 2002) and (Mammen *et al.* 2008)), studying the dynamics of patterns of dengue transmission

((Teixeira *et al.* 2002), (Morrison *et al.* 2010) and (Endy *et al.* 2011)) or describing a model that takes into account the role of human movement in the transmission dynamics of vector-borne pathogens (Stoddard *et al.* 2009). Earlier cluster investigation methods were designed as an alternative approach to the commonly used prospective cohort study method for investigating the natural history of dengue virus infection in South-East Asia and Latin America ((Beckett *et al.* 2005) and (Reyes *et al.* 2010)). Although different study designs have demonstrated the feasibility of identification of inapparent dengue cases, it remains difficult to recruit these subjects. We designed our study to include family household investigation in order to identify a group of inapparent dengue-infected subjects and to compare them with symptomatic dengue-infected and non-dengue-infected subjects living in the same family household. The study design was based on family household recruitment specifically in order to collect data and biological samples, and to study secondarily the host susceptibility to dengue infection and disease. Unlike studies based on cohorts from hospital referrals, this multi-country study captured dengue cases ranging from inapparent infections, through mild disease to severe dengue fever, using definitions of clinical cases and diagnostic methodology standardized across the four sites. The period of inclusion, from July 2006 to June 2007, spanned the dengue season at each site, although incidence of dengue was low that year in French Guiana.

The main objective of this study was to identify dengue infections and particularly inapparent infections among dengue patients' household family members in South-East Asia and Latin America. Based on our data, we estimated the proportion to be about 45Most of the dengue cases studied had symptomatic infections, covering the spectrum of disease from dengue fever to severe dengue cases. We also identified inapparent infections in the population. We observed dengue-infected subjects classified as DIC and some of their HHM without acute dengue infection but with a positive IgM detection, suggesting an early convalescent phase after dengue infection with no clinical symptoms. In this study we identified 29 inapparent dengue infections but we believe this number underestimates the proportion of inapparent dengue cases because we were not able to take blood samples from nonsymptomatic subjects at Home Visit 2.

We postulated that dengue is transmitted to members of the DIC's family household during the period of the index subject's infection, and thus designed our study to detect inapparent dengue infections with a home visit organized shortly after identification of DIC. Obviously, we cannot confirm whether the index subject's DIC was always the source of infection in other family members, but we can postulate that a non-hospitalized DIC who remains at home during acute illness represents a potential source of DENV transmission to Aedes. According to our study design, clustering of cases within a household could be the result of a single or very few infected mosquitoes biting different household members during a short period of time, perhaps within a single gonotrophic cycle as previously suggested ((Mammen *et al.* 2008) and (De Benedictis *et al.* 2003)). This is also consistent with a previous observation that over periods from 1 to 3 days, dengue cases were clustered within short distances, i.e., within a household (Morrison *et al.* 1998). No mosquito captures were, however, conducted in our study to identify DENV-positive Aedes mosquitoes. DENV sequencing would help resolve the extent of localized transmission.

We characterized subjects with acute dengue infection using virus isolation and detection of the genome. We also used NS1 antigen detection, a more recently recognised diagnostic tool. As for many tropical infectious diseases, there is an urgent need for validated diagnostic tools for dengue. In parallel with the virological techniques, we evaluated detection of the NS1 antigen with the Platelia Dengue NS1 Ag test. In this study, this test was found to have good sensitivity (83.6% ; 95% CI : 78.5-88.6) and specificity (98.9% ; 95% CI : 96.6-99.9) in both Asia and Latin America, as reported in previous studies ((Dussart *et al.* 2006), (Chuansumrit *et al.* 2008) and (Lima *et al.* 2010)). A recent multi-country study observed unequal sensitivity between geographical regions that remains unexplained, suggesting further assessments are needed ((Guzman *et al.* 2010)). The use of viral detection antigen is particularly useful during the first five days of illness with NS1 assays that are significantly more sensitive for primary than secondary dengue ((Dussart *et al.* 2008), (Lima *et al.* 2010) and (Tricou *et al.* 2010)). However, NS1 antigen could be detected in only 20% of inapparent DENV-infection. This finding suggests that NS1 antigen may have a role in dengue disease pathogenesis and also indicates that this test cannot be relied upon for detection of inapparent dengue infection.

By comparing HHM not infected with dengue with those presenting with inapparent dengue infection, we showed that neutrophil and monocyte counts were early indirect biological markers of dengue infection, whereas platelet counts and the frequency of IgG detection at the first visit did not differ between the two groups (Table 3). A comparison of inapparent dengueinfected HHM with symptomatic dengue-infected subjects showed that lymphocyte counts and detection of the NS1 antigen differed significantly between these two groups (Table 4). Moreover, the NS1 antigen was detected during the acute phase in most of the dengue cases tested, and the sensitivity of this test was even higher in severe dengue cases (26/28, Table S1), possibly reflecting higher viral loads. These findings may indirectly reflect the progression of the immune response to DENV, leading in some cases to severe acute lymphopenia and a lack of virological control, with high rates of NS1 antigen circulation in the blood that may be correlated with high-level or prolonged viremia ((Thomas *et al.* 2008) and (Tricou *et al.* 2010)). Severe dengue cases were also more likely to be male, to have lower monocyte counts or normal liver enzyme levels, and to be infected with DENV-2, although quantitative RT-PCR did no permit study of the magnitude of the viremia. We showed that half of the severe dengue cases had not previously been infected with DENV, as confirmed by the occurrence of DENV IgG seroconversion during convalescent phase (Thomas *et al.* 2008). In all dengue-infected subjects, including inapparent, we observed a decrease in neutrophil and monocyte counts. On one hand, it may suggest a direct effect of dengue illness on hematopoiesis, although such an effect is in conflict with data reported elsewhere in the literature (Balsitis *et al.* 2009). On the other hand, DENV is detected in peripheral monocytes during acute disease, and the infection of monocytes leads to cytokine production, suggesting that virus-monocyte interactions are relevant to pathogenesis ((Halstead & O'Rourke 1977), (Hase *et al.* 1989) and (Neves-Souza *et al.* 2005)). Moreover, DENV can induce apoptosis in monocytes, and this may lead to decreases in the number of these cells in severe dengue cases (Torrentes-Carvalho *et al.* 2009). In this study we only observed severe dengue cases in South- East Asia. Disease severity and pathogenesis remain largely unexplained and certainly related to complex interactions of several factors, including virus strain, immune response to previous dengue infection and host genetic background. The introduction of the Asian 1 DENV-2 genotype into the Americas in the 1980s led to the emergence of severe dengue cases on this continent. Following this introduction a new genotype emerged, named Asian/American DENV-2 genotype ((Twiddy *et al.* 2005), (Oliveira *et al.* 2010) and (Vu *et al.* 2010)). During the study period, this Asian/American genotype was circulating in French Guiana (Philippe Dussart, personal data) and probably in the north of Brazil, however DENV-2 did not cause an outbreak and we did not report any severe dengue case among Brazilian subjects.

Two constraints of the study design deserve mention. All methods (biological markers, virological testing, NS1 antigen detection and IgM serology) were standardized across the four reference laboratories, with the exception of the IgG ELISA. As a consequence, we were unable to calculate the IgM/IgG ratio ((Innis *et al.* 1989) and (Shu *et al.* 2003)). However, as the intention was to include dengue cases during the acute phase of infection, this ratio was not a crucial endpoint for the study. Another constraint of this study was that we did not include infants and children below 24 months of age in the DENFRAME project. However, several previous reports already provide insight into the epidemiology of dengue in this specific population ((Hammond *et al.* 2005), (Pengsaa *et al.* 2006), (Chau *et al.* 2009) and (Capeding *et al.* 2010)).

These findings confirm the complexity of dengue disease in humans and the need to strengthen multidisciplinary research efforts to improve our understanding not only of virus transmission but also host responses to DENV in various human populations. It will therefore be interesting, based on clinical data and biological samples collected in this study, to further evaluate the host susceptibility to dengue infection and disease using family-based association analyses. Moreover, we think that technological transfer of standardized diagnostic methods in laboratories based in tropical countries is essential if we are to estimate disease burden and to optimize vector control interventions. Together with improvements in clinical care for dengue patients and better understanding of dengue pathogenesis, the development of a preventive vaccine and antiviral drugs would complete the arsenal of weapons for combating dengue worldwide.

6.5 Appendix

TAB. 6.1 – Characteristics of dengue index cases (DIC, n=215).

	Acute serum samples (n=215)											
	Latin America (n=101)						South-East Asia (n=114)*					
	Negative IgG (n=14)			Positive IgG (n=87)			Negative IgG (n=14)			Positive IgG (n=87)		
	SD	DF	NC	SD	DF	NC	SD	DF	NC	SD	DF	NC
Dengue type												
DENV-1	-	3(50.0)	-	-	8(11.4)	3(17.6)	4(26.7)	20(46.5)	-	2(16.7)	14(40.0)	-
DENV-2	-	-	8(100.0)	-	13(18.7)	1(5.9)	7(46.7)	12(27.9)	-	7(58.3)	6(17.2)	-
DENV-3	-	3(50.0)	-	-	47(67.1)	13(76.5)	3(20.0)	9(21.0)	1(50.0)	1(8.3)	11(31.4)	2(50.0)
DENV-4	-	-	-	-	-	-	-	1(2.3)	-	-	2(5.7)	1(25.0)
MD	-	-	-	-	2(2.8)	-	1(6.6)	1(2.3)	1(50.0)	2(16.7)	2(5.7)	1(25.0)
Age group (years)												
$[2-7]$	-	-	-	-	3(4.3)	2(11.8)	2(13.3)	21(48.9)	2(100.0)	2(16.7)	13(37.1)	-
$]7-10]$	-	1(16.7)	1(12.5)	-	-	1(5.9)	3(20.0)	9(20.9)	-	4(33.3)	6(17.1)	1(25.0)
> 10	-	5(83.3)	7(87.5)	-	67(95.7)	13(76.4)	10(66.7)	13(30.2)	-	6(50.0)	16(45.7)	3(75.0)
MD	-	-	-	-	-	1(5.9)	-	-	-	-	-	-

Note : SD : severe dengue. DF : dengue fever. NC : non classifiable. MD : missing data. (.) : percentage. * : For 3 subjects infected by DENV-2, data related to IgG status were missing : 2 dengue fever cases and 1 severe dengue case. Distribution of DIC is provided by region in relation to the presence of WHO criteria for severe dengue and IgG status during the acute phase.

TAB. 6.2 – Distribution of the participants in the clinical study (n=590).

	Brazil	French Guiana	Cambodia	Vietnam	Total
	n=134(%)	n=28(%)	n=180(%)	n=248(%)	n=590(%)
	[IgG+/IgG2]	[IgG+/IgG2]	[IgG+/IgG2]	[IgG+/IgG2]	[IgG+/IgG2]
Non DENV-infected subjects	47(15.4)	9(3.0)	98(32.1)	151(49.5)	305 (51.7)
	[44/3]	[3/6]	[95/3]	[97/54]	[239/66]
Missing IgG data	1	-	-	1	2(0.3)
Early convalescent phase or convalescent phase (HHM only)	4(6.5)	3(4.9)	22(36.1)	32(52.5)	61(10.3)
	[4/0]	[2/1]	[22/0]	[25/7]	[53/8]
Missing IgG data	-	-	-	1	1(0.2)
DENV-infected at the acute phase (DIC+HHM)	82(37.6)	16(7.4)	60(27.5)	60(27.5)	218(37.0)
Symptomatic	[69/6]	[3/10]	[30/19]	[16/36]	[118/71]
Missing IgG data	-	-	-	3	3(0.5)
Inapparent dengue infection	[6/1]	[1/2]	[8/3]	[3/5]	[18/11]

<u>Note</u> : All participants were identified at Visit 1 for Dengue Index Cases (DIC) and at Home Visit 1 for dengue-infected household members (HHM). Their distribution is presented by country, according to DENV-infected status and IgG status. (.) : percentage.

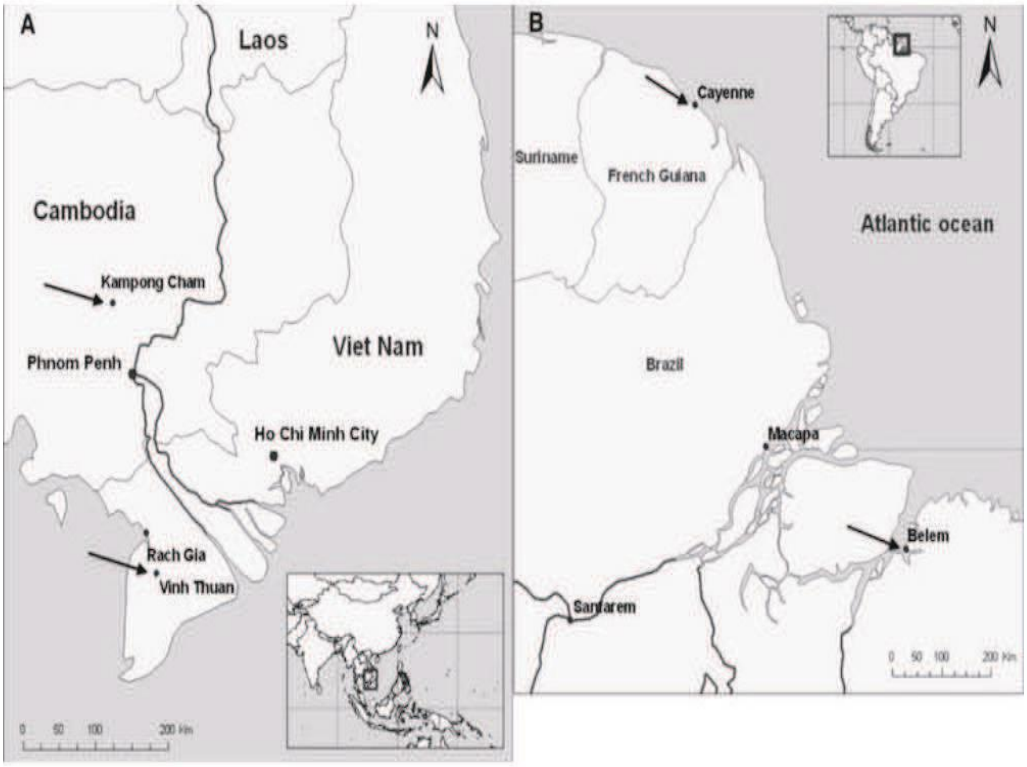

FIG. 6.1 – Clinical sites.

TAB. 6.3 – Main characteristics of subjects with inapparent dengue infections compared to non-dengue-infected subjects among Household members.

	NDI	IDI	Crude OR	95% CI	P*	Adj. OR	95% CI	P
Sex								
Male	135(44.0)	16(55.2)	1					
Female	172(56.0)	13(44.8)	0.64	[0.3 − 1.4]	0.25			
Age (years)								
[2 − 7]	16(5.2)	5(17.2)	1			1		
]7 − 10]	17(5.5)	2(6.9)	0.38	[0.1 − 2.2]	0.28	0.79	[0.1 − 6.5]	0.83
> 10	274(89.3)	22(75.9)	0.26	[0.1 − 0.7]	0.015	0.41	[0.1 − 1.8]	0.25
Weight-based Z-score								
$[-1, 1]$	89(29.0)	6(20.7)	1					
< -1	195(63.5)	21(72.4)	1.6	[0.6 − 4.1]	0.33			
> 1	23(7.5)	2(6.9)	1.3	[0.2 − 6.8]	0.76			
Hematocrit (%)								
≤ 36	93(30.3)	7(24.1)	1					
> 36	212(69.1)	22(75.9)	1.38	[0.6 − 3.3]	0.48			
Missing data	2(0.6)	-						
Platelets ($\times 10^9/L$)								
> 100	296(96.4)	26(89.7)	1			1		
≤ 100	10(3.3)	3(10.3)	3.42	[0.9 − 13.2]	0.075	1.71	[0.2 − 12.3]	0.6
Missing data	1(0.3)	-						
Neutrophils ($\times 10^9/L$)								
> 2	288(93.8)	18(62.1)	1			1		
≤ 2	18(5.9)	11(37.9)	9.8	[4 − 23.8]	< 0.0001	7.75	[2.5 − 24]	< 0.0001
Missing data	1(0.3)	-						
Lymphocytes ($\times 10^9/L$)								
> 2	243(79.2)	15(51.7)	1			1		
≤ 2	63(20.5)	14(48.3)	3.6	[1.6 − 7.8]	0.001	2.08	[0.7 − 5.6]	0.15
Missing data	1(0.3)	-						
Monocytes ($\times 10^9/L$)								
> 0.2	298(97.1)	23(79.3)	1			1		
≤ 0.2	8(2.6)	6(20.7)	9.72	[3.1 − 30]	< 0.0001	9.1	[1.8 − 44]	0.006
Missing data	1(0.3)	-						
ASAT^a **(UI/L)**								
≤ 30	225(73.3)	17(58.6)	1			1		
> 30	81(26.4)	11(37.9)	1.8	[0.8 − 4]	0.15	1.96	[0.7 − 5.2]	0.17
Missing data	1(0.3)	1(3.5)						
ALAT^b **(UI/L)**								
≤ 35	261(85.0)	22(75.9)	1					
> 35	45(14.7)	6(20.7)	1.58	[0.6 − 4.1]	0.35			
Missing data	1(0.3)	1(3.4)						
Bilirubin (μ**mol**)								
≤ 17	262(85.3)	24(82.8)	1					
> 17	42(13.7)	3(10.3)	0.78	[0.2 − 2.7]	0.69			
Missing data	3(1.0)	2(6.9)						
IgG at Visit 1								
Negative	66(21.5)	11(37.9)	1			1		
Positive	239(77.8)	18(62.1)	0.45	[0.2 − 1.0]	0.051	0.37	[0.1 − 1.04]	0.06
Missing data	2(0.7)	-						

Note : NDI : non dengue-infected. IDI : inapparent dengue infection. (.) : percentage. * : Potential confounders with a P value of less than 0.20 in univariate analysis were retained for the final multivariable analyses. In this table : age, platelets, neutrophils, lymphocytes, ASAT and IgG at Visit 1. ^{a}ASAT : Aspartate amino transferase. ^{b}ALAT : Alanine amino transferase. Univariate and multivariable logistic regression were used for analyses.

TAB. 6.4 – Main characteristics of subjects with inapparent dengue infections compared to symptomatic dengue-infected subjects.

	SDI	IDI	Crude OR	95% CI	P*	Adj. OR	95% CI	P
Sex								
Male	103(53.6)	16(55.2)	1					
Female	89(46.4)	13(44.8)	0.94	[0.4 − 2.1]	0.88			
Age (years)								
[2 − 7]	38(19.8)	5(17.2)	1			1		
]7 − 10]	27(14.1)	2(6.9)	0.56	[0.1 − 3.1]	0.51			
> 10	127(66.1)	22(75.9)	1.32	[0.5 − 3.7]	0.6			
Weight-based Z-score								
[−1, 1]	75(39.1)	6(20.7)	1			1		
< -1	102(53.1	21(72.4)	2.57	[0.9 − 6.7]	0.052	2.54	[0.6 − 10.4]	0.20
> 1	15(7.8)	2(6.9)	1.66	[0.3 − 9.1]	0.55	4.11	[0.4 − 43]	0.24
Hematocrit (%)								
≤ 36	38(19.8)	7(24.1)	1					
> 36	154(80.2)	22(75.9)	0.77	[0.3 − 1.9]	0.59			
Platelets ($\times 10^9/L$)								
> 100	126(65.6)	26(89.7)	1			1		
≤ 100	66(34.4)	3(10.3)	0.22	[0.1 − 0.7]	0.016	0.23	[0.4 − 1.4]	0.11
Neutrophils ($\times 10^9/L$)								
> 2	76(39.6)	18(62.1)	1			1		
≤ 2	116(60.4)	11(37.9)	0.4	[0.2 − 0.9]	0.026	0.5	[0.15 − 1.6]	0.25
Lymphocytes ($\times 10^9/L$)								
> 2	16(8.3)	15(51.7)	1			1		
≤ 2	176(91.7)	14(48.3)	0.08	[0.03 − 0.2]	< 0.0001	0.09	[[0.02 − 0.4]	0.001
Monocytes ($\times 10^9/L$)								
> 0.2	114(59.4)	23(79.3)	1			1		
≤ 0.2	78(40.6)	6(20.7)	0.38	[0.1 − 0.9]	0.045	0.65	[[0.16 − 2.7]	0.56
ASAT[a] (UI/L)								
≤ 30	75(39.1)	17(58.6)	1			1		
> 30	117 (60.9)	11(37.9)	0.4	[0.2 − 0.9	0.034	0.4	[0.1 − 1.5]	0.17
Missing data	-	1(3.5)						
ALAT[b] (UI/L)								
≤ 35	112(58.3)	22(75.9)	1			1		
> 35	80(41.7)	6(20.7)	0.38	[0.15 − 0.9]	0.046	0.52	[0.14 − 1.9]	0.33
Missing data	-	1(3.4)						
Bilirubin (μ**mol)**								
≤ 17	175(91.1)	24(82.8)	1					
> 17	14(7.3)	3(10.3)	1.56	[0.4 − 5.8]	0.51			
Missing data	3(1.6)	2(6.9)						
DENV type								
DENV-1	50(26.0)	5(17.2)	1					
DENV-2	50(26.0)	7(24.2)	1.4	[0.4 − 4.7]	0.59			
DENV-3	79(41.2)	13(44.8)	1.64	[0.5 − 4.9]	0.37			
DENV-4	3(1.6)	-						
Missing data	10(5.2)	4(13.8)						
IgG at Visit 1								
Negative	71(37.0)	11(37.9)	1					
Positive	118(61.4)	18(62.1)	0.98	[0.4 − 2.2]	0.97			
Missing data	3(1.6)	-						
NS1 antigen								
Negative	21(10.9)	23(79.3)	1			1		
Positive	171(89.1)	6(20.7)	0.03	[0.01 − 0.1]	< 0.0001	0.05	[0.01 − 0.2]	< 0.0001

Note : SDI : symptomatic dengue-infected. IDI : inapparent dengue infection. In this table : weight-based Z-score, platelets, neutrophils, lymphocytes, monocytes, ASAT, ALAT and NS1 antigen. [a]ASAT : Aspartate amino transferase. [b]ALAT : Alanine amino transferase.

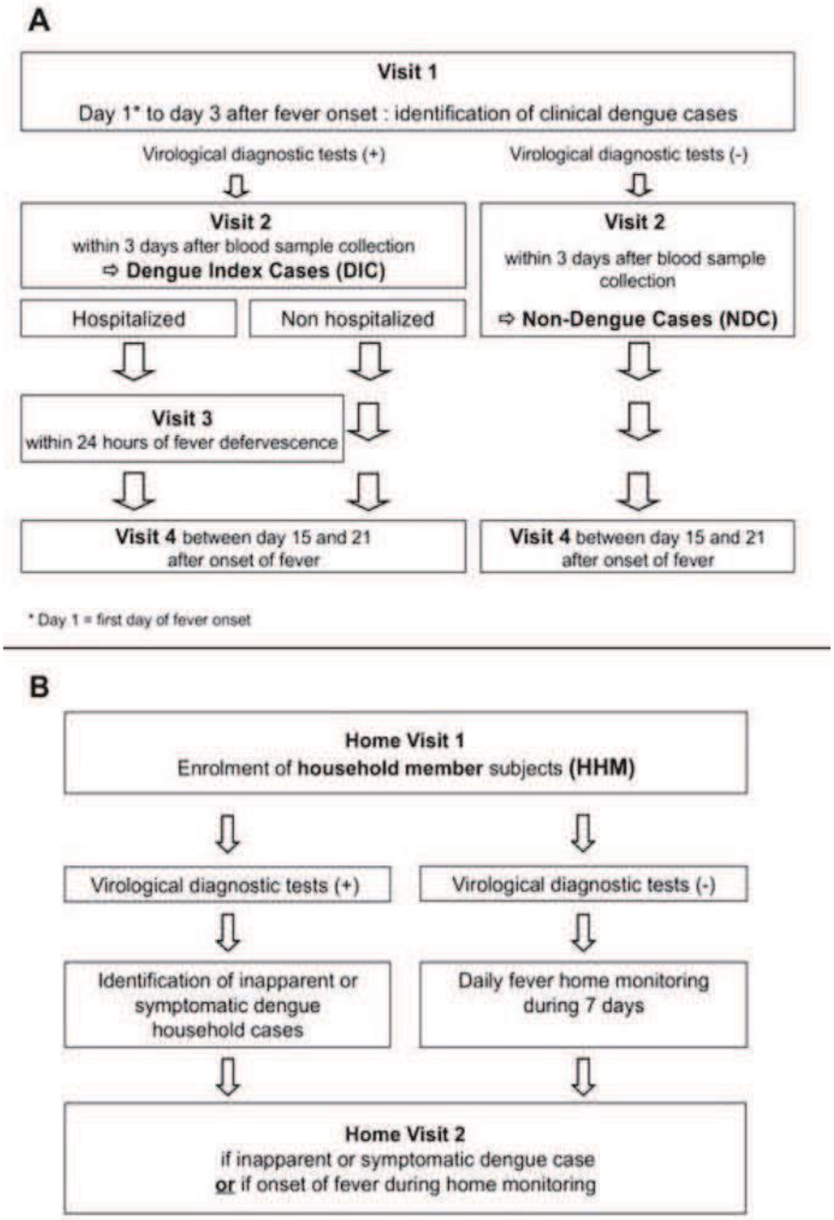

FIG. 6.2 – Study design charts.

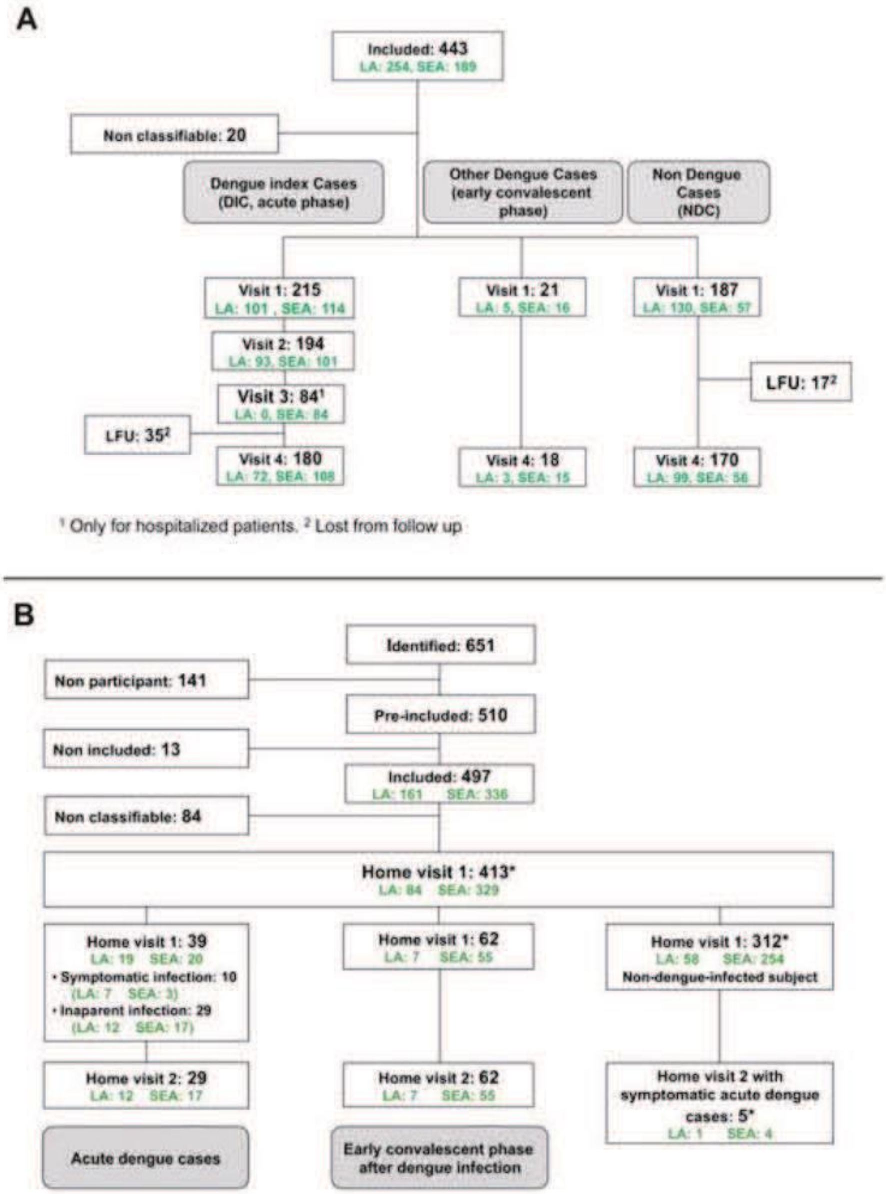

FIG. 6.3 – Flowcharts for the recruitment of participants at each step.

Chapitre 7

Application aux données de DENFRAME

Sommaire

7.1 Introduction

Dans cette application, nous considérons une étude de la dengue. La dengue est une infection virale survenant majoritairement dans les régions tropicales et subtropicales. Le virus de la dengue comprend quatre sérotypes (DENV-1, DENV-2, DENV-3, DENV-4), et se transmet par la piqûre du moustique de la fièvre jaune issu de la famille *Aedes aegypti* et du moustique tigre *Aedes albopticus*. Les deux types de moustiques sont des insectes diurnes. L'agent pathogène fait partie des *flavivirus*, notamment responsables de la fièvre jaune, de la méningo-encéphalite verno-estivale (MEVE) ainsi que du chikungunya. Le symptôme classique de la dengue est une fièvre élevée soudaine (comme pour la grippe), évoluant puis disparaissant sans

causer de complications en règle générale. L'évolution de la dengue peut être dangereuse lorsqu'elle s'accompagne d'hémorragies internes. Cette forme de complication entraîne encore beaucoup de décès, en particulier chez les enfants. La seule prévention possible consiste à prendre des mesures générales contre les piqûres d'insectes. Aucun vaccin n'est disponible à l'heure actuelle. La dengue est une maladie infectieuse à déclaration obligatoire.
L'infection par la dengue confère une immunité partielle et transitoire contre une infection ultérieure du même virus (voir (Dussart *et al.* 2012)).

7.2 Description des données

Nous considérons dans cette étude une base de données de taille $n = 528$, qui a été construite avec des individus recrutés au Brésil, au Cambodge, en Guyanne Française et au Vietnam. Selon les critères de sélection les individus sont classés cas index et membres de la fratrie (*household members* : tout individu ayant des liens biologiques avec un individu infecté (*DIC : Dengue Index Case*) et ayant habité au moins 7 jours avec celui-ci dans le même site).

Les individus qui répondent aux critères de sélection suivants sont classés cas index :
- Signature du consentement (ou signature d'un représentant pour les adolescents âgés de moins de 18 ans) ;
- Habiter dans un ménage où habitent plus de 2 personnes durant au moins les 7 derniers jours avant l'apparition de la maladie ;
- Être âgé au moins âgé de 2 ans ;
- Avoir une température supérieure à 38°C ;

- Apparition de fièvre dans les 72h ;
- Une des manifestations cliniques (autre que la fièvre) évoquant la dengue : maux de tête sévères, myalgie, éruption cutanée, douleurs rétro-orbitaires ou saignements ;
- Habiter dans un site accessible au moniteur.

L'inclusion se fait en deux étapes. Les cas de dengue tout d'abord. Ensuite les autres membres de leurs familles (père, mère, fratrie, etc...). Pour plus de détails, on peut se référer à la figure 6.3 dans le chapitre 6.

Chaque individu $i(1 \leq i \leq 528)$ est diagnostiqué pour l'infection par la dengue et est codé comme suit :

$$Y_i = \begin{cases} 1 & \text{si l'individu i est infecté} \\ 0 & \text{sinon} \end{cases}$$

Notons que si $Y_i = 0$, alors l'individu i peut être immune à l'instant du diagnostic (à cause, par exemple, d'une immunité temporaire acquise après une infection précédente) ou susceptible à l'infection de la dengue, bien que non infecté encore. Notre but est d'estimer le risque d'infection de ces individus, en utilisant cette base de données contenant les covariables[1] suivantes : *âge* (une variable continue bornée) et *poids* (codé 0 dans le cas de sous-poids et 1 sinon).

Une description détaillée des données est faite dans le chapitre précédent.

[1]L'étude pourrait être plus poussée en ajoutant dans le modèle d'autres covariables telles que, par exemple, les données cliniques (maux de tête, myalgie, température etc...), les données de laboratoire parmi lesquelles les prélèvements sanguins, des données sur les signes biologiques, les sérotypes de la dengue (DEN-1, DEN-2, DEN-3 et DEN-4), les tests confirmant les cas de dengue, présence antérieure de dengue et confirmation du laboratoire etc....

Dans la suite, nous présentons le modèle conjoint à utiliser et les résultats de l'analyse.

7.3 Modèles et résultats

Nous ajustons tout d'abord aux données aux données un modèle de régression logistique standard. Dans cette analyse "naïve", chaque individu non infecté est considéré comme susceptible. Le modèle final (4.1)-(4.2)-(4.3) est donné par

$$\begin{cases} \text{logit } \mathbf{P}(Y=1|\text{âge, poids}) = \beta_1 + \beta_2\text{âge} + \beta_3\text{poids} \\ \text{logit } \mathbf{P}(S=1) = \theta_1. \end{cases} \tag{7.1}$$

Puis nous ajustons le modèle conjoint final (4.1)-(4.2)-(4.3), où

$$\begin{cases} \text{logit } \mathbb{P}(Y=1|\text{âge, poids}, S=1) = \beta_1 + \beta_2\text{âge} + \beta_3\text{poids} \\ \text{logit } \mathbb{P}(S=1|\text{ poids}) = \theta_1 + \theta_2\text{poids}. \end{cases} \tag{7.2}$$

Les résultats sont présentés dans le tableau 7.1. Puis, nous estimons les paramètres β_1, β_2 et β_3 en utilisant la méthodologie développée dans la section 4.2 du chapitre 4. Notons tout d'abord que l'éventuelle immunité conférée par une infection précédente est juste transitoire, donc il n'y a aucune raison pour qu'un individu agé (qui a donc été exposé plus longtemps au risque d'infection par la dengue) ait une probabilité de devenir immune plus élevée qu'un individu moins âgé. En fait, la susceptibilité d'un individu à une infection par la dengue dépendra plutôt du fait que cet individu bénéficie ou non d'une quelconque action préventive (campagnes de sensibilisation contre la dengue, utilisation d'insecticides, élimination des eaux usées,...). De telles

TAB. 7.1 – Analyse des données de la dengue

	analyse naïve		modèle (4.1)-(4.2)-(4.3)	
paramètre	estimateur	écart-type	estimateur	écart-type
β_1	1.552	0.255	7.654	1.485
β_2	-0.055	0.007	-0.131	0.020
β_3	-0.813	0.207	-4.501	1.059
θ_1			0.497	0.159

informations ne sont pas disponibles dans notre base de données.
L'âge a été pris comme étant la variable V dans la condition C4 de la section 4.2 du chapitre 4.

Le test de Wald de l'hypothèse "$\theta_2 = 0$" dans le modèle (7.2) n'est pas significatif, nous avons retiré la variable *poids* du modèle pour la susceptibilité. Les résultats de cette modélisation sont donnés dans le tableau 7.1.

L'exécution de cette procédure de modélisation conjointe fournit l'estimation suivante : $1 - \exp(0.497)/(1 + \exp(0.497)) \approx 0.38$ pour la probabilité d'être immune. Alors, comme on pouvait s'y attendre, les probabilités d'infection estimées à partir de notre approche de modélisation sont plus élevées que celles obtenues à partir du modèle de régression logistique standard qui ne tient pas compte d'une possible immunité. Par exemple, les probabilités d'infection pour les individus âgés de 30 ans et 10 ans, avec un poids "normal", sont respectivement estimées à 0.29 et 0.55 (par le modèle de régression logistique standard) et à 0.31 et 0.86 (par notre approche).

On s'attend à ce que les sujets qui souffrent de sous-poids (ceux qui sont en dessous de leur poids normal) aient des risques d'infection élevés. Les probabilités d'infection pour les sujets en sous-poids âgés de 30 ans et de 10 ans sont respectivement estimées à 0.48 et 0.73 (par le modèle de régression logistique standard) et à 0.97 et 0.99 (par notre approche). Bien que les deux approches fournissent les mêmes conclusions qualitatives : la probabilité d'infection par la dengue est plus élevée pour les personnes très jeunes et qui sont en sous-poids (causé par la malnutrition, par exemple), elles diffèrent par-contre sur leurs estimations du risque d'infection. Notre approche tient compte d'une éventuelle immunité conférée par une infection antérieure et donc, il est raisonnable de penser que les probabilités d'infection provenant de ces estimations fournissent une image plus réaliste du risque d'infection pour cette base de données. En particulier, les estimations fournies par notre approche suggèrent que le sous-poids constitue un facteur de risque majeur de l'infection par la dengue, indépendamment de l'âge.

Conclusion et perspectives

Dans cette thèse, nous nous sommes intéressés à l'inférence statistique dans le modèle de régression logistique avec fraction immune. Notre approche, originale, qui consiste à proposer un modèle de mélange de deux modèles de régression logistiques s'avère plutôt efficace dans la pratique. Elle permet de prendre en compte une éventuelle immunité d'une partie de la population, même si les informations correspondant aux individus immunes sont manquantes.

Nous avons proposé une procédure d'estimation par maximum de vraisemblance dans le modèle de régression logistique avec fraction immune. Nous avons dans un premier temps décrit le modèle à excès de zéros proposé, puis nous en avons établi l'identifiabilité, sous un ensemble d'hypothèses aisément interprétables. Puis, nous avons montré l'existence de l'estimateur du maximum de vraisemblance dans ce modèle. Nous avons ensuite montré la consistance de cet estimateur, et nous avons enfin établi sa normalité asymptotique.

Nous avons comparé sur des données simulées les résultats de notre approche à ceux résultant d'une analyse naïve, dans laquelle nous supposons que les individus non infectés sont tous susceptibles. Cette comparaison suggère que la modélisation conjointe du statut d'infection et du statut d'immunité donne des estimations moins biaisées, même lorsque la proportion d'individus immunes est importante.

Nous avons ensuite proposé d'appliquer cette procédure de modélisation conjointe à l'analyse statistique de données sur la dengue. Les résultats montrent que les probabilités d'infection estimées à partir de notre approche de modélisation sont plus élevées que celles obtenues à partir d'un modèle de régression logistique standard qui ne tient pas compte d'une possible immunité. En particulier, les estimations fournies par notre approche suggèrent que le sous-poids constitue un facteur de risque majeur de l'infection par la dengue, indépendamment de l'âge.

La procédure d'estimation que nous avons proposée suppose néanmoins que le modèle pour le statut d'immunité est bien spécifié. Il serait donc intéressant d'étudier l'effet d'une mauvaise spécification de ce modèle (en particulier, de la fonction de lien) sur les résultats de nos analyses. Les techniques et résultats établis par (Czado & Santner 1992) peuvent être très utiles pour ce problème. Il serait également intéressant d'étudier le modèle de régression logistique avec fraction immune, en présence d'un très grand nombre de variables explicatives. Des études récentes (voir (Huang *et al.* 2008) et (Meier *et al.* 2008)) ont considéré le problème de l'estimation dans le modèle de régression logistique standard (sans fraction immune) lorsque le nombre de covariables est supérieur à la taille de l'échantillon, comme dans les études génétiques par exemple. Mais le problème reste entier en présence d'immunité.

Dans un deuxième temps, nous nous sommes intéressés à la construction de bandes de confiance simultanées pour la probabilité d'infection, dans le modèle de régression logistique avec fraction immune. Nous avons proposé trois méthodes de constructions de bandes de confiance pour la fonction de régression. La première méthode utilise la propriété de normalité asymptotique de l'estimateur du maximum de vraisemblance, et une approximation par une loi du χ^2 pour approcher le quantile. La deuxième utilise également la propriété de normalité asymptotique de l'estimateur du maximum de vraisemblance et est basée sur une égalité classique de (Landau & Sheep 1970). Cette inégalité nous a permis d'estimer la loi du supremum de processus gaussien. La troisième méthode repose sur des simulations, pour estimer le quantile approprié de la loi du supremum d'un processus gaussien. La construction des bandes de confiance via un processus gaussien demande une estimation de la fonction de covariance de l'estimateur.

A titre de travail futur, il serait intéressant d'étendre notre étude des bandes de confiance aux régions de confiance, en présence de plusieurs variables explicatives. Les méthodes utilisées pour la construction de bandes de confiance simultanées peuvent également être utilisées pour la construction de régions de confiance en présence de plusieurs covariables. On peut consulter (Li *et al.* 2008) à ce propos. La construction de régions de confiance lorsque le nombre de covariables est supérieur à 1 constitue une tâche difficile. Notre approche pourrait être étendue à cet effet.

Quatrième partie

Annexes

Annexe A

Rappels et preuves complémentaires

A.1 Rappels mathématiques

Théorème A.1.1 *(Théorème de Slutsky) Soit W_n et Z_n deux suites de variables aléatoires tels que $W_n \longrightarrow W$ en loi et $Z_n \longrightarrow c$ en probabilité, où c est une constante. Alors*

$$Z_n W_n \longrightarrow cW \text{ en loi}, \quad Z_n + W_n \longrightarrow c + W \text{ en loi}.$$

Théorème A.1.2 *(Delta-Méthode) Soit X_n une suite de vecteurs aléatoires dans $\mathbb{R}^p$ satisfaisant $\sqrt{n}(X_n - \theta) \longrightarrow \mathcal{N}(0, V)$ en loi. Soit $F : \mathbb{R}^p \longrightarrow \mathbb{R}$ différentiable en θ. Alors $\sqrt{n}(F(X_n) - F(\theta)) \longrightarrow \mathcal{N}(0, [\nabla F(\theta)]^\top V \nabla F(\theta))$ en loi.*

A.2 Bootstrap

Le *bootstrap* représente un outil puissant dans l'inférence statistique. Sa motivation ((Efron 1982), (Efron & Tibshirani 1993)) est d'approcher par simulation (*Monte Carlo*) la distribution d'un estimateur lorsque l'on ne connaît pas la loi de l'échantillon ou, plus souvent lorsque l'on ne peut pas supposer qu'elle est gaussienne. L'objectif est de remplacer des hypothèses probabiliste pas toujours vérifiées ou même invérifiables par des simulations et donc beaucoup de calcul.

Le principe fondamental de cette technique de ré-échantillonnage est de substituer à la distribution de probabilité inconnue F, dont est issu l'échantillon d'apprentissage, la distribution empirique $\hat{F}$ qui donne un poids $1/n$ à chaque réalisation. Ainsi, on obtient un échantillon de taille n, dit échantillon bootstrap, selon la distribution empirique $\hat{F}$ par n tirages aléatoires avec remise parmi les n observations initiales.

Il est facile de construire un grand nombre d'échantillons bootstrappés sur lesquels on calcule l'estimateur concerné. La loi simulée de cet estimateur est une approximation asymptotiquement convergente sous des hypothèses raisonnables[1] de la loi de l'estimateur. Cette approximation fournit ainsi des estimations du biais, de la variance, donc d'un risque quadratique, et même des intervalles de confiance du paramètre considéré sans hypothèse (de normalité par exemple) sur la vraie loi.

A.3 Preuves complémentaires

Preuve du corollaire 4.4.6 *Soit ϕ une application de $\mathbb{R}^k$ à valeurs dans $\mathbb{R}^p$ définie par $\phi(x,y) = x$ avec $x \in \mathbb{R}^p$ et $y \in \mathbb{R}^q$. ϕ est linéaire et continue de matrice M. En appliquant la delta méthode multivariée au théorème 4.4.1, on obtient*

$$\sqrt{n}(\phi(\widehat{\psi}_n) - \phi(\psi)) \xrightarrow{\mathscr{L}} \mathscr{N}(0, M\mathscr{I}_{\psi}^{-1}(\psi)M^{\top}).$$

[1] échantillon indépendant de même loi et estimateur indépendant de l'ordre des observations

Ce qui donne, par suite

$$\sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow{\mathscr{L}} \mathscr{N}(0, M\mathscr{I}_\psi^{-1}(\psi)M^\top).$$

Preuve du théorème 4.4.7 *On suppose le Corollaire 4.4.6 vérifié. Posons* $F(x) = e^x/1+e^x$*. F est différentiable. En appliquant la Delta-Méthode à la fonction F, au point* $\widehat{\beta}_n^\top x$*,* $\sqrt{n}(\widehat{p}_n(\mathbf{x}) - p(\mathbf{x}))$ *converge en loi vers une variable Gaussienne de moyenne nulle et de variance*

$$[\nabla F(\theta)]^\top \mathrm{var}(\widehat{\beta}_n)\nabla F(\theta) = \exp(2\beta^\top \mathbf{x}).\mathbf{x}^\top M\mathscr{I}_\psi^{-1}M^\top \mathbf{x}/(1+\exp(\beta^\top \mathbf{x}))^4.$$

Preuve du Lemme 5.3.1.

La matrice V^{-1} est réelle, symétrique et, définie positive, tout comme V. Il existe donc une matrice $A \in \mathscr{M}(m \times m)$ telle que $V^{-1} = A^\top A$. *En effet, une matrice réelle symétrique étant diagonalisable sur* $\mathbb{R}$ *admet la décomposition spectrale* $M = P^\top DP$ *:* P*, la matrice dont les colonnes sont les vecteurs propres, peut être choisie orthogonale ;* D *est la matrice diagonale des valeurs propres. Si* M *est définie positive, les valeurs propres sont positives ou nulles. Il suffit alors de prendre* $A = \sqrt{D}P$ *pour obtenir* $M = A^\top A$. Donc, $Z^\top V^{-1} Z = \|AZ\|^2$. Le résultat est alors une conséquence du fait que le vecteur aléatoire AZ suit une loi normale multivariée de moyenne 0 et de matrice de variance I_m, puisque, en effet,

$$\mathrm{var}(AZ) = A\mathrm{var}(Z)A^\top = AVA^\top = AA^{-1}(A^\top)^{-1}A^\top = I_m.$$

Preuve du Lemme 5.3.2. Le point 1 du Lemme est obtenu en posant $U = BX$, et $V = BY$ et en appliquant l'inégalité de Cauchy Schwartz sur U et V. De manière similaire, on montre le point 2 en choisissant $U = (B^{-1})^\top Y$ et $V = BX$.

Annexe B

Résultats complets des simulations

B.1 Résultats de simulations du chapitre 4

B.1.1 Modèle

Nous considérons les modèles suivants pour le statut d'infection :

$$\begin{cases} \log\left(\frac{\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}{1-\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}\right) = \beta_1 + \beta_2 X_{i2} + \beta_3 Z_{i2} + \beta_4 Z_{i3} + \beta_5 Z_{i4} & \text{if } S_i = 1 \\ \mathbb{P}(Y=1|\mathbf{X}_i,S_i) = 0 & \text{if } S_i = 0 \end{cases} \tag{2.1}$$

et pour le statut d'immunité :

$$\log\left(\frac{\mathbb{P}(S=1|\mathbf{Z}_i)}{1-\mathbb{P}(S=1|\mathbf{Z}_i)}\right) = \theta_1 + \theta_2 Z_{i2} + \theta_3 Z_{i3} + \theta_4 Z_{i4}, \tag{2.2}$$

où $X_{i2} \sim N(0,1)$, $Z_{i2} \sim N(1,1)$, et Z_{i3} et Z_{i4} sont des variables indicatrices construites à partir d'une variable qualitative à trois catégories. Notons que la variable X_{i2} joue le rôle de la variable continue V dans la condition C4 dans le chapitre 4.

Les modèles suivants sont considérés :

– modèle $\mathscr{M}_1$: $\beta = (-1.7, -2, -3.4, 5, 0.3)^\top$: approximativement 30% des susceptibles sont infectés.

- modèle $\mathscr{M}_2$: $\beta = (1.5, -2.3, 2.5, -3.5, 0.5)^\top$: approximativement 70% des susceptibles sont infectés.
- modèle $\mathscr{M}_3$: $\beta = (-1.7, -2.8, 0, -0.7, 1.1)^\top$: approximativement 30% des susceptibles sont infectés.
- modèle $\mathscr{M}_4$: $\beta = (1.5, -2, 0, 3.5, -4)^\top$: approximativement 70% des susceptibles sont infectés.

B.1.2 Résultats

Tab. B.1: Modèle $\mathscr{M}_1$: $\beta = (-1.7, -2, -3.4, 5, .3)^\top$.

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
Pourcentage d'immunes = 25%, $\theta = (.71, 1, 2, -3)$									
100	-1.688	-2.019	-3.417	4.648	0.322	0.667	1.154	2.153	-3.088
	(1.625)	(1.160)	(1.704)	(1.866)	(1.880)	(2.660)	(1.462)	(1.715)	(2.663)
	[1.330]	[0.689]	[1.387]	[1.525]	[1.524]	[2.199]	[1.161]	[1.410]	[2.156]
		0.631*	0.539*	0.316*	0.013*				
500	-1.710	-2.006	-3.392	4.991	0.311	0.677	1.076	2.079	-2.994
	(0.966)	(0.427)	(0.923)	(1.167)	(1.581)	(0.849)	(0.804)	(1.144)	(2.090)
	[0.734]	[0.330]	[0.686]	[0.907]	[1.329]	[0.711]	[0.643]	[0.910]	[1.663]
		0.997*	0.993*	0.984*	0.090*				
1000	-1.702	-2.004	-3.398	4.968	0.305	0.697	1.046	2.026	-2.997
	(0.579)	(0.297)	(0.584)	(0.797)	(0.843)	(0.749)	(0.623)	(0.779)	(1.127)
	[0.456]	[0.233]	[0.412]	[0.612]	[0.716]	[0.611]	[0.477]	[0.605]	[0.910]
		1*	0.998*	1*	0.107*				
1500	-1.720	-1.998	-3.410	4.971	0.305	0.709	1.035	2.013	-3.002
	(0.492)	(0.272)	(0.474)	(0.649)	(0.794)	(0.607)	(0.475)	(0.614)	(0.979)
	[0.384]	[0.206]	[0.332]	[0.484]	[0.675]	[0.487]	[0.361]	[0.484]	[0.778]
		1*	1*	1*	0.095*				
Pourcentage d'immunes = 50%, $\theta = (-.3, -1, 2.1, 1)$									
100	-1.767	-2.105	-3.341	5.334	0.317	-0.377	-1.123	2.212	1.090
	(2.068)	(1.013)	(1.783)	(2.557)	(2.758)	(2.672)	(1.965)	(2.156)	(2.851)
	[1.682]	[0.784]	[1.464]	[2.150]	[2.312]	[2.141]	[1.265]	[1.760]	[2.366]

n	$\widehat{\beta}_n$: $\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_n$: $\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
		0.335*	0.337*	0.127*	0*				
500	-1.716	-2.081	-3.576	5.217	0.294	-0.342	-1.092	2.097	1.078
	(1.417)	(0.545)	(0.930)	(1.704)	(1.704)	(1.079)	(0.760)	(1.518)	(1.746)
	[1.085]	[0.452]	[0.743]	[1.383]	[1.320]	[0.875]	[0.554]	[1.177]	[1.401]
		1*	1*	0.839*	0.057*				
1000	-1.701	-2.076	-3.529	5.015	0.304	-0.318	-1.056	2.105	1.031
	(0.780)	(0.391)	(0.627)	(1.072)	(1.113)	(0.743)	(0.473)	(1.028)	(1.160)
	[0.650]	[0.316]	[0.495]	[0.866]	[0.892]	[0.607]	[0.352]	[0.770]	[0.905]
		1*	1*	0.984*	0.045*				
1500	-1.698	-2.013	-3.482	5.007	0.303	-0.315	-1.023	2.103	1.024
	(0.694)	(0.294)	(0.489)	(0.857)	(0.865)	(0.609)	(0.384)	(0.885)	(0.926)
	[0.568]	[0.242	[0.381]	[0.694]	[0.685]	[0.497]	[0.290]	[0.642]	[0.721]
		1*	1*	0.999*	0.057*				
Pourcentage d'immunes = 75%, $\theta = (.4, -1, -.6, -2)$									
100	-1.661	-2.131	-3.387	4.830	0.332	0.410	-1.059	-0.610	-2.158
	(2.139)	(2.127)	(2.803)	(3.283)	(3.661)	(3.554)	(1.995)	(3.118)	(2.974)
	[1.754]	[1.720]	[2.394]	[2.849]	[2.811]	[2.823]	[1.380]	[2.587]	[2.511]
		0.043*	0.064*	0.005*	0*				
500	-1.673	-2.075	-3.435	4.987	0.325	0.407	-1.060	-0.607	-1.921
	(1.436)	(1.012)	(1.614)	(2.455)	(2.198)	(1.295)	(0.598)	(1.651)	(1.884)
	[1.103]	[0.848]	[1.337]	[2.039]	[1.765]	[1.046]	[0.409]	[1.290]	[1.471]
		0.747*	0.787*	0.641*	0.046*				
1000	-1.545	-2.053	-3.399	5.024	0.309	0.405	-1.045	-0.604	-1.980
	(0.847)	(0.783)	(1.157)	(2.069)	(1.394)	(0.940)	(0.344)	(1.006)	(1.044)
	[0.669]	[0.619]	[0.899]	[1.586]	[1.125]	[0.737]	[0.253]	[0.775]	[0.843]
		0.994*	0.970*	0.895*	0.083*				
1500	-1.595	-2.017	-3.410	5.024	0.306	0.404	-1.035	-0.605	-1.997
	(0.708)	(0.630)	(0.909)	(0.895)	(1.234)	(0.749)	(0.259)	(0.866)	(0.860)
	[0.543]	[0.492]	[0.679]	[0.738]	[0.991]	[0.606]	[0.196]	[0.661]	[0.679]
		1*	0.993*	0.937*	0.089*				

TAB. B.2: $\mathscr{M}_2 : \beta = (1.5, -2.3, 2.5, -3.5, .5)$.

n	$\widehat{\beta}_n$: $\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_n$: $\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
Pourcentage d'immunes = 25%, $\theta = (.71, 1, 2, -3)$									

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
100	1.512	-2.369	2.522	-3.518	0.562	0.725	1.191	2.061	-2.896
	(1.413)	(1.165)	(1.189)	(2.180)	(1.853)	(0.979)	(0.830)	(2.728)	(1.646)
	[1.190]	[0.917]	[0.949]	[1.810]	[1.582]	[0.789]	[0.636]	[2.309]	[1.264]
		0.800*	0.814*	0.271*	0.236*				
500	1.508	-2.313	2.520	-3.497	0.514	0.714	1.076	2.045	-2.959
	(0.905)	(0.635)	(0.568)	(1.198)	(0.633)	(0.419)	(0.434)	(1.617)	(0.652)
	[0.724]	[0.478]	[0.435]	[0.937]	[0.545]	[0.335]	[0.268]	[1.290]	[0.454]
		0.993*	0.993*	0.991*	0.629*				
1000	1.499	-2.297	2.508	-3.502	0.512	0.712	1.071	2.025	-2.985
	(0.569)	(0.488)	(0.398)	(0.908)	(0.557)	(0.308)	(0.387)	(1.178)	(0.365)
	[0.453]	[0.335]	[0.286]	[0.663]	[0.479]	[0.241]	[0.204]	[0.941]	[0.273]
		0.999*	0.998*	0.997*	0.732*				
1500	1.499	-2.299	2.497	-3.503	0.504	0.708	1.050	2.012	-2.985
	(0.339)	(0.372)	(0.322)	(0.701)	(0.522)	(0.257)	(0.337)	(0.983)	(0.289)
	[0.331]	[0.252]	[0.224]	[0.508]	[0.447]	[0.204]	[0.174]	[0.766]	[0.225]
		1*	1*	1*	0.764*				
Pourcentage d'immunes = 50%, $\theta = (-.3, -1, 2.1, 1)$									
100	1.526	-2.328	2.339	-3.336	0.488	-0.332	-1.107	2.179	1.053
	(1.887)	(1.824)	(2.170)	(2.570)	(2.181)	(0.902)	(1.204)	(1.946)	(1.567)
	[1.577]	[1.507]	[1.806]	[2.174]	[1.654]	[0.679]	[0.809]	[1.391]	[1.158]
		0.411*	0.228*	0.101*	0.060*				
500	1.517	-2.295	2.635	-3.397	0.537	-0.284	-0.983	2.127	1.041
	(0.956)	(0.772)	(0.687)	(1.352)	(0.826)	(0.317)	(0.472)	(0.466)	(1.043)
	[0.775]	[0.580]	[0.530]	[1.045]	[0.648]	[0.255]	[0.244]	[0.361]	[0.678]
		0.999*	0.963*	0.924*	0.339*				
1000	1.517	-2.303	2.563	-3.408	0.512	-0.296	-1.023	2.110	1.026
	(0.650)	(0.531)	(0.467)	(0.962)	(0.573)	(0.207)	(0.157)	(0.310)	(0.532)
	[0.518]	[0.390]	[0.364]	[0.701]	[0.454]	[0.169]	[0.123]	[0.243]	[0.366]
		1*	0.998*	0.999*	0.399*				
1500	1.498	-2.299	2.531	-3.365	0.513	-0.295	-1.012	2.112	0.995
	(0.473)	(0.389)	(0.355)	(0.768)	(0.455)	(0.181)	(0.122)	(0.260)	(0.339)
	[0.384]	[0.281]	[0.281]	[0.553]	[0.373]	[0.146]	[0.095]	[0.208]	[0.253]
		1*	1*	1*	0.451*				
Pourcentage d'immunes = 75%, $\theta = (.4, -1, -.6, -2)$									
100	1.489	-2.384	2.521	-3.458	0.554	0.420	-1.116	-0.557	-2.123
	(2.356)	(2.356)	(2.263)	(2.793)	(2.326)	(1.163)	(1.602)	(1.665)	(1.904)

n	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
	[1.934]	[1.979]	[1.944]	[2.389]	[1.902]	[0.875]	[1.121]	[1.232]	[1.450]
		0.109*	0.070*	0.106*	0.055*				
500	1.515	-2.381	2.523	-3.482	0.548	0.418	-1.047	-0.633	-2.084
	(1.282)	(1.405)	(1.701)	(1.910)	(1.757)	(0.394)	(0.396)	(0.651)	(0.665)
	[1.047]	[1.133]	[1.317]	[1.527]	[1.420]	[0.322]	[0.268]	[0.482]	[0.491]
		0.929*	0.899*	0.756*	0.353*				
1000	1.490	-2.295	2.522	-3.515	0.521	0.415	-1.031	-0.627	-1.997
	(0.862)	(0.916)	(1.141)	(1.228)	(1.282)	(0.243)	(0.185)	(0.345)	(0.398)
	[0.694]	[0.709]	[0.793]	[0.958]	[1.013]	[0.198]	[0.142]	[0.271]	[0.308]
		0.999*	0.999*	0.929*	0.367*				
1500	1.508	-2.304	2.497	-3.512	0.513	0.415	-1.023	-0.613	-1.997
	(0.721)	(0.683)	(0.867)	(0.957)	(0.961)	(0.221)	(0.149)	(0.293)	(0.299)
	[0.573]	[0.505]	[0.575]	[0.728]	[0.765]	[0.181]	[0.119]	[0.218]	[0.236]
		1*	1*	0.973*	0.359*				

Dans le but d'évaluer le niveau d'un test de type Wald d'hypothèses de la forme H_0 : $\beta_l = 0$ dans le modèle ZIB (4.1)-(4.2) du chapitre 4, nous simulons des échantillons à partir des modèles $\mathscr{M}_3$ et $\mathscr{M}_4$. Les résultats sont respectivement donnés dans les tableaux B.3 et B.4. Dans ces tableaux, * représente la puissance empirique du test de Wald (lorsque $\beta_l \neq 0$) et † représente le niveau empirique (lorsque $\beta_l = 0$).

Tab. B.3: $\mathscr{M}_3 : \beta = (-1.7, -2.8, 0, -.7, 1.1)$.

n	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
Pourcentage d'immunes = 25%, $\theta = (.71, 1, 2, -3)$									
100	-1.732	-2.918	0.017	-0.764	1.190	0.754	1.169	1.857	-2.975
	(1.630)	(0.631)	(1.310)	(1.721)	(1.857)	(2.681)	(1.686)	(1.983)	(2.877)
	[1.258]	[0.502]	[0.851]	[1.297]	[1.569]	[2.080]	[1.240]	[1.675]	[2.325]
		0.818*	0.033†	0.053*	0.052*				
500	-1.688	-2.908	-0.016	-0.751	1.114	0.729	1.164	2.139	-3.106
	(0.570)	(0.386)	(0.352)	(0.741)	(1.231)	(0.837)	(0.634)	(1.494)	(1.162)
	[0.441]	[0.312]	[0.196]	[0.465]	[0.956]	[0.669]	[0.449]	[1.224]	[0.898]

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
		0.998*	0.055†	0.301*	0.174*				
1000	-1.698	-2.853	-0.004	-0.713	1.079	0.726	1.080	2.083	-3.088
	(0.385)	(0.272)	(0.145)	(0.383)	(0.781)	(0.553)	(0.363)	(1.103)	(0.771)
	[0.304]	[0.217]	[0.115]	[0.301]	[0.608]	[0.442]	[0.269]	[0.913]	[0.596]
		1*	0.055†	0.494*	0.294*				
1500	-1.704	-2.837	-0.004	-0.705	1.110	0.716	1.057	2.071	-3.087
	(0.301)	(0.210)	(0.126)	(0.277)	(0.650)	(0.452)	(0.292)	(0.906)	(0.680)
	[0.234]	[0.168]	[0.101]	[0.224]	[0.489]	[0.359]	[0.205]	[0.743]	[0.503]
		1*	0.049†	0.707*	0.479*				
Pourcentage d'immunes = 50%, $\theta = (-.3, -1, 2.1, 1)$									
100	-1.776	-2.912	-0.039	-0.776	1.203	-0.336	-1.116	1.994	1.108
	(1.879)	(0.986)	(1.824)	(1.782)	(2.056)	(2.904)	(2.260)	(2.630)	(2.878)
	[1.542]	[0.807]	[1.404]	[1.493]	[1.709]	[2.143]	[1.533]	[2.053]	[2.129]
		0.472*	0.076†	0.008*	0.018*				
500	-1.753	-2.918	-0.030	-0.768	1.194	-0.279	-0.974	2.197	1.035
	(1.191)	(0.590)	(0.490)	(1.361)	(1.307)	(0.752)	(0.806)	(1.312)	(1.053)
	[0.919]	[0.481]	[0.371]	[1.056]	[1.021]	[0.590]	[0.456]	[0.929]	[0.747]
		1*	0.126†	0.108*	0.196*				
1000	-1.718	-2.853	0.005	-0.719	1.127	-0.288	-1.003	2.149	1.021
	(0.647)	(0.417)	(0.293)	(0.875)	(0.899)	(0.509)	(0.522)	(0.833)	(0.654)
	[0.525]	[0.335]	[0.224]	[0.682]	[0.710]	[0.414]	[0.259]	[0.591]	[0.495]
		1*	0.084†	0.148*	0.295*				
1500	-1.696	-2.824	-0.002	-0.705	1.117	-0.303	-1.020	2.119	1.021
	(0.551)	(0.329)	(0.310)	(0.669)	(0.701)	(0.387)	(0.304)	(0.561)	(0.490)
	[0.442]	[0.259]	[0.181]	[0.517]	[0.557]	[0.314]	[0.186]	[0.423]	[0.385]
		0.999*	0.078†	0.190*	0.383*				
Pourcentage d'immunes = 75%, $\theta = (.4, -1, -.6, -2)$									
100	-1.684	-2.948	-0.027	-0.792	1.215	0.497	-1.188	-0.587	-2.120
	(2.086)	(1.581)	(1.912)	(1.939)	(2.648)	(3.491)	(2.037)	(2.879)	(2.731)
	[1.689]	[1.313]	[1.591]	[1.621]	[2.224]	[2.297]	[1.493]	[2.183]	[2.215]
		0.127*	0.027†	0.013*	0.003*				
500	-1.774	-2.898	-0.028	-0.746	1.197	0.476	-0.923	-0.592	-1.905
	(1.392)	(0.908)	(0.993)	(1.651)	(1.898)	(0.952)	(1.042)	(1.396)	(1.664)
	[0.976]	[0.750]	[0.752]	[1.215]	[1.567]	[0.720]	[0.616]	[0.959]	[1.156]
		0.932*	0.162†	0.141*	0.084*				
1000	-1.745	-2.851	-0.004	-0.746	1.162	0.473	-0.925	-0.595	-1.927

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
	(0.731)	(0.631)	(0.587)	(0.934)	(1.430)	(0.648)	(0.797)	(0.742)	(0.978)
	[0.540]	[0.512]	[0.435]	[0.715]	[1.157]	[0.477]	[0.381]	[0.545]	[0.678]
		1^*	$0.125^\dagger$	0.205^*	0.143^*				
1500	-1.733	-2.831	-0.004	-0.746	1.098	0.461	-0.970	-0.595	-1.960
	(0.596)	(0.514)	(0.360)	0.677	(1.289)	(0.513)	(0.447)	(0.572)	(0.845)
	[0.407]	[0.410]	[0.282]	[0.538]	[1.002]	[0.383]	[0.247]	[0.417]	[0.545]
		1^*	$0.114^\dagger$	0.247^*	0.186^*				

TAB. B.4: $\mathscr{M}_4 : \beta = (1.5, -2, 0, 3.5, -4)$.

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
Pourcentage d'immunes $= 25\%$, $\theta = (.71, 1, 2, -3)$									
100	1.663	-2.215	-0.004	3.303	-3.512	0.793	0.939	2.216	-2.951
	(1.209)	(1.236)	(1.592)	(1.775)	(1.895)	(1.185)	(1.108)	(1.304)	(1.989)
	[1.005]	[0.975]	[1.125]	[1.377]	[1.567]	[0.926]	[0.856]	[1.050]	[1.604]
		0.364^*	$0.026^\dagger$	0.364^*	0.117^*				
500	1.580	-2.230	0.039	3.614	-3.776	0.671	1.132	2.130	-2.943
	(0.991)	(0.959)	(0.442)	(0.842)	(1.528)	(0.486)	(0.745)	(1.041)	(1.805)
	[0.793]	[0.660]	[0.339]	[0.679]	[1.108]	[0.391]	[0.454]	[0.713]	[1.324]
		0.954^*	$0.137^\dagger$	0.935^*	0.778^*				
1000	1.548	-2.096	0.028	3.559	-3.988	0.685	1.051	2.081	-2.954
	(0.664)	(0.712)	(0.256)	(0.565)	(1.098)	(0.326)	(0.419)	(0.654)	(1.241)
	[0.519]	[0.429]	[0.198]	[0.452]	[0.820]	[0.262]	[0.256]	[0.427]	[0.882]
		1^*	$0^\dagger$	1^*	0.909^*				
1500	1.514	-2.036	0.010	3.538	-3.997	0.707	1.031	2.031	-2.975
	(0.560)	(0.589)	(0.206)	(0.431)	(0.958)	(0.312)	(0.299)	(0.420)	(0.871)
	[0.429]	[0.314]	[0.161]	[0.349]	[0.697]	[0.246]	[0.192]	[0.318]	[0.669]
		0.965^*	$0.077^\dagger$	0.965^*	0.958^*				
Pourcentage d'immunes $= 50\%$, $\theta = (-.3, -1, 2.1, 1)$									
100	1.472	-1.931	0.022	3.378	-3.427	-0.384	-0.913	2.286	1.276
	(1.790)	(1.663)	(2.130)	(1.998)	(2.231)	(0.979)	(1.487)	(1.607)	(1.563)
	[1.546]	[1.447]	[1.530]	[1.691]	[1.946]	[0.690]	[0.924]	[1.193]	[1.284]
		0.114^*	$0.184^\dagger$	0.163^*	0.005^*				
500	1.484	-2.034	0.004	3.428	-3.446	-0.339	-0.934	2.155	1.054
	(1.236)	(1.247)	(0.800)	(1.180)	(1.826)	(0.303)	(0.411)	(1.225)	(1.196)

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
	[1.020]	[0.936]	[0.518]	[0.999]	[1.505]	[0.248]	[0.232]	[0.633]	[0.945]
		0.878*	0.235†	0.734*	0.608*				
1000	1.490	-1.956	0.008	3.475	-3.482	-0.325	-0.976	2.148	1.052
	(0.809)	(0.877)	(0.378)	(0.960)	(1.565)	(0.206)	(0.237)	(0.653)	(0.847)
	[0.652]	[0.603]	[0.282]	[0.812]	[1.187]	[0.166]	[0.130]	[0.401]	[0.658]
		0.994*	0.206†	0.959*	0.845*				
1500	1.490	-1.987	-0.001	3.492	-3.763	-0.308	-0.982	2.092	1.032
	(0.570)	(0.637)	(0.322)	(0.784)	(0.989)	(0.178)	(0.208)	(0.475)	(0.649)
	[0.458]	[0.421]	[0.238]	[0.662]	[0.769]	[0.145]	[0.106]	[0.299]	[0.513]
		1*	0.172†	0.989*	0.892*				
Pourcentage d'immunes = 75%, $\theta = (.4, -1, -.6, -2)$									
100	1.462	-1.936	-0.012	3.380	-3.520	0.508	-1.137	-0.710	-2.238
	(1.937)	(2.112)	(2.375)	(2.509)	(2.581)	(1.050)	(1.710)	(1.880)	(1.705)
	[1.643]	[1.790]	[1.833]	[2.211]	[2.146]	[0.791]	[1.018]	[1.329]	[1.445]
		0.042*	0.143†	0.007*	0.012*				
500	1.456	-1.939	-0.020	3.453	-3.466	0.485	-0.933	-0.674	-2.156
	(1.493)	(1.418)	(0.820)	(1.900)	(2.061)	(0.478)	(0.410)	(0.644)	(1.395)
	[1.202]	[1.162]	[0.633]	[1.585]	[1.626]	[0.372]	[0.268]	[0.450]	[1.156]
		0.449*	0.297†	0.116*	0.231*				
1000	1.477	-1.947	0.014	3.480	-3.785	0.462	-0.951	-0.645	-2.051
	(1.059)	(1.084)	(0.531)	(1.521)	(1.601)	(0.339)	(0.300)	(0.409)	(1.196)
	[0.851]	[0.857]	[0.419]	[1.222]	[1.273]	[0.259]	[0.170]	[0.291]	[0.969]
		0.912*	0.258†	0.489*	0.450*				
1500	1.482	-1.975	-0.011	3.492	-3.801	0.437	-0.961	-0.642	-2.021
	(0.741)	(0.731)	(0.368)	(1.250)	(1.091)	(0.236)	(0.214)	(0.345)	(1.050)
	[0.597]	[0.561]	[0.290]	[0.969]	[0.905]	[0.190]	[0.121]	[0.242]	[0.838]
		0.971*	0.268†	0.692*	0.558*				

Lorsqu'il n'y a aucun individu immune dans l'échantillon et qu'on le sait avant d'analyser les données, un modèle de régression logistique standard peut être ajusté aux données. Les résultats alors obtenus sont intéressants, car ils fournissent un point de repère pour évaluer les performances de l'estimateur du maximum de vraisemblance de β dans le modèle ZIB.

Le tableau B.5 donne les résultats fournis par un modèle de régression logistique standard sur les modèles $\mathscr{M}_1$ et $\mathscr{M}_3$, lorsqu'il y a absence d'individus immunes dans l'échantillon.

Ensuite, nous comparons ces résultats à ceux obtenus par la méthode d'analyse "naïve" où :

– nous considérons tout individu i tel que $\{Y_i = 0\}$ comme susceptible mais non infecté (c'est-à-dire nous ignorons une éventuelle immunité de cet individu),

– nous appliquons un modèle de régression logistique standard aux données ainsi obtenues.

Les résultats de cette analyse "naïve" pour le modèle $\mathscr{M}_1$ sont donnés dans le tableau B.6.

Tab. B.6: $\mathscr{M}_1$: $\beta = (-1.7, -2, -3.4, 5, .3)$ quand on ignore l'immunité.

	$\widehat{\beta}_n$				
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$
Pourcentage d'immunes = 25%					
100	-1.093	-0.376	0.115	2.358	-1.288
	(3.448)	(3.792)	(4.844)	(5.860)	(4.619)
	[2.082]	[2.048]	[3.721]	[5.131]	[2.526]
		0.032*	0.294*	0.171*	0.497*
500	-1.376	-0.085	0.265	1.815	-1.572
	(3.609)	(1.929)	(4.058)	(5.159)	(2.845)
	[2.179]	[1.917]	[3.751]	[4.657]	[2.227]
		0.053*	0.559*	0.446*	0.701*
1000	-1.290	-0.158	0.158	2.410	-0.966
	(2.944)	(2.182)	(3.565)	(4.871)	(2.760)
	[2.019]	[1.921]	[3.558]	[4.304]	[1.843]
		0.046*	0.624*	0.535*	0.751*
1500	-1.171	-0.112	0.162	1.962	-1.044
	(2.138)	(1.902)	(3.570)	(4.885)	(2.659)
	[1.867]	[1.890]	[3.562]	[4.437]	[1.847]

	$\widehat{\beta}_n$				
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$
		0.052*	0.652*	0.570*	0.778*
Pourcentage d'immunes = 50%					
100	-1.953	-0.228	-1.048	3.165	0.171
	(3.843)	(4.118)	(5.234)	(5.194)	(0.506)
	[1.745]	[2.181]	[3.101]	[4.347]	[0.397]
		0.028*	0.508*	0.437*	0.139*
500	-1.611	-0.237	-0.663	2.754	0.426
	(1.583)	(2.603)	(3.193)	(5.193)	(0.458)
	[1.231]	[1.902]	[2.845]	[4.383]	[0.392]
		0.046*	0.704*	0.658*	0.457*
1000	-1.755	-0.148	-0.617	1.274	0.432
	(1.348)	(1.871)	(2.836)	(4.088)	(0.434)
	[1.175]	[1.855]	[2.800]	[3.725]	[0.371]
		0.043*	0.759*	0.731*	0.587*
1500	-1.446	-0.124	-0.621	1.425	0.487
	(1.281)	(1.891)	(2.800)	(3.632)	(0.433)
	[1.143]	[1.878]	[2.778]	[3.574]	[0.378]
		0.051*	0.799*	0.755*	0.632*
Pourcentage d'immunes = 75%					
100	-1.665	-0.284	-1.095	1.434	0.178
	(4.897)	(4.546)	(5.127)	(6.948)	(6.963)
	[2.514]	[2.293]	[2.981]	[5.974]	[2.119]
		0.037*	0.515*	0.083*	0.237*
500	-1.746	-0.484	-1.032	0.751	0.446
	(3.847)	(3.748)	(4.757)	(6.987)	(6.769)
	[2.091]	[2.028]	[2.936]	[5.969]	[1.901]
		0.041*	0.696*	0.305*	0.596*
1000	-1.520	-0.261	-0.745	0.252	0.321
	(2.883)	(4.496)	(2.679)	(5.839)	(3.468)
	[1.857]	[2.110]	[2.655]	[5.662]	[1.194]
		0.038*	0.788*	0.481*	0.696*
1500	-1.510	-0.120	-0.757	0.132	0.315
	(2.550)	(1.902)	(2.671)	(6.016)	(3.658)
	[1.791]	[1.887]	[2.649]	[5.659]	[1.427]
		0.041*	0.801*	0.567*	0.727*

TAB. B.5 – Résultats des modèles $\mathscr{M}_1$ et $\mathscr{M}_3$ s'il y a absence d'immunité.

n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$
$\mathscr{M}_1 : \beta = (-1.7, -2, -3.4, 5, .3)$					
100	-1.887	-2.474	-4.126	5.747	0.349
	(1.197)	(1.098)	(1.456)	(1.851)	(1.825)
	[0.938]	[0.750]	[1.031]	[1.437]	[1.418]
		0.985*	0.996*	0.953*	0.033*
500	-1.749	-2.072	-3.537	5.186	0.317
	(0.472)	(0.282)	(0.442)	(0.727)	(0.586)
	[0.366]	[0.217]	[0.332]	[0.558]	[0.469]
		1*	1*	1*	0.067*
1000	-1.724	-2.027	-3.449	5.066	0.302
	(0.318)	(0.188)	(0.275)	(0.456)	(0.436)
	[0.253]	[0.149]	[0.216]	[0.362]	[0.348]
		1*	1*	1*	0.121*
1500	-1.715	-2.020	-3.437	5.053	0.298
	(0.253)	(0.152)	(0.218)	(0.372)	(0.340)
	[0.199]	[0.121]	[0.169]	[0.297]	[0.273]
		1*	1*	1*	0.145*
$\mathscr{M}_3 : \beta = (-1.7, -2.8, 0, -.7, 1.1)$					
100	-1.881	-2.937	0.002	-0.753	1.293
	(0.830)	(0.550)	(0.385)	(0.969)	(1.057)
	[0.659]	[0.453]	[0.297]	[0.746]	[0.803]
		1*	$0.044^{\dagger}$	0.137*	0.244*
500	-1.740	-2.875	-0.002	-0.718	1.118
	(0.348)	(0.299)	(0.144)	(0.367)	(0.397)
	[0.272]	[0.228]	[0.115]	[0.289]	[0.314]
		1*	$0.047^{\dagger}$	0.540*	0.826*
1000	-1.728	-2.823	-0.001	-0.697	1.116
	(0.237)	(0.190)	(0.095)	(0.243)	(0.267)
	[0.188]	[0.151]	[0.078]	[0.197]	[0.212]
		1*	$0.054^{\dagger}$	0.809*	0.989*
1500	-1.711	-2.823	-0.001	-0.702	1.104
	(0.195)	(0.159)	(0.081)	(0.202)	(0.229)
	[0.154]	[0.125]	[0.065]	[0.162]	[0.184]
		1*	$0.047^{\dagger}$	0.935*	0.998*

Comme nous l'avons dit antérieurement, dans le cadre de la régression logistique, il est naturellement intéressant d'estimer la probabilité de survenue de l'événement d'intérêt $p(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}, S = 1)$ pour un $\mathbf{x}$ donné. Dans le tableau B.7 ci dessous, nous étudions les propriétés numériques de l'estimateur $\widehat{p}_n(\mathbf{x})$ (nous nous restreignons à une seule valeur de $p(\mathbf{x})$ pour chacun des modèles $\mathscr{M}_1$ et $\mathscr{M}_2$). Pour chaque configuration des paramètres de simulation, nous obtenons les moyennes (sur les N échantillons) des estimations des probabilités d'infection $p(\mathbf{x}) = 0.250$ (modèle $\mathscr{M}_1$) et $p(\mathbf{x}) = 0.343$ (modèle $\mathscr{M}_2$). Nous obtenons également la racine carrée de l'erreur quadratique moyenne et l'erreur absolue moyenne correspondantes. Nous déterminons aussi les probabilités de couverture empiriques des intervalles de confiance de niveau asymptotique 95% pour $p(\mathbf{x})$, et les longueurs moyennes de ces intervalles.

TAB. B.7 – Probabilités estimées $p(\mathbf{x})$ pour les modèles $\mathscr{M}_1$ ($p(\mathbf{x}) = 0.250$) et $\mathscr{M}_2$ ($p(\mathbf{x}) = 0.343$).

	0% d'immune		25% d'immune		50% d'immune		75% d'immune	
n	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_1$	$\mathscr{M}_2$
100	0.229	0.335	0.267	0.363	0.264	0.353	0.266	0.362
	(0.118)	(0.117)	(0.191)	(0.164)	(0.264)	(0.215)	(0.360)	(0.367)
	[0.097]	[0.095]	[0.149]	[0.132]	[0.202]	[0.174]	[0.293]	[0.318]
	0.964⋆	0.959⋆	0.964⋆	0.978⋆	0.987⋆	0.949⋆	0.951⋆	0.831⋆
	0.407∓	0.424∓	0.596∓	0.523∓	0.719∓	0.627∓	0.859∓	0.756∓
500	0.246	0.343	0.261	0.356	0.263	0.358	0.255	0.354
	(0.046)	(0.049)	(0.079)	(0.069)	(0.110)	(0.084)	(0.188)	(0.201)
	[0.037]	[0.039]	[0.056]	[0.054]	[0.083]	[0.068]	[0.150]	[0.160]
	0.957⋆	0.959⋆	0.930⋆	0.898⋆	0.921⋆	0.943⋆	0.891⋆	0.701⋆
	0.182∓	0.192∓	0.216∓	0.228∓	0.343∓	0.319∓	0.572∓	0.482∓
1000	0.247	0.342	0.255	0.352	0.258	0.351	0.254	0.350
	(0.034)	(0.035)	(0.051)	(0.052)	(0.071)	(0.061)	(0.137)	(0.134)
	[0.027]	[0.028]	[0.037]	[0.039]	[0.055]	[0.049]	[0.106]	[0.106]
	0.948⋆	0.954⋆	0.887⋆	0.875⋆	0.931⋆	0.932⋆	0.894⋆	0.651⋆
	0.128∓	0.136∓	0.149∓	0.164∓	0.248∓	0.226∓	0.436∓	0.359∓
1500	0.249	0.343	0.252	0.348	0.253	0.352	0.251	0.348
	(0.028)	(0.028)	(0.037)	(0.038)	(0.062)	(0.047)	(0.106)	(0.108)
	[0.022]	[0.023]	[0.028]	[0.028]	[0.044]	[0.038]	[0.085]	[0.084]
	0.945⋆	0.958⋆	0.907⋆	0.877⋆	0.932⋆	0.929⋆	0.904⋆	0.641⋆
	0.105∓	0.112∓	0.119∓	0.133∓	0.199∓	0.185∓	0.373∓	0.301∓

<u>Note</u> : $(\cdot)$: racine carrée de l'erreur quadratique moyenne. $[\cdot]$: erreur absolue moyenne. $\star$: probabilité de couverture empirique. $\mp$: longueur moyenne des intervalles de confiance. Pour chaque pourcentage d'immunes, les pourcentages d'infectés parmi les susceptibles sont respectivement 30% ($\mathscr{M}_1$) et 70% ($\mathscr{M}_2$). Tous les résultats sont basés sur 1500 réplications.

Nous avons également évalué la qualité de l'approximation normale de la distribution asymptotique de $\widehat{\beta}_n$. Pour chaque configuration des paramètres de simulation, nous obtenons les histogrammes des $\widehat{\beta}_{n,l}^{(j)}$ $(j = 1, \ldots, N)$, avec les Q-Q plots correspondants. Les résultats sont obtenus pour β_2 et β_3 dans le modèle $\mathscr{M}_1$. Pour les autres paramètres du modèle, nous obtenons des résultats similaires.

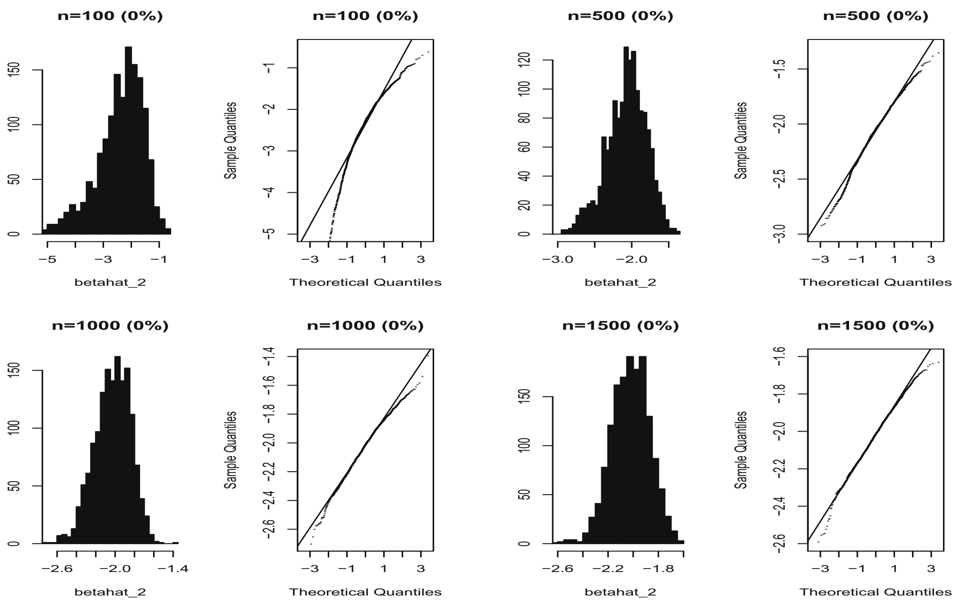

FIG. B.1 – Histogrammes et Q-Q plots pour $\widehat{\beta}_{2,n}$ dans le modèle $\mathscr{M}_1$, sans individus immunes dans l'échantillon (le pourcentage d'immunes est indiqué entre parenthèses). n est la taille d'échantillon. Tous les résultats sont basés sur 1500 jeux de données simulés.

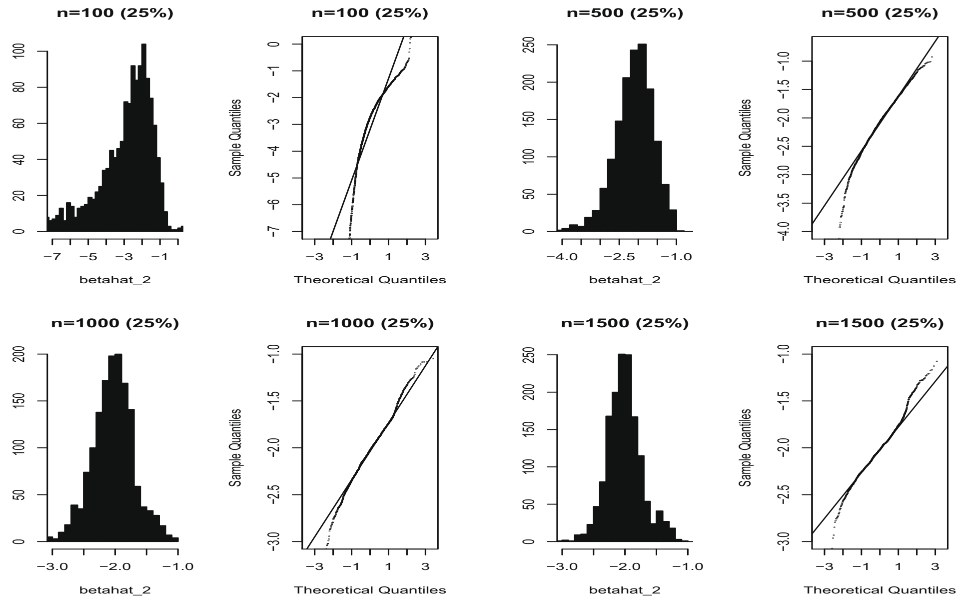

FIG. B.2 – Histogrammes et Q-Q plots pour $\widehat{\beta}_{2,n}$ dans le modèle $\mathscr{M}_1$, avec 25% d'immunes.

Finalement, nous étudions la qualité de l'approximation normale de la distribution de $\widehat{p}_n(\mathbf{x})$. Les histogrammes et Q-Q plots de $\widehat{p}_n^{(j)}(\mathbf{x})$ $(j = 1, \ldots, N)$ sont obtenus pour une valeur de $p(\mathbf{x}) = 0.250$ dans le modèle $\mathscr{M}_1$, et sont donnés par les figures B.9 à B.12.

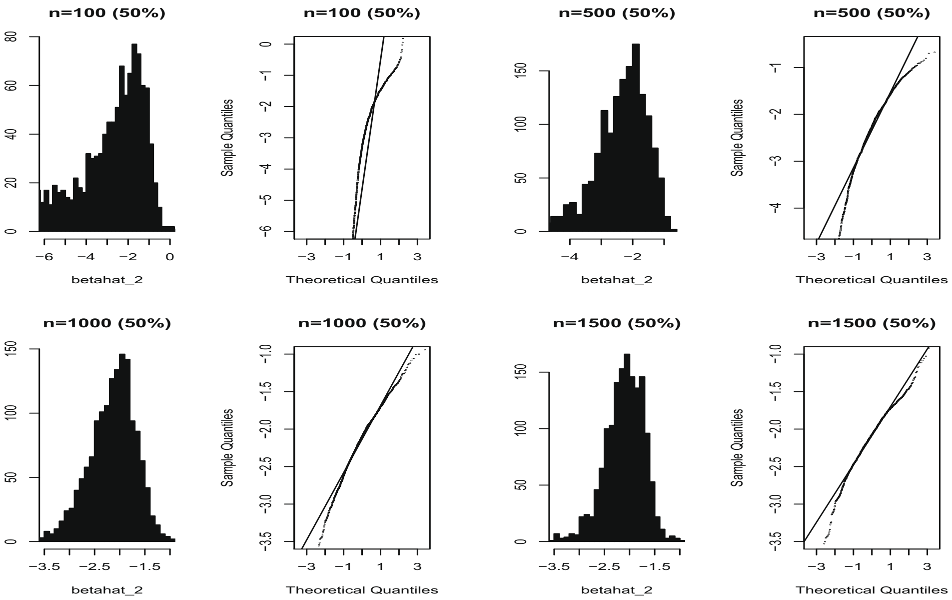

FIG. B.3 – Histogrammes et Q-Q plots pour $\widehat{\beta}_{2,n}$ dans le modèle $\mathscr{M}_1$, avec 50% d'immunes.

B.2 Résultats de simulations du chapitre 5

Dans cette partie, nous présentons les résultats de simulations du modèle $\mathscr{M}_2$ décrit dans le chapitre 5.

Les résultats pour 50% d'immunes (n=100, 500, 1000, 1500) sont en cours et seront aussitôt rajoutés dans la version vraiment définitive du manuscrit.

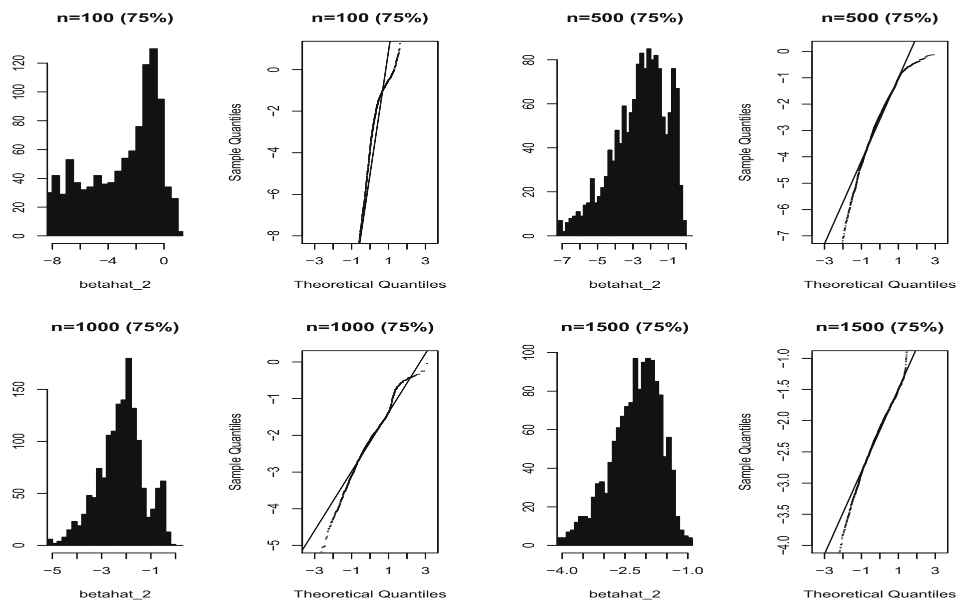

FIG. B.4 – Histogrammes et Q-Q plots pour $\widehat{\beta}_{2,n}$ dans le modèle $\mathscr{M}_1$, avec 75% d'immunes.

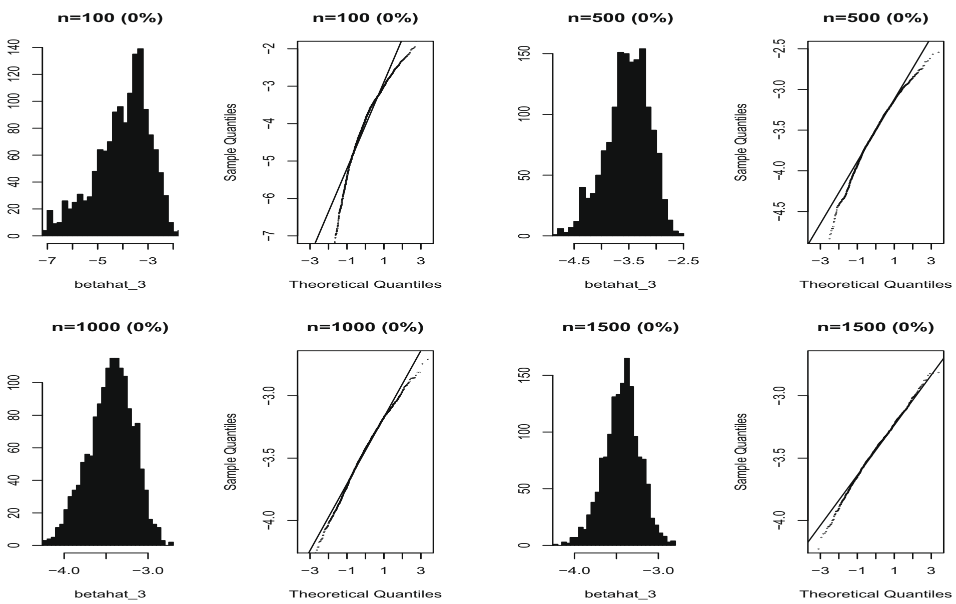

FIG. B.5 – Histogrammes et Q-Q plots pour $\widehat{\beta}_{3,n}$ dans modèle $\mathscr{M}_1$, sans individus immunes dans l'échantillon.

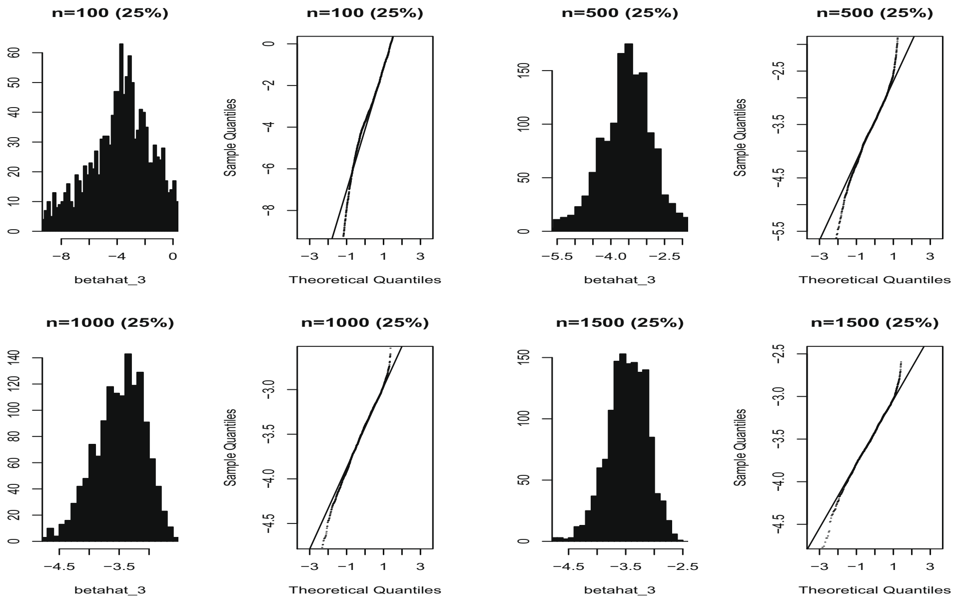

FIG. B.6 – Histogrammes et Q-Q plots pour $\widehat{\beta}_{3,n}$ dans modèle $\mathscr{M}_1$, avec 25% d'immunes.

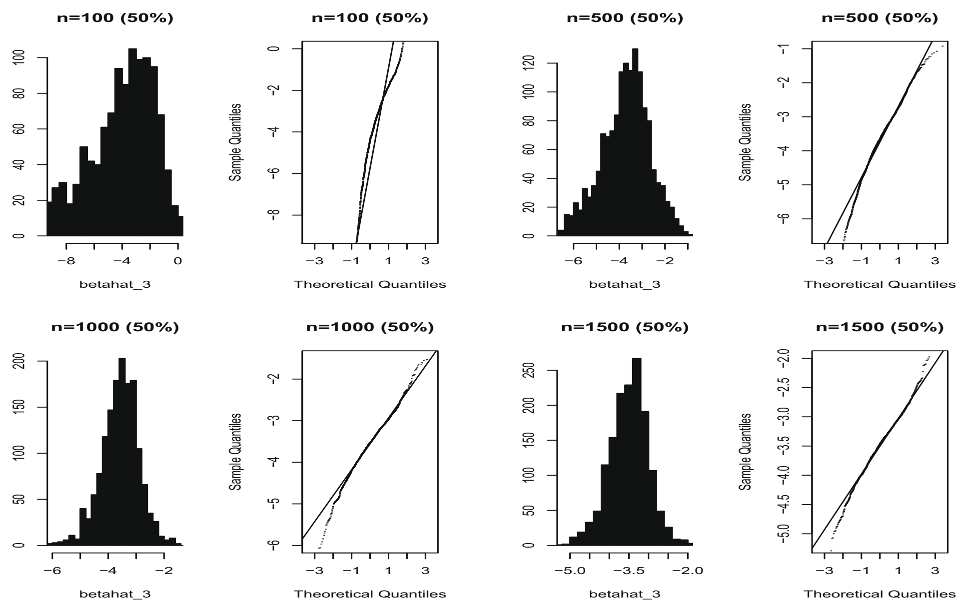

FIG. B.7 – Histogrammes et Q-Q plots pour $\widehat{\beta}_{3,n}$ dans modèle $\mathscr{M}_1$, avec 50% d'immunes.

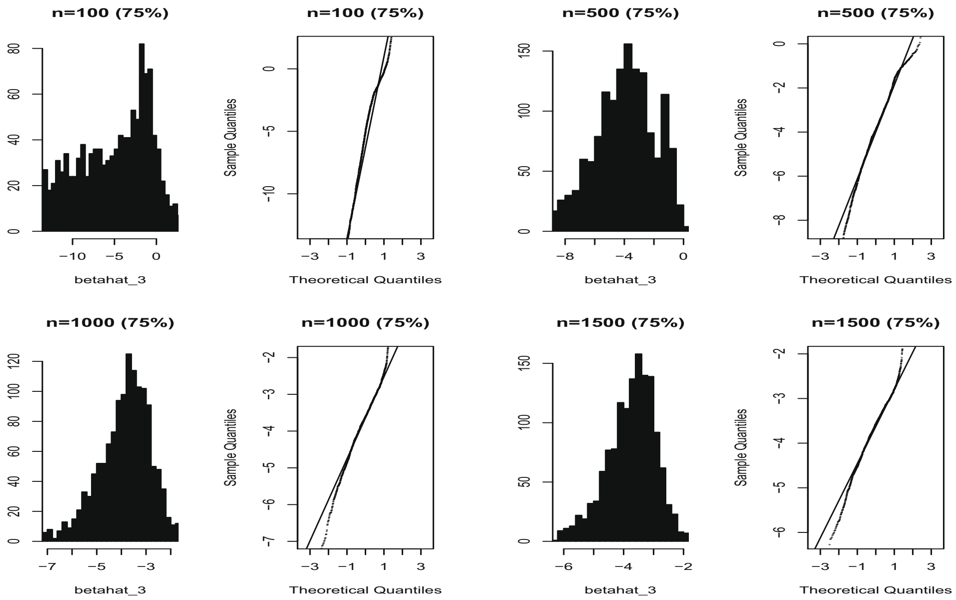

FIG. B.8 – Histogrammes et Q-Q plots pour $\widehat{\beta}_{3,n}$ dans le modèle $\mathscr{M}_1$, avec 75% d'immunes.

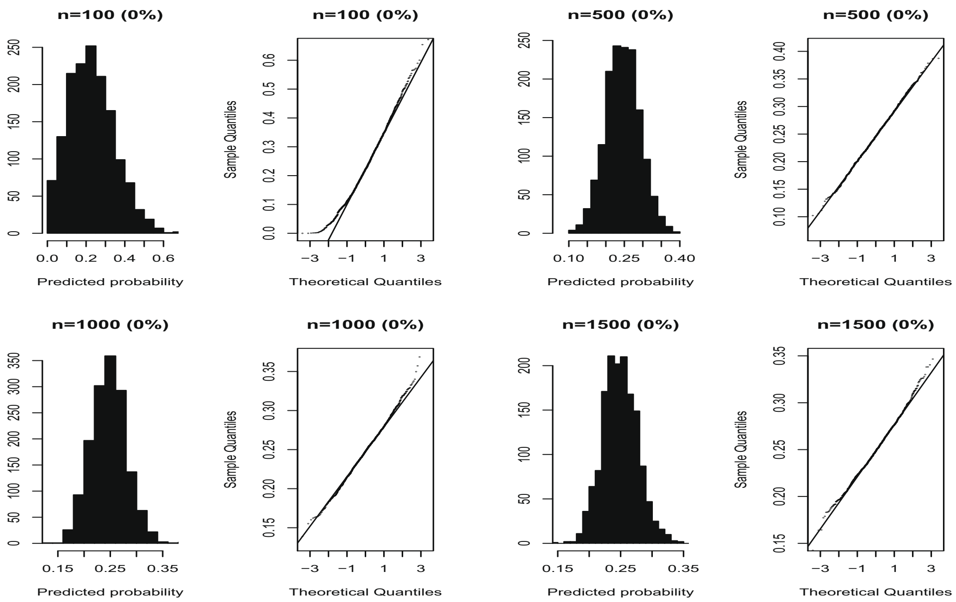

FIG. B.9 – Histogrammes et Q-Q plots pour $\widehat{p}_n(\mathbf{x})$ dans le modèle $\mathscr{M}_1$, sans individus immunes dans l'échantillon.

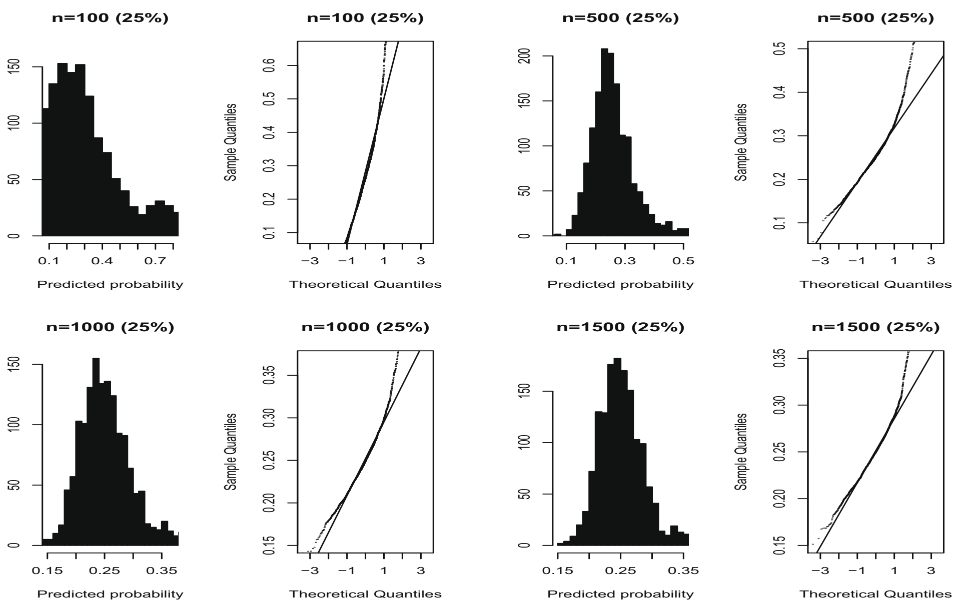

FIG. B.10 – Histogrammes et Q-Q plots pour $\widehat{p}_n(\mathbf{x})$ dans le modèle $\mathscr{M}_1$, avec 25% d'immunes.

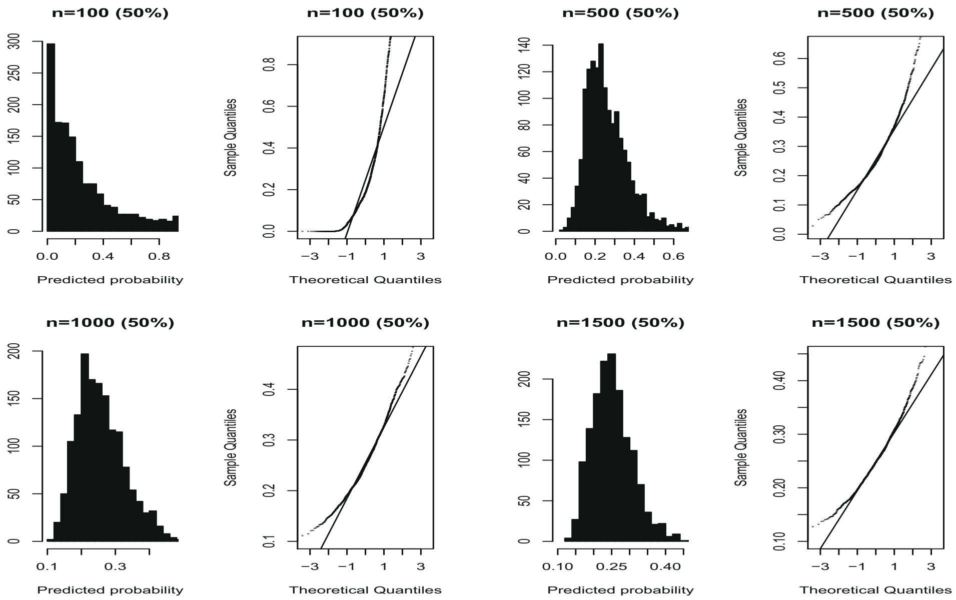

FIG. B.11 – Histogrammes et Q-Q plots pour $\widehat{p}_n(\mathbf{x})$ dans le modèle $\mathscr{M}_1$, avec 50% d'immunes.

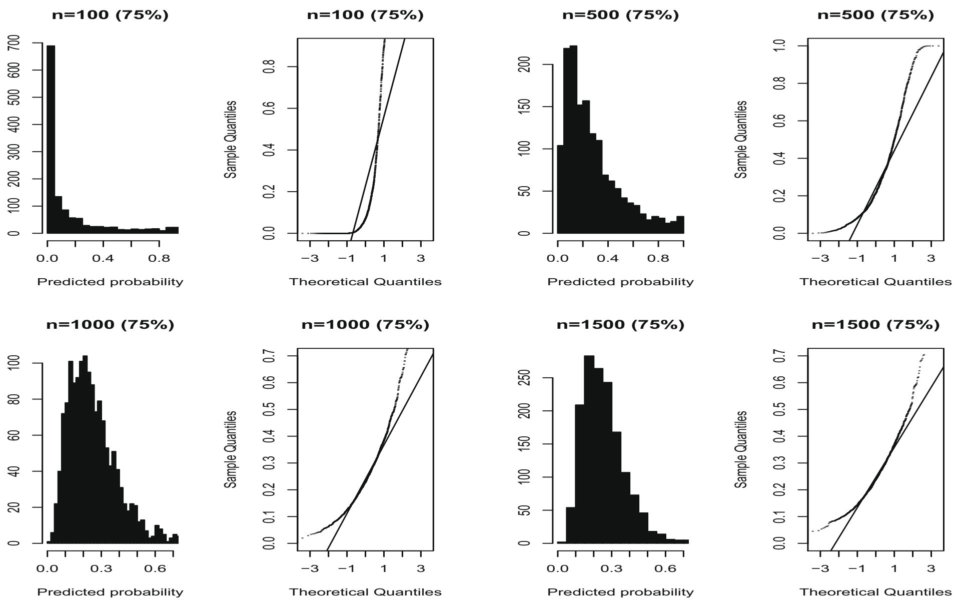

FIG. B.12 – Histogrammes et Q-Q plots pour $\widehat{p}_n(\mathbf{x})$ dans le modèle $\mathscr{M}_1$, avec 75% d'immunes.

TAB. B.8 – Modèle $\mathscr{M}_2$: $\beta = (.5, 1, -1)$ avec 25% d'immunes.

		Méthode 1		Méthode 2		Méthode 3	
$1-\alpha$	n	Couverture^a	Largeur	Couverture^a	Largeur	Couverture^a	Largeur
			Pourcentage d'immunes $= 25\%, \theta = (1.6, -1)$				
0.99	500	0.831	0.362^+	0.815	0.344^+	0.992	0.692^+
			$0.261^\dagger$		$0.244^\dagger$		$0.474^\dagger$
			$0.343^\star$		$0.326^\star$		$0.816^\star$
			$0.468^\mp$		$0.447^\mp$		$0.889^\mp$
	1000	0.775	0.219^+	0.762	0.209^+	0.982	0.526^+
			$0.139^\dagger$		$0.131^\dagger$		$0.282^\dagger$
			$0.214^\star$		$0.204^\star$		$0.635^\star$
			$0.292^\mp$		$0.280^\mp$		$0.729^\mp$
	1500	0.735	0.171^+	0.725	0.164^+	0.983	0.350^+
			$0.101^\dagger$		$0.095^\dagger$		$0.180^\dagger$
			$0.169^\star$		$0.161^\star$		$0.368^\star$
			$0.232^\mp$		$0.223^\mp$		$0.507^\mp$
0.95	500	0.734	0.274^+	0.719	0.263^+	0.994	0.598^+
			$0.179^\dagger$		$0.169^\dagger$		$0.366^\dagger$
			$0.259^\star$		$0.248^\star$		$0.712^\star$
			$0.366^\mp$		$0.353^\mp$		$0.800^\mp$
	1000	0.662	0.172^+	0.644	0.166^+	0.973	0.443^+
			$0.100^\dagger$		$0.095^\dagger$		$0.205^\dagger$
			$0.167^\star$		$0.160^\star$		$0.528^\star$
			$0.235^\mp$		$0.228^\mp$		$0.639^\mp$
	1500	0.638	0.136^+	0.621	0.131^+	0.964	0.287^+
			$0.073^\dagger$		$0.069^\dagger$		$0.135^\dagger$
			$0.133^\star$		$0.128^\star$		$0.289^\star$
			$0.189^\mp$		$0.183^\mp$		$0.427^\mp$
0.90	500	0.631	0.239^+	0.619	0.233^+	0.983	0.543^+
			$0.150^\dagger$		$0.145^\dagger$		$0.307^\dagger$
			$0.222^\star$		$0.216^\star$		$0.644^\star$
			$0.323^\mp$		$0.314^\mp$		$0.746^\mp$
	1000	0.588	0.151^+	0.572	0.147^+	0.950	0.403^+
			$0.083^\dagger$		$0.080^\dagger$		$0.172^\dagger$
			$0.145^\star$		$0.141^\star$		$0.472^\star$
			$0.209^\mp$		$0.204^\mp$		$0.589^\mp$
	1500	0.570	0.120^+	0.561	0.116^+	0.941	0.255^+
			$0.062^\dagger$		$0.060^\dagger$		$0.109^\dagger$
			$0.117^\star$		$0.114^\star$		$0.251^\star$
			$0.168^\mp$		$0.164^\mp$		$0.385^\mp$

Note : $^+$: moyenne, † : 1er quartile, * : médiane, $^\mp$: 3ème quartile.

TAB. B.9 – Modèle $\mathscr{M}_2$: $\beta = (.5, 1, -1)$ avec 50% d'immunes

		Method 1		Method 2		Method 3	
$1-\alpha$	n	Couverture[a]	Largeur	Couverture[a]	Largeur	Couverture[a]	Largeur
				Pourcentage d'immunes $= 50\%, \theta = (-1, 2)$			
0.99	500	0.690	0.479^+	0.672	0.457^+	0.996	0.830^+
			$0.339^\dagger$		$0.320^\dagger$		$0.696^\dagger$
			$0.435^\star$		$0.412^\star$		$0.927^\star$
			$0.636^\mp$		$0.611^\mp$		$0.974^\mp$
	1000	0.685	0.301^+	0.672	0.287^+	0.995	0.664^+
			$0.201^\dagger$		$0.188^\dagger$		$0.391^\dagger$
			$0.276^\star$		$0.263^\star$		$0.812^\star$
			$0.405^\mp$		$0.387^\mp$		$0.884^\mp$
	1500	0.679	0.226^+	0.665	0.216^+	0.987	0.552^+
			$0.143^\dagger$		$0.134^\dagger$		$0.273^\dagger$
			$0.213^\star$		$0.206^\star$		$0.677^\star$
			$0.306^\mp$		$294^\mp$		$0.774^\mp$
0.95	500	0.574	0.362^+	0.555	0.348^+	0.986	0.767^+
			$0.231^\dagger$		$0.219^\dagger$		$0.599^\dagger$
			$0.317^\star$		$0.303^\star$		$0.871^\star$
			$0.501^\mp$		$0.482^\mp$		$0.934^\mp$
	1000	0.607	0.226^+	0.589	0.218^+	0.984	0.581^+
			$0.136^\dagger$		$0.129^\dagger$		$0.301^\dagger$
			$0.207^\star$		$0.198^\star$		$0.714^\star$
			$0.312^\mp$		$0.301^\mp$		$0.805^\mp$
	1500	0.618	0.174^+	0.609	0.168^+	0.975	0.465^+
			$0.098^\dagger$		$0.093^\dagger$		$0.205^\dagger$
			$0.163^\star$		$0.157^\star$		$0.552^\star$
			$0.243^\mp$		$0.235^\mp$		$0.675^\mp$
0.90	500	0.533	0.311^+	0.525	0.302^+	0.986	0.728^+
			$0.189^\dagger$		$0.182^\dagger$		$0.539^\dagger$
			$0.266^\star$		$0.258^\star$		$0.835^\star$
			$0.437^\mp$		$0.426^\mp$		$0.911^\mp$
	1000	0.511	0.198^+	0.504	0.193^+	0.977	0.524^+
			$0.113^\dagger$		$0.109^\dagger$		$0.251^\dagger$
			$0.179^\star$		$0.173^\star$		$0.632^\star$
			$0.277^\mp$		$0.269^\mp$		$0.745^\mp$
	1500	0.517	0.153^+	0.510	0.149^+	0.960	0.429^+
			$0.082^\dagger$		$0.079^\dagger$		$0.174^\dagger$
			$0.141^\star$		$0.137^\star$		$0.503^\star$
			$0.215^\mp$		$0.210^\mp$		$0.637^\mp$

Note : $^+$: moyenne, † : 1er quartile, * : médiane, $^\mp$: 3ème quartile.

TAB. B.10 – Analyse "naïve" : Modèle $\mathscr{M}_2$: $\beta = (.5, 1, -1)$ avec 25% et 50% d'immunes.

		Method 1		Method 2		Method 3	
$1-\alpha$	n	Couverture^a	Largeur	Couverture^a	Largeur	Couverture^a	Largeur
		Pourcentage d'immunes $= 25\%, \theta = (1.6, -1)$					
0.99	500	0.001	0.201	0	0.193	0.001	0.201
	1000	0	0.134	0	0.129	0	0.133
	1500	0	0.109	0	0.105	0	0.108
0.95	500	0.162	0	0	0.157	0	0.163
	1000	0	0.110	0	0.107	0	0.110
	1500	0	0.089	0	0.086	0	0.089
0.90	500	0	0.145	0	0.142	0	0.145
	1000	0	0.098	0	0.096	0	0.098
	1500	0	0.079	0	0.078	0	0.079
		Pourcentage d'immunes $= 50\%, \theta = (-1, 2)$					
0.99	500	0	0.277	0	0.268	0	0.271
	1000	0	0.196	0	0.189	0	0.190
	1500	0	0.158	0	0.153	0	0.153
0.95	500	0	0.229	0	0.222	0	0.224
	1000	0	0.160	0	0.155	0	0.156
	1500	0	0.130	0	0.127	0	0.127
0.90	500	0	0.205	0	0.201	0	0.201
	1000	0	0.143	0	0.140	0	0.139
	1500	0	0.116	0	0.113	0	0.113

Note : a : probabilité de couverture.

Bibliographie

[Adler & Taylor 2007] R. J. Adler et J.E. Taylor. Random fields and geometry. Springer Monographs in Mathematics, 2007. 69

[Agarwal *et al.* 2002] D. Agarwal, A. Gelfand et S. Citron-Pousty. *Zero-inflated models with application to spatial count data.* Environ. Ecol. Statist., vol. 9 (4), pages 341–355, 2002. 18

[Agresti 2002] A. Agresti. Categorical data analysis. John Wiley & Sons, Inc., 2002. 27

[Al-Saidy *et al.* 2003] O.M. Al-Saidy, W.W. Piegorsch, R.W. West et D.K. Nitcheva. *Confidence bands for low-dose risk estimation with quantal response data.* Biometrics, vol. 59, pages 1056–1062, 2003. 66

[Alcon *et al.* 2002] S. Alcon, A. Talarmin, M. Debruyne, V. Falconar A. Deubel et al. *Enzyme-linked immunosorbent assay specific to Dengue virus type 1 non-structural protein NS1 reveals circulation of the antigen in the blood during the acute phase of disease in patients experiencing primary or secondary infections.* J. Clin. Microbiol., vol. 40, pages 376–381, 2002. 94

[Aldo *et al.* 2011] M. G. Aldo, M. H. Elizabeth, M.M.O. Edwin et H.L. Víctor. *On estimation and influence diagnostics for zero-inflated negative binomial regression models.* Comput. Statist. Data Anal., vol. 55, pages 1304–1318, 2011. 20

[Aminot & Damon 2002] I. Aminot et M.N. Damon. *Régression logistique : intérêt* *dans l'analyse de données relatives aux pratiques médicales*. Revue Médicale de l'Assurance Maladie, vol. 33, pages 137–143, 2002. 7

[Azaïs & Bardet 2006] J.M. Azaïs et J.M. Bardet. Le modèle linéaire par l'exemple : Régression, analyse de la variance et plans d'expérience illustrés avec r, sas et splus. Dunod, 2006. 5

[Azaïs *et al.* 2010] J.M. Azaïs, S. Bercu, J. Fort, A. Lagnoux et P. Lé. *Simultaneous confidence bands in curve prediction applied to load curves*. J. Roy. Statist. Soc., vol. 59, pages 889–904, 2010. 66

[Balsitis *et al.* 2009] S.J. Balsitis, J. Coloma, G. Castro, A. Alava, D. Flores, et al. *(2009) Tropism of dengue virus in mice and humans defined by viral nonstructural protein 3-specific immunostaining*. Am. J. Trop. Med. Hyg., vol. 80, pages 416–424, 2009. 109

[Beckett *et al.* 2005] C.G. Beckett, H. Kosasih, I. Faisal, T.R. Nurhayati et al. *Early detection of dengue infections using cluster sampling around index cases*. Am. J. Trop. Med. Hyg., vol. 72, pages 777–782, 2005. 94, 106

[Bhargava & Spurrier 2004] P. Bhargava et J.D. Spurrier. *Exact confidence bands for comparing two regression lines with a control regression line on a fixed interval.* Biometrical J., vol. 46, pages 720–730, 2004. 66

[Blacksell *et al.* 2008] S.D. Blacksell, M.P. Jr. Mammen, S. Thongpaseuth, R.V. Gibbons, R.G. Jarman et al. *Evaluation of the Panbio dengue virus non-structural 1 antigen detection and immunoglobulin M antibody enzyme-linked immunosorbent assays for the diagnosis of acute dengue infections in Laos*. Diagn. Microbiol. Infect. Dis., vol. 60, pages 43–49, 2008. 94

[Blischke 1978] W.R. Blischke. *Mixtures of distributions.* Intertional Encyclopedia of Statistics, vol. Vol.1, W.H. Kruskal and J.M. Tanur (Eds.). New York : The Free Press, pages 174–180, 1978. 26

[Bohning *et al.* 1999] D. Bohning, E. Dietz, P. Schlattmann, L. Mendonca et U. Kirchner. *The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology.* J. Amer. Stat. Assoc., pages 195–209, 1999. 18

[Bohrer 1973] R. Bohrer. *A multivariate t probability integral.* Biometrika, vol. 60, pages 647–654, 1973. 66

[Brand *et al.* 1973] R. Brand, D. Pinnock et K. Jackson. *Large sample confidence bands for the logistic response curve and its inverse.* Amer. Statist., vol. 27, pages 157–160, 1973. 72, 73

[Capeding *et al.* 2010] R.Z. Capeding, J.D. Brion, M.M. Caponpon, R.V. Gibbons, R.G. Jarman et al. *The incidence, characteristics, and presentation of dengue virus infections during infancy.* Am. J. Trop. Med. Hyg., vol. 82, pages 330–336, 2010. 110

[Carroll *et al.* 1995] R. J. Carroll, D. Ruppert et L. A. Stefanski. Measurement error in nonlinear models. Chapman and Hall, New York, 1995. 23

[Casella & Strawderman 1980] G. Casella et W.E. Strawderman. *Confidence bands for linear-regression with restricted predictor variables.* J. Amer. Statist. Assoc., vol. 75, pages 862–868, 1980. 66

[Chau *et al.* 2009] T.N. Chau, N.T. Hieu, K.L. Anders, M. Wolbers, le B. Lien et al. *Dengue virus infections and maternal antibody decay in a prospective birth* *cohort study of Vietnamese infants*. J. Infect. Dis., vol. 200, pages 1893–1900, 2009. 110

[Chuansumrit *et al.* 2008] A. Chuansumrit, W. Chaiyaratana, V. Pongthanapisith, K Tangnararatchakit, S. Lertwongrath et al. *The use of dengue nonstructural protein 1 antigen for the early diagnosis during the febrile stage in patients with dengue infection*. Pediatr. Infect. Dis. J., vol. 27, pages 43–48, 2008. 108

[Claeskens & Van Keilegom 2003] G. Claeskens et I. Van Keilegom. *Bootstrap confidence bands for regression curves and their derivatives*. Ann. Statist., vol. 31, pages 1852–1884, 2003. 78

[Consul & Famoye 1992] P. C. Consul et F. Famoye. *Generalized Poisson regression model*. Comm. Statist. Theory Methods, vol. 21, pages 89–109, 1992. 18

[Cowling *et al.* 1996] A. Cowling, P. Hall et Phillips M.J. *Bootstrap confidence regions for the intensity of a Poisson point process*. J. Amer. Statist. Assoc., vol. 91(436), pages 1516–1524, 1996. 78

[Czado & Santner 1992] C. Czado et T. J. Santner. *The effect of link misspecification on binary regression inference*. J. Statist. Plann. Inference, vol. 33, pages 213–231, 1992. 58, 126

[Czado *et al.* 2007] V. Czado C.and Erhardt, A. Min et S. Wagner. *Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates*. Statist. Model., vol. 7(2), pages 125–153, 2007. 21, 22

[De Benedictis *et al.* 2003] J. De Benedictis, E. Chow-Shaffer, A. Costero, G.G. Clark, J.D. Edman et al. *Identification of the people from whom engor-*

ged Aedes aegypti took blood meals in Florida, Puerto Rico, using polymerase chain reaction-based DNA profiling. Am. J. Trop. Med. Hyg., vol. 68, pages 437–446, 2003. 107

[Dempster *et al.* 1977] A. Dempster, N. Laird et D. Rubin. *Maximum likelihood from incomplete data via the em algorithm (with discussion)*. J. Roy. Statist. Soc. Ser. B, vol. 39, pages 1–38, 1977. 23

[Dietz & Bohning 2000] E. Dietz et D. Bohning. *On estimation of the Poisson parameter in zero-modified Poisson models*. Comput. Statist. Data Anal., vol. 34, pages 441–459, 2000. 34

[Diop *et al.* 2011] A. Diop, A. Diop et J.-F. Dupuy. *Maximum likelihood estimation in the logistic regression model with a cure fraction*. Electronic J. Statist., vol. 5, pages 460–483, 2011. 67, 68

[Dudek *et al.* 2008] A. Dudek, M. Goéwin et J. Leśkow. *Simultaneous Confidence Bands for the Integrated Hazard Function*. Comput. Statist., vol. 23(1), pages 41–62, 2008. 78

[Dussart *et al.* 2006] P. Dussart, B. Labeau, G. Lagathu, P. Louis, M.R. Nunes et al. *Evaluation of an enzyme immunoassay for detection of dengue virus NS1 antigen in human serum*. Clin. Vac. Im., vol. 13, pages 1185–1189, 2006. 94, 108

[Dussart *et al.* 2008] P. Dussart, L. Petit, B. Labeau, L. Bremand, A. Leduc et al. *Evaluation of two new commercial tests for the diagnosis of acute dengue virus infection using NS1 antigen detection in human serum*. PLoS. Negl. Trop. Dis., vol. 2, page e230, 2008. 94, 108

[Dussart *et al.* 2012] P. Dussart, L. Baril, L. Petit, L. Beniguel, L. C. Quang, S. Ly, R. do Socorro Azevedo, J.-B. Meynard, S. Vong, L. Chartier, A. Diop, O. Sivuth, V. Duong, C. M. Thang, M. Jacobs, A. Sakuntabhai, M. R. Texeira Nunes, V. T. Que Huong, P. Buchy et P. F. Vasconcelos. *Study of dengue cases and the members of their households : a familial cluster analysis in the multinational DENFRAME project*. PLoS. Negl. Trop. Dis., vol. 6(1) : e1482, 2012. 120

[Efron & DiCiccio 1996] B. Efron et T.J. DiCiccio. *Bootstrap Confidence Intervals*. Statist. Sci., vol. 11(3), pages 189–228, 1996. 78

[Efron & Tibshirani 1993] B. Efron et R. Tibshirani. An introduction to the bootstrap. Chapman & Hall, 1993. 131

[Efron 1982] B. Efron. *The jackknife, the bootstrap, and other resampling plans*. In Society for Industrial and Applied Mathematics. 1982. 131

[Eicker 1966] F. Eicker. *A multivariate central limit theorem for random linear vector forms*. Ann. Math. Statist., vol. 37, pages 1825–1828, 1966. 49

[Endy *et al.* 2002] T.P. Endy, S. Chunsuttiwat, A. Nisalak, DH. Libraty, S Green et et al. *Epidemiology of inapparent and symptomatic acute dengue virus infection : a prospective study of primary school children in Kamphaeng Phet*. Thailand. Am. J. Epidemiol, vol. 156, pages 40–51, 2002. 93, 105

[Endy *et al.* 2011] T.P. Endy, K.B. Anderson, A. Nisalak, I.K. Yoon, S. Green et al. *Determinants of inapparent and symptomatic dengue infection in a prospective study of primary school children in Kamphaeng Phet, Thailand*. PloS. Neg. Top. Dis., vol. 5, page e975, 2011. 106

[Fahrmeir & Kaufmann 1985] L. Fahrmeir et H. Kaufmann. *Consistency and Asymptotic normality of the Maximum Likelihood Estimator in Generalized Linear Model.* Ann. Statist., vol. 13, pages 342–368, 1985. 7, 12, 33

[Fahrmeir & Tutz 2001] L. Fahrmeir et G. Tutz. Multivariate statistical modelling based on generalized linear models. Springer Series in Stat, 2001. 15

[Famoye & Singh 2006] F. Famoye et K. P. Singh. *Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data.* J. Data Sci., vol. 4, pages 117–130, 2006. 18, 34

[Fang *et al.* 2005] H.-B. Fang, G. Li et J. Sun. *Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model.* Scand. J. Statist., vol. 32, pages 59–75, 2005. 37

[Farewell & Sprott 1988] V.T. Farewell et d. Sprott. *The use of a mixture model in the analysis of count data.* Biometrics, vol. 44, pages 1191–1194, 1988. 27

[Follmann & Lambert 1991a] D.A. Follmann et D. Lambert. *Generalysing logistic regression by nonparametric mixing.* J. Amer. Stat. Assoc., vol. 84, pages 295–300, 1989,1991. 27

[Follmann & Lambert 1991b] D.A. Follmann et D. Lambert. *Identifiability for nonparametric mixtures of logistic regressions.* J. Statist. Plann. Inference, vol. 27, pages 375–381, 1991. 40

[Fong & Yip 1993] D.Y.T. Fong et P. Yip. *An EM algorithm for a mixture model of count data.* Statist. Probab. Lett., vol. 17, pages 53–60, 1993. 18

[Fong & Yip 1995] D.Y.T. Fong et Yip. *A note on information loss in analysing a mixture model of count data.* Comm. Statist. Theory Methods, vol. 24, pages 3197–3209, 1995. 18

[Givens & Hoeting 2005] G.H. Givens et J.A. Hoeting. Computational statistics. John Wiley & Sons, Inc., 2005. 11

[Gouriéroux & Monfort 1981] G. Gouriéroux et A. Monfort. *Asymptotic Properties of the Maximum Likelihood Estimator in Dichotomous Logit Models.* J. Econom., vol. 17, pages 83–97, 1981. 7, 12, 33, 40, 44, 45, 47

[Govindarajulu 2001] Z. Govindarajulu. Statistical techniques in bioassay. Karger, 2001. 69

[Greene 1994] W. Greene. *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models.* Working Paper, Dep of Eco, New York University, 1994. 20

[Greenland 2000] S. Greenland. *Principles of multilevel modelling.* In. J. Epidemiology, vol. 29, pages 158–167, 2000. 100

[Gubler *et al.* 1984] D.J. Gubler, G. Kuno, G.E. Sather, M. Velez et A. Oliver. *Mosquito cell cultures and specific monoclonal antibodies in surveillance for dengue viruses.* Am. J. Trop. Med. Hyg., vol. 33, pages 158–165, 1984. 98

[Guyon 2001] X. Guyon. Statistique et économétrie - du modèle linéaire aux modèles non-linéaires. Ellipses Marketing, 2001. 8, 40, 43

[Guzman & Kouri 2002] M.G. Guzman et G. Kouri. *Dengue : an update.* Lancet Infect. Dis., vol. 2, pages 33–42, 2002. 93

[Guzman *et al.* 2010] M.G. Guzman, T. Jaenisch, R. Gaczkowski, V.T. Ty Hang, S.D. Sekaran et al. *Multi-country evaluation of the sensitivity and specificity of two commercially-available NS1 ELISA assays for dengue diagnosis*. PloS. Neg. Top. Dis., vol. 4, page e81, 2010. 108

[Hall & Shen 2010] D.B. Hall et J. Shen. *Robust Estimation for Zero-Inflated Poisson Regression.* Scand. J. Statist., vol. 37, pages 237–252, 2010. 22, 23, 24

[Hall 2000] D. Hall. *Zero-inflated Poisson and binomial regression with random effects : a case study.* Biometrics, vol. 56 (4), pages 1030–1039, 2000. 18, 25, 34

[Halstead & O'Rourke 1977] S.B. Halstead et E.J. O'Rourke. *Dengue viruses and mononuclear phagocytes. I. Infection enhancement by non-neutralizing antibody.* J. Exp. Med., vol. 146, pages 201–217, 1977. 109

[Hammond *et al.* 2005] S.N. Hammond, A. Balmaseda, L. Perez, Y. Tellez, S.I. Saborio et al. *Differences in dengue severity in infants, children, and adults in a 3-year hospital-based study in Nicaragua.* Am. J. Trop. Med. Hyg., vol. 73, pages 1063–1070, 2005. 110

[Hase *et al.* 1989] T. Hase, P.L. Summers et K.H. Eckels. *Flavivirus entry into cultured mosquito cells and human peripheral blood monocytes*. Arch. Virol., vol. 104, pages 129–143, 1989. 109

[Hauck 1983] W. Hauck. *A note on confidence bands for the logistic response curve*. Amer. Statist., vol. 37, pages 158–160, 1983. 66, 73

[He *et al.* 2010] X. He, H. Xue et N.Z. Shi. *Sieve Maximum Likelihood Estimation for Doubly Semiparametric Zero-Inflated Poisson Models*. J. Multivar. Anal., vol. 101(9), pages 2026–2038, 2010. 24

[Hilbe 2007] J.M. Hilbe. Negative binomial regression. Cambridge University Press, 2007. 19

[Hilbe 2009] J.M. Hilbe. Logistic regression models. Chapman and Hall : Boca Raton, 2009. 7, 33

[Hommel *et al.* 1998] D. Hommel, A. Talarmin, V. Deubel, J.M. Reynes, M.T. Drouet et al. *Dengue encephalitis in French Guiana*. Res. Virol., vol. 149, pages 235–238, 1998. 94

[Hosmer & Lemeshow 2000] D.W. Hosmer et S. Lemeshow. Applied logistic regression. Editions Wiley, 2000. 7, 33

[Huang *et al.* 2008] J. Huang, S. Ma et Zhang C. H. *The iterated lasso for high-dimensional logistic regression*. Rapport technique, Technical report No. 392, The University of Iowa, 2008. 58, 126

[Innis *et al.* 1989] B.L. Innis, A. Nisalak, S. Nimmannitya, S. Kusalerdchariya, V. Chongswasdi et al. *An enzyme-linked immunosorbent assay to characterize dengue infections where dengue and Japanese encephalitis co-circulate*. Am. J. Trop. Med. Hyg., vol. 40, pages 418–427, 1989. 110

[Kalayanarooj *et al.* 1997] S. Kalayanarooj, D.W. Vaughn, S. Nimmannitya, S. Green et S. et al. Suntayakorn. *Early clinical and laboratory indicators of acute dengue illness*. J. Infect. Dis., vol. 176, pages 313–321, 1997. 94

[Kelley & Anderson 2008] M. E. Kelley et S. J. Anderson. *Zero inflation in ordinal data : incorporating susceptibility to response through the use of a mixture model.* Statist. Med., vol. 27, pages 3674–3688, 2008. 18, 27, 34, 41

[Kemp & Kemp 1988] C.D. Kemp et A.W. Kemp. *Rapid estimation for discrete distributions.* The Statistician, vol. 37, pages 243–255, 1988. 25

[Lam *et al.* 2006] K.F. Lam, H. Xue et Y.B. Cheung. *Semiparametric analysis of zero-inflated count data.* Biometrics, vol. 62, pages 996–1003, 2006. 24, 34

[Lambert 1992] D. Lambert. *Zero-Inflated Poisson Regression Models with an Application to Defects in Manufacturing.* Technometrics, vol. 34, pages 1–14, 1992. 18, 20, 22, 34

[Lanciotti *et al.* 1992] R.S. Lanciotti, C.H. Calisher, D.J. Gubler, G.J. Chang et A.V. Vorndam. *Rapid detection and typing of dengue viruses from clinical samples by using reverse transcriptase-polymerase chain reaction.* J. Clin. Microbiol., vol. 30, pages 545–551, 1992. 98

[Landau & Sheep 1970] H. Landau et L.A. Sheep. *On the supremum of a gaussian process.* Sankhya, vol. Serie A 32, pages 369–378, 1970. xii, xiv, 70, 75, 76, 127

[Lee *et al.* 2006] A. H. Lee, K. Wang, J. A. Scott, K. K. W. Yau et G. J. McLachlan. *Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros.* Statist. Methods in Med Res, vol. 15, pages 47–61, 2006. 34

[Li & Datta 2001] G. Li et S. Datta. *A bootstrap approch to non-parametric re-*

gression for right censored data. Ann. Inst. Statist. Math., vol. 53(4), pages 708–729, 2001. 78

[Li *et al.* 2008] J. Li, E.V. Nordheim, C. Zhang et C.E. Lehner. *Estimation and Confidence Regions for Multi-Dimensional Effective Dose.* Biometrical J., vol. 50, pages 110–122, 2008. 127

[Li *et al.* 2010] J. Li, C. Zhang, K. Doksum et E. Nordheim. *Simultaneous confidence intervals for semiparametric logistics regression and confidence regions for the multi-dimensional effective dose.* Statist. Sinica, vol. 20, pages 637–659, 2010. 78, 83

[Lima *et al.* 2010] Mda R. Lima, R.M. Nogueira, H.G. Schatzmayr et F.B. dos Santos. *Comparison of three commercially available dengue NS1 antigen capture assays for acute diagnosis of dengue in Brazil.* PLoS. Negl. Trop. Dis., vol. 4, page e738, 2010. 108

[Liu *et al.* 2004] W. Liu, M. Jamshidian et Y. Zhang. *Multiple comparison of several regression models.* J. Amer. Stat. Assoc., vol. 99, pages 395–403, 2004. 66

[Liu *et al.* 2005] W. Liu, M. Jamshidian, Y. Zhang et J. Donnelly. *Simulation-based simultaneous confidence bands for a multiple linear regression model when the covariates are constrained.* J. Comput. Graph. Statist., vol. 14(2), pages 459–484, 2005. 67

[Liu *et al.* 2007] W. Liu, M. Jamshidian, Y. Zhang, F. Bretz et X. Han. *Pooling batches in drug stability study by using constant width simultaneous confidence bands.* Statist. Med., vol. 26(14), pages 2759–2771, 2007. 66

[Liu 2011] W. Liu. Simultaneous inference in regression. Chapman & Hall, 2011. 66, 69

[Lu 2008] W. Lu. *Maximum likelihood estimation in the proportional hazards cure model.* Ann. Inst. Statist. Math., vol. 60, pages 545–574, 2008. 37

[Lu 2010] W. Lu. *Efficient estimation for an accelerated failure time model with a cure fraction.* Statist. Sinica, vol. 20, pages 661–674, 2010. 37

[Mackenzie *et al.* 2004] J.S. Mackenzie, D.J. Gubler et L.R. Petersen. *Emerging flaviviruses : the spread and resurgence of Japanese encephalitis, West Nile and dengue viruses.* Nat. Med., vol. 10, pages 98–109, 2004. 94

[Mak 1993] K. T. Mak. *Solving Non-Linear Estimation Equations.* J. Roy. Stat. Soc. Ser. B, vol. 55, pages 945–955, 1993. 11

[Mammen *et al.* 2008] M.P. Mammen, C. Pimgate, C.J. Koenraadt, A.L. Rothman, J. Aldstadt et al. *Spatial and temporal clustering of dengue virus transmission in Thai villages.* PLoS. Med., vol. 5, page e205, 2008. 94, 105, 107

[Mandel & Betensky 2008] M. Mandel et R.A. Betensky. *Simultaneous Confidence Intervals Based on the Percentile Bootstrap Approach.* Comput. Statist. Data Anal., vol. 52(4), pages 2158–2165, 2008. 78

[Marcus & Sheep 1971] M.B. Marcus et L.A. Sheep. *Sample behavior of Gaussian processes.* Proc. Sixth Berkeley Symp. Math. Statist. Probab, vol. 2, pages 423–442, 1971. 75

[Margolin *et al.* 1989] B.H. Margolin, B.S. Kim et K.J. Risko. *The Ames salmo-*

nella/microsome mutagenicity assay : issues of inference and validation. J. Amer. Statist. Assoc., vol. 84, pages 651–661, 1989. 26

[McCullagh & Nelder 1989] P. McCullagh et J.A. Nelder. Generalized linear models. Chapman and Hall, London, chapman and hall, london édition, 1989. 1

[Meier *et al.* 2008] L. Meier, S. Van de Geer et P. Buhlmann. *The group Lasso for logistic regression*. J. Roy. Statist. Soc. Ser. B, vol. 70, pages 53–71, 2008. 58, 126

[Morrison *et al.* 1998] A.C. Morrison, A. Getis, M. Santiago, J.G. Rigau-Perez et P. Reiter. *Exploratory space-time analysis of reported dengue cases during an outbreak in Florida, Puerto Rico, 1991-1992*. Am. J. Trop. Med. Hyg., vol. 58, pages 287–298, 1998. 107

[Morrison *et al.* 2010] A.C. Morrison, S.L. Minnick, C. Rocha, B.M. Forshey et S.T. Stoddard. *Epidemiology of dengue virus in Iquitos, Peru 1999 to 2005 : interepidemic and epidemic patterns of transmission*. PLoS. Negl. Trop. Dis., vol. 4, page e670, 2010. 106

[Murgue *et al.* 1999] B. Murgue, X. Deparis, O. Chungue E. Cassar et C. Roche. *Dengue : an evaluation of dengue severity in French Polynesia based on an analysis of 403 laboratory-confirmed cases*. Trop. Med. Int. Health, vol. 4, pages 765–773, 1999. 94

[Naiman 1987] D.Q. Naiman. *Simultaneous confidence-bounds in multiple-regression using predictor variable constraints*. J. Amer. Stat. Assoc., vol. 82, pages 214–219, 1987. 67

[Naiman 1990] D.Q. Naiman. *On volumes of tubular neighborhoods of spherical polyhedra and statistical inference*. Ann. Statist., vol. 18, pages 685–716, 1990. 67

[Nelder & Wedderburn 1972] J.A. Nelder et R. W. M. Wedderburn. *Generalized linear models*. J. Roy. Statist. Soc. Ser. A, vol. 135, pages 370–384, 1972. 3

[Neumann & Polzehl 1998] N.H. Neumann et J. Polzehl. *Simultaneous bootstrap confidence bands in non-parametric regression*. J. Nonparametr. Stat., vol. 9, pages 307–333, 1998. 78

[Neves-Souza *et al.* 2005] P.C. Neves-Souza, E.L. Azeredo, S.M. Zagne, R. Valls-de Souza, S.R. Reis et al. *Inducible nitric oxide synthase (iNOS) expression in monocytes during acute Dengue Fever in patients and during in vitro infection*. BMC Infect. Dis., vol. 5, page 64, 2005. 109

[Nunes *et al.* 2011] M.R. Nunes, J.P. Neto, S.M. Casseb, K.N. Nunes, L.C. Martins et al. *Evaluation of an immunoglobulin M-specific capture enzyme-linked immunosorbent assay for rapid diagnosis of dengue infection*. J. Virol. Methods, vol. 171, pages 13–21, 2011. 98

[Oliveira *et al.* 2010] M.F. Oliveira, J.M. Galvao Araujo, O.C. Jr. Ferreira, D.F. Ferreira, D.B. Lima et al. *Two lineages of dengue virus type 2*. Brazil. Emerg. Infect. Dis., vol. 16, pages 576–578, 2010. 109

[Pengsaa *et al.* 2006] K. Pengsaa, C. Luxemburger, A. Sabchareon, K. Limkittikul, S. Yoksan et al. *Dengue virus infections in the first 2 years of life and the kinetics of transplacentally transferred dengue neutralizing antibodies in thai children*. J. Infect. Dis., vol. 194, pages 1570–1576, 2006. 110

[Piegorsch *et al.* 2005] W.W. Piegorsch, R.W. West, W. Pan et R. Kodell. *Low dose risk estimation via simultaneous statistical inferences*. J. Roy. Statist. Soc. Ser. C, vol. 54(1), pages 245–258, 2005. 66

[Pradhan & Leung 2006] N.C. Pradhan et P. Leung. *A Poisson and negative binomial regression model of sea turtle interactions in Hawaii's longline fishery*. Fish. Res., vol. 78, pages 309–322, 2006. 21

[Reyes *et al.* 2010] M. Reyes, J.C. Mercado, K. Standish, J.C. Matute, O. Ortega et al. *Index cluster study of dengue virus infection in Nicaragua*. Am. J. Trop. Med. Hyg., vol. 83, pages 683–689, 2010. 106

[Rico-Hesse 2007] R. Rico-Hesse. *Dengue virus evolution and virulence models*. Clin. Infect. Dis., vol. 44, pages 1462–1466, 2007. 94

[Ridout *et al.* 2001] M. Ridout, J. Hinde et C. G. B. Demétrio. *A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives*. Biometrics, vol. 57, pages 219–223, 2001. 34

[Rosen *et al.* 2000] O. Rosen, W. X. Jiang et M. A. Tanner. *Mixtures of marginal models*. Biometrika, vol. 87, pages 391–404, 2000. 23

[Sakuntabhai *et al.* 2005] A. Sakuntabhai, C. Turbpaiboon, I. Casademont, A. Chuansumrit, T. Lowhnoo et al. *A variant in the CD209 promoter is associated with severity of dengue disease*. Nat. Genet., vol. 37, pages 507–513, 2005. 94

[Seppanen & Uusipaikka 1992] E. Seppanen et E. Uusipaikka. *Confidence bands for linear-regression over restricted regions*. Scand. J. Statist., vol. 19, pages 73–81, 1992. 66

[Shu & Huang 2004] P.Y. Shu et J.H. Huang. *Current advances in dengue diagnosis*. Clin. Diagn. Lab. Im., vol. 11, pages 642–650, 2004. 99

[Shu *et al.* 2003] P.Y. Shu, L.K. Chen, S.F. Chang, Y.Y. Yueh, L. Chow et al. *Comparison of capture immunoglobulin M (IgM) and IgG enzyme-linked immunosorbent assay (ELISA) and nonstructural protein NS1 serotype-specific IgG ELISA for differentiation of primary and secondary dengue virus infections*. Clin. Diagn. Lab. Im., vol. 10, pages 622–630, 2003. 110

[Silva *et al.* 2010] L.K. Silva, R.E. Blanton, A.R. Parrado, P.S. Melo, V.G. Morato et al. *Dengue hemorrhagic fever is associated with polymorphisms in JAK1*. Eur. J. Hum. Genet., vol. 18, pages 1221–1227, 2010. 94

[Spurrier 1999] J.D. Spurrier. *Exact confidence bounds for all contrasts of three or more regression lines*. J. Amer. Stat. Assoc., vol. 94, pages 483–488, 1999. 66

[Stoddard *et al.* 2009] S.T. Stoddard, A.C. Morrison, G.M. Vazquez-Prokopec, V. Paz Soldan, T.J. Kochel et al. *The role of human movement in the transmission of vector-borne pathogens*. PloS. Neg. Top. Dis., vol. 3, page e481, 2009. 106

[Sun & Loader 1994] J. Sun et C.R. Loader. *Simultaneous confidence bands for linear regression and smoothing*. Ann. Statist., vol. 22, pages 1328–1346, 1994. 67

[Sun *et al.* 1999] J. Sun, J. Raz et J.J. Faraway. *Simultaneous confidence bands for growth and response curves*. Statist. Sinica, vol. 9(3), pages 679–698, 1999. 66

[Sun *et al.* 2000] J. Sun, C.R. Loader et W.P. McCormick. *Confidence bands in generalized linear models.* Ann. Statist., vol. 28, pages 429–460, 2000. 67

[Teicher 1960] H. Teicher. *On the mixture of distributions.* Ann. Math. Statist., vol. 31, pages 55–73, 1960. 26

[Teicher 1963] H. Teicher. *Identifiability of finite mixtures.* Ann. Math. Statist., vol. 34, pages 1265–1269, 1963. 26

[Teixeira *et al.* 2002] Mda G. Teixeira, Costa Mda C. Barreto M.L., L.D. Ferreira, P.F. Vasconcelos et al. *Dynamics of dengue virus circulation : a silent epidemic in a complex urban area.* Trop. med. and in health, vol. TM & IH 7, pages 757–762, 2002. 106

[Thomas *et al.* 2008] L. Thomas, O. Verlaeten, A. Cabie, S. Kaidomar, V. Moravie et al. *Influence of the dengue serotype, previous dengue infection, and plasma viral load on clinical presentation and outcome during a dengue-2 and dengue-4 coepidemic.* Am. J. Trop. Med. Hyg., vol. 78, pages 990–998, 2008. 94, 109

[Torrentes-Carvalho *et al.* 2009] A. Torrentes-Carvalho, E.L. Azeredo, S.R. Reis, A.S. Miranda, M. Gandini et al. *Dengue-2 infection and the induction of apoptosis in human primary monocytes.* Mem. Inst. Oswaldo Cruz, vol. 104, pages 1091–1099, 2009. 109

[Tricou *et al.* 2010] V. Tricou, H.T. Vu, N.V. Quynh, C.V. Nguyen, H.T. Tran, et al. *Comparison of two dengue NS1 rapid tests for sensitivity, specificity and relationship to viraemia and antibody responses.* BMC Infect. Dis., vol. 10, page 142, 2010. 108, 109

[Twiddy *et al.* 2005] S.S. Twiddy, J.J. Farrar, N. Vinh Chau, B. Wills, E.A. Gould et al. *Phylogenetic relationships and differential selection pressures among genotypes of dengue-2 virus*. Virology, vol. 298, pages 63–72, 2005. 109

[Vu *et al.* 2010] T.T. Vu, E.C. Holmes, V. Duong, T.Q. Nguyen, T.H. Tran et al. *Emergence of the Asian 1 genotype of dengue virus serotype 2 in viet nam : in vivo fitness advantage and lineage replacement in South-East Asia*. PloS. Neg. Top. Dis., vol. 4, page e757, 2010. 109

[Wang 1994] P. Wang. *Mixed regression models for discrete data*. PhD thesis, University of British Columbia, Vancouver, 1994. 27

[Watts *et al.* 1999] D.M. Watts, K.R. Porter, P. Putvatana, B. Vasquez, C. Calampa et al. *Failure of secondary infection with American genotype dengue 2 to cause dengue haemorrhagic fever*. Lancet, vol. 354, pages 1431–1434, 1999. 94

[WHO 2009] WHO. *Dengue : guidelines for diagnosis, treatment, prevention and control*. Rapport technique, Geneva : World Health Organization., 2009. 93, 94

[Working & Hotelling 1929] H. Working et H. Hotelling. *Applications of the theory of error to the interpretation of trends*. J. Amer. Stat. Assoc., vol. 24, pages 73–85, 1929. 66

[Xiang *et al.* 2007] L. Xiang, A. H. Lee, K. K. W. Yau et G. J. McLachlan. *A score test for overdispersion in zero-inflated Poisson mixed regression model*. Statist. Med., vol. 1608-1622, page 26, 2007. 34

[Yip 1988] P. Yip. *Inference about the mean of a Poisson distribution in the presence* *of a nuisance parameter*. Australian J. Statist., vol. 30, pages 299–306, 1988. 18

[Yip 1991] P. Yip. *Conditional inference for a mixture model for the analysis of count data*. Comm. Statist. Theory Methods, vol. 20, pages 2045–2057, 1991. 18

[Young *et al.* 2000] P.R. Young, P.A. Hilditch, C. Bletchly et W. Halloran. *An antigen capture enzyme-linked immunosorbent assay reveals high levels of the dengue virus protein NS1 in the sera of infected patients*. J. Clin. Microbiol., vol. 38, pages 1053–1057, 2000. 94

[Zhang & Peng 2010] W. Zhang et H. Peng. *Simultaneous confidence band and hypothesis test in generalised varying-coefficient models*. J. Mulivar. Anal., vol. 101, pages 1656–1680, 2010. 66, 78, 83

[Zhou & Tu 2000] X. Zhou et W. Tu. *Confidence intervals for the mean of diagnostic test charge data containing zeros*. Biometrics, vol. 56 (4), pages 1118–1125, 2000. 18