Modélisation d'une probabilité d'infection en présence d'immunité

Diop A.[†], Diop A.[†], Dupuy J.-F.[*]

[†]: Laboratoire de Mathématiques, Université de Saint-Louis, Sénégal
[*]: Laboratoire Mathématiques, Image et Applications, La Rochelle

Séminaire de l'équipe Biostatistique, Recherche Clinique & Mesures Subjectives en Santé - Nantes, 28/01/2011

Outline of the talk

1. Logistic regression: introduction and the problem of immunes
2. Estimation in logistic regression with immunes
3. Simulation results
4. Discussion

1. Logistic regression: introduction and the problem of immunes
2. Estimation in logistic regression with immunes
3. Simulation results
4. Discussion

Logistic regression for binary responses

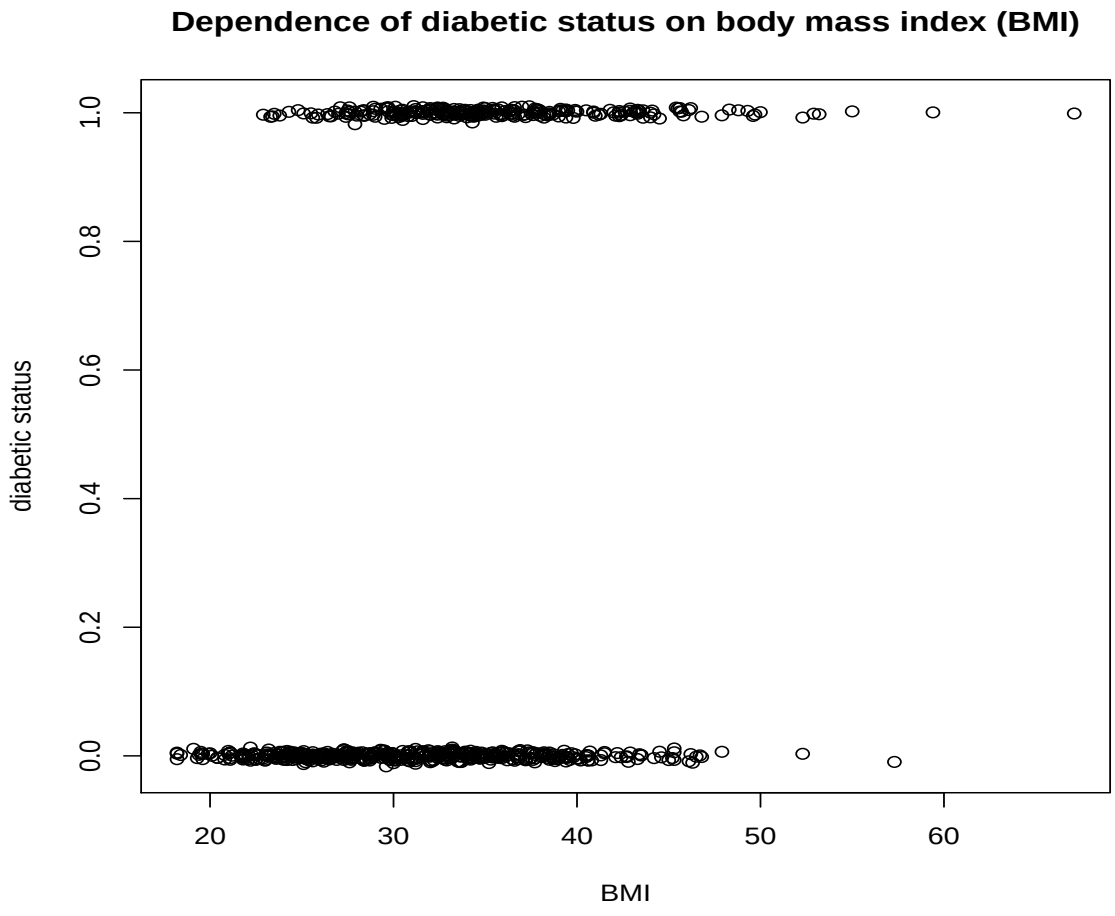
Logistic regression deals with the statistical analysis of binary 0 or 1 data.

Example: Study on diabetes among 768 females Pima Indians in Phoenix, USA.

$\hookrightarrow$ investigates the effect of risk factors (or covariates) on the fact of being diabetic (1) or not (0):

1. age of the woman
2. body mass index (BMI = weight[kg]/height[m]^2)
3. ...

Binary response data: an example



It seems that the number of 1's tends to increase relative to the number of 0's with increasing BMI. Thus BMI seems to be a positiver risk factor for getting diabete.

Logistic regression: the mathematical formulation

The logistic regression model is appropriate for analyzing such data:

$$\mathbb{P}(Y_i = 1|X_{i1}, X_{i2}) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})}$$

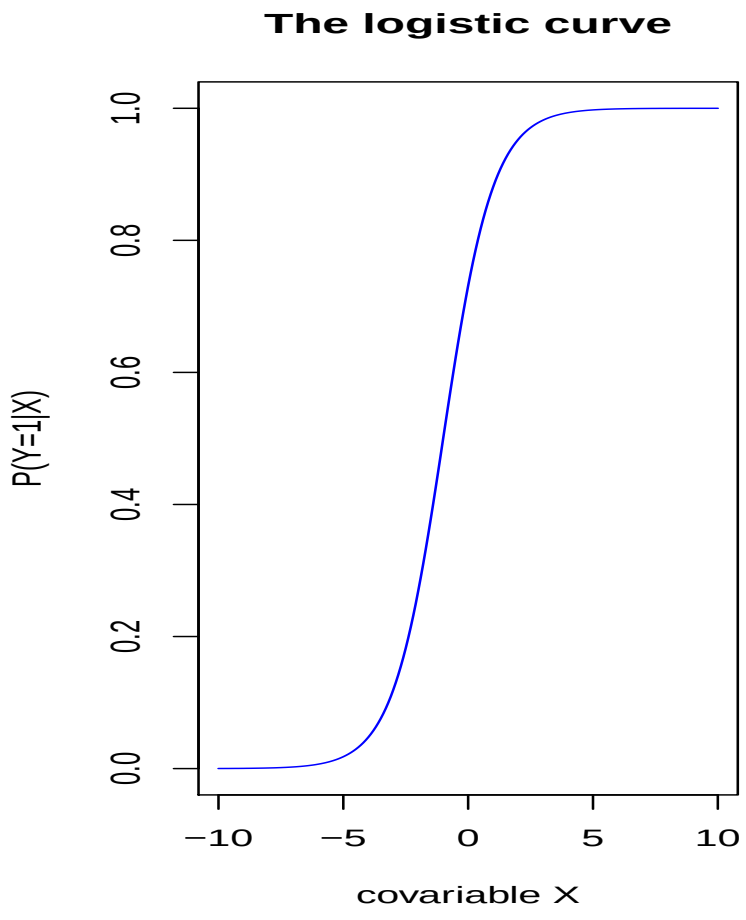
with $X_{i1} =$ age, $X_{i2} =$ BMI, for the i-th woman.

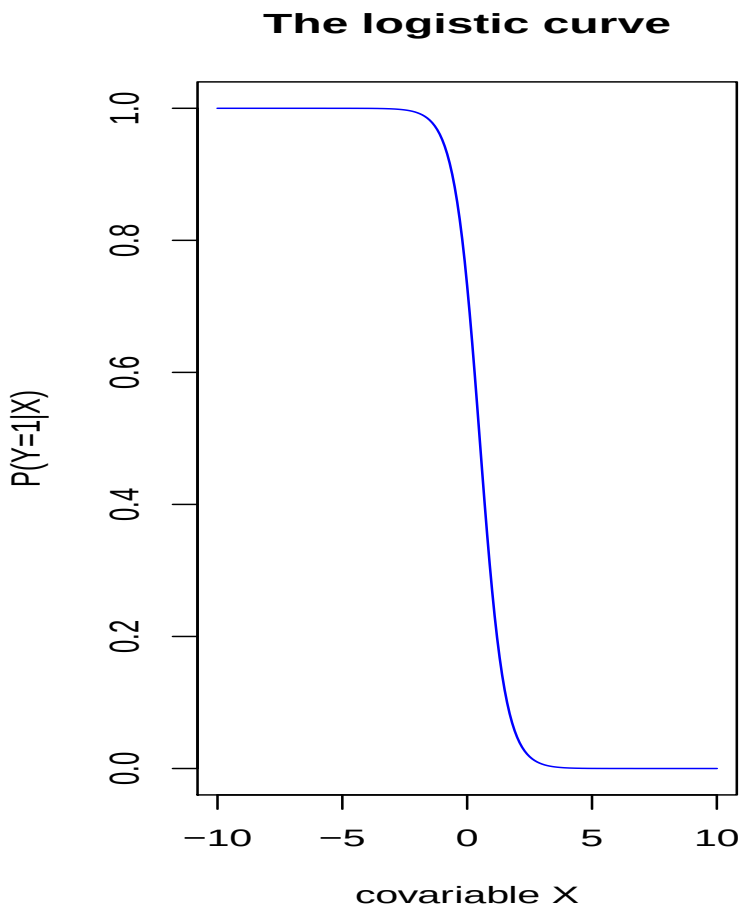
The combination $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$ is called the linear predictor.

An alternative way of writing this model is as:

$$\underbrace{\log\left(\frac{\mathbb{P}(Y_i = 1|X_{i1}, X_{i2})}{1 - \mathbb{P}(Y_i = 1|X_{i1}, X_{i2})}\right)}_{\text{logit}(\mathbb{P}(Y_i=1|X_{i1},X_{i2}))} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

The logistic curve: examples





Statistical analysis of binary data

Objectives: From a sample of n individuals $(Y_i, X_{i1}, \ldots, X_{i2})$:

1. estimate the parameters β_j,
2. test hypothesis about the β_j, such as $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$,
3. predict a particular probability $\pi_i = \mathbb{P}(Y_i = 1 | X_{i1}, X_{i2})$ as:

$$\widehat{\pi}_i = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2})}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2})}$$

Maximum likelihood estimation

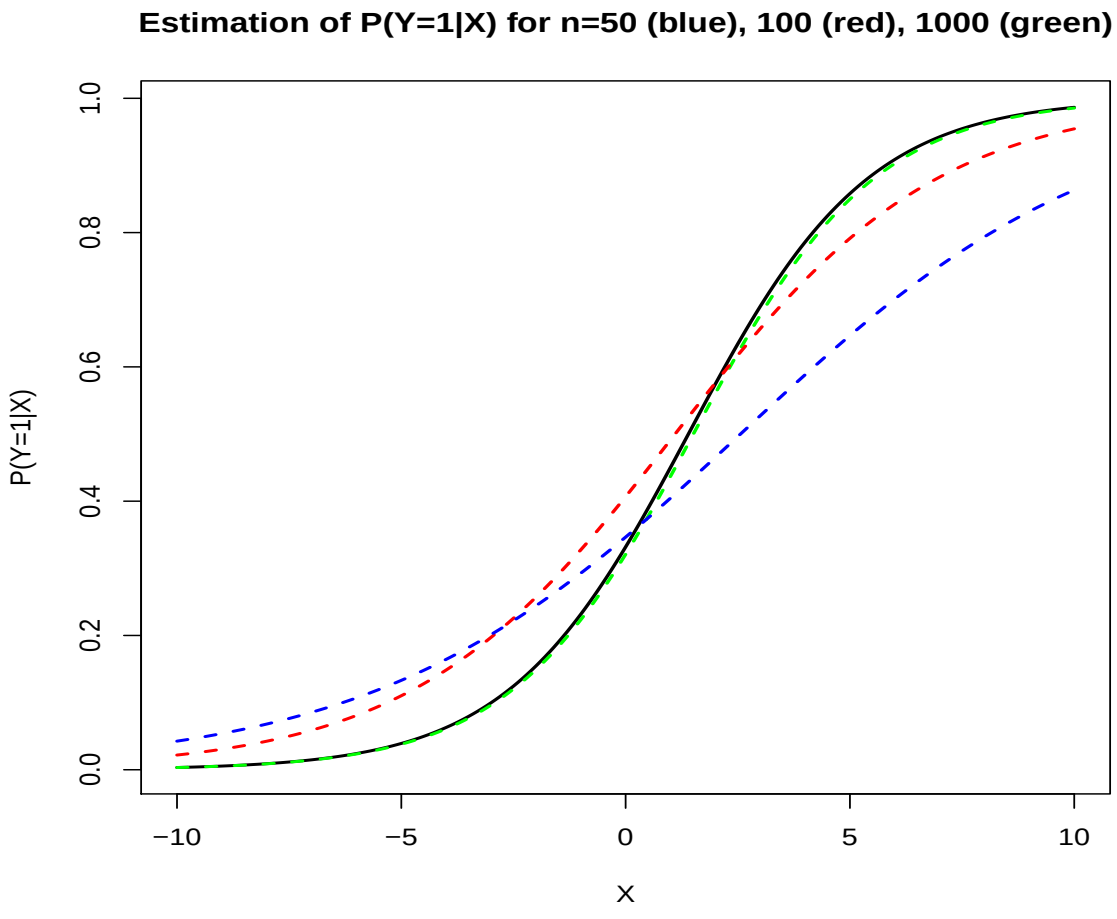
β is estimated by the value $\widehat{\beta}_n$ which maximizes the likelihood

$$L_n(\beta) = \prod_{i=1}^{n} \left[\mathbb{P}(Y_i = 1|X_{i1}, X_{i2})\right]^{Y_i} \left[\mathbb{P}(Y_i = 0|X_{i1}, X_{i2})\right]^{1-Y_i}$$

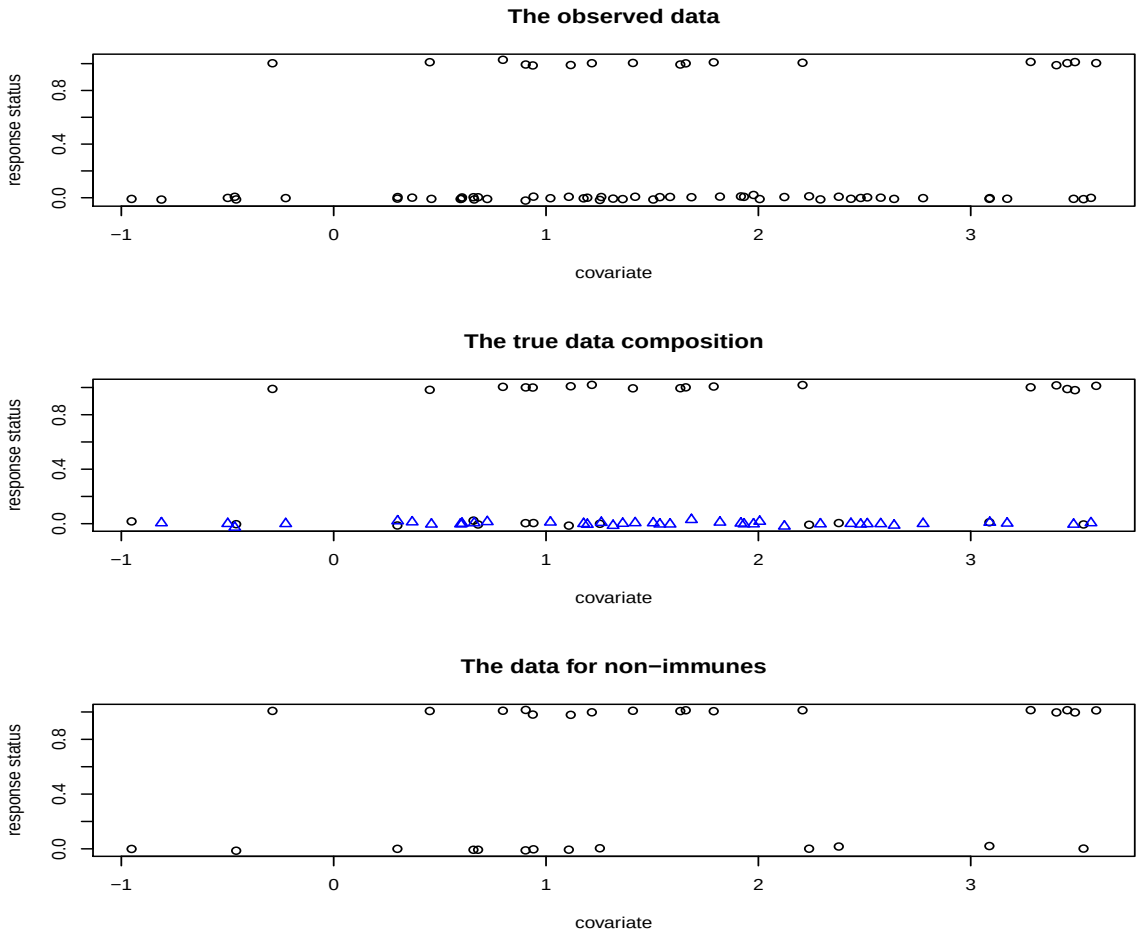
This estimator has some nice properties:

1. $\widehat{\beta}_n$ "is closer and closer" to the unknown β,
2. $\widehat{\beta}_n$ is distributed approximately as a normal law $\Rightarrow$ confidence intervals, p-values.

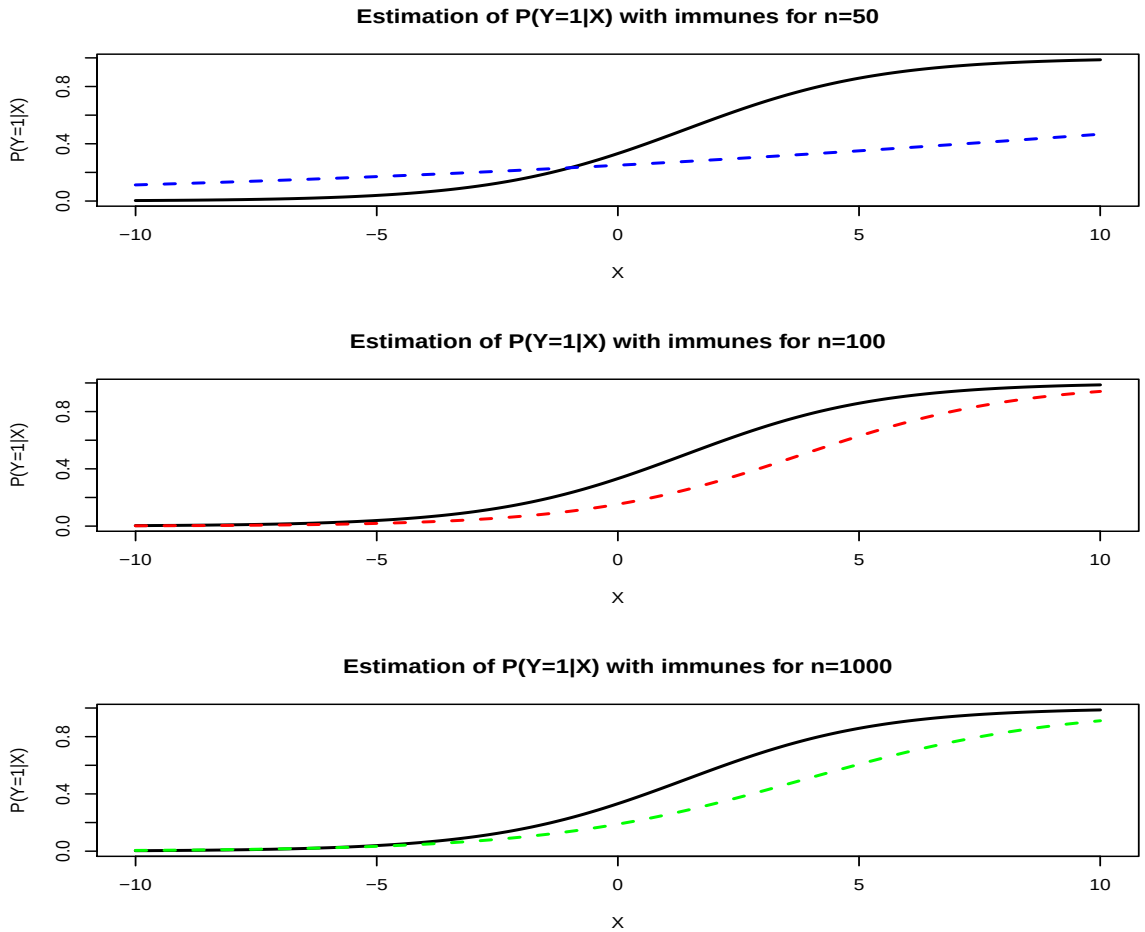
Estimation of the logistic curve



The problem of immunes



Estimation of the logistic curve with immunes



1. Logistic regression: introduction and the problem of immunes
2. Estimation in logistic regression with immunes
3. Simulation results
4. Discussion

The proposed estimation procedure

With immunity, we are in fact led to estimate the β_j in the model:

$$\begin{cases} \mathbb{P}(Y_i = 1 | X_{i1}, X_{i2}, \mathbf{S_i = 1}) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})} \\ \\ \mathbb{P}(Y_i = 1 | X_{i1}, X_{i2}, \mathbf{S_i = 0}) = 0 \end{cases}$$

where $S_i = 0$ if the patient is immune and $S_i = 1$ otherwise (susceptible).

This problem falls within the general context of zero-inflated regression:

- zero-inflated Poisson: Lambert (1992), Dietz and Bohning (2000), Lam (2006), Xiang *et al.* (2007),...
- zero-inflated binomial: Hall (2000),...
- zero-inflated proportional odds: Kelley and Anderson (2008)

The proposed estimation procedure

Estimation is still possible if we can model the probability of being cured, for example, by a logistic regression model:

$$\mathbb{P}(S_i = 1|Z_{i1}, Z_{i2}) = \frac{\exp(\theta_0 + \theta_1 Z_{i1} + \theta_2 Z_{i2})}{1 + \exp(\theta_0 + \theta_1 Z_{i1} + \theta_2 Z_{i2})}$$

since then:

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

and β and θ are estimated by maximizing the likelihood

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

The identifiability issue

Some important regularity conditions:

- The covariates are bounded. The $X_{i1}, X_{i2}, \ldots$ are linearly independent. The $Z_{i1}, Z_{i2}, \ldots$ are linearly independent

  $\hookrightarrow$ "classical conditions" for standard logistic regression.

- There exists one continous covariate V which is in $\mathbf{X}_i$ but not in $\mathbf{Z}_i$. Moreover, at the model-building stage, it is known that V is in $\mathbf{X}_i$.

Parameter exchangeability

Recall that

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

If $\mathbf{X}_i=\mathbf{Z}_i$ (contain the same covariates) then: $L_n(\beta,\theta)=L_n(\theta,\beta)=$

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

For example, $\beta=(1,3)$ and $\theta=(2,3.5)$. β and θ are exchangeable and cannot be identified from the data.

$\Rightarrow$ the model is non-identifiable. No convergent estimation procedure can exist.

Assuming there is one covariate in $\mathbf{X}_i$ which is not in $\mathbf{Z}_i$ $\Rightarrow$ $\mathbf{X}_i\neq\mathbf{Z}_i$.

Linear predictors (l.p.) exchangeability

The same likelihood value

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

can arise from the following two models:

$$\begin{cases}\mathbb{P}(Y_i=1|\mathbf{X}_i,S_i=1)=\frac{e^{\beta'\mathbf{X}_i}}{1+e^{\beta'\mathbf{X}_i}}\\ \mathbb{P}(S_i=1|\mathbf{Z}_i)=\frac{e^{\theta'\mathbf{Z}_i}}{1+e^{\theta'\mathbf{Z}_i}}\end{cases}\qquad\begin{cases}\mathbb{P}(Y_i=1|\mathbf{X}_i,S_i=1)=\frac{e^{\theta'\mathbf{Z}_i}}{1+e^{\theta'\mathbf{Z}_i}}\\ \mathbb{P}(S_i=1|\mathbf{Z}_i)=\frac{e^{\beta'\mathbf{X}_i}}{1+e^{\beta'\mathbf{X}_i}}\end{cases}$$

The l.p. $\beta'\mathbf{X}_i$ and $\theta'\mathbf{Z}_i$ are exchangeable and the sub-models for Y_i and S_i cannot be identified from the data $\Rightarrow$ the model is non-identifiable.

Knowing, prior to model fitting, which l.p. the covariate V is attached to will force each l.p. to be attached to the correct sub-model.

Identifiability of finite mixture of logistic regressions

The condition that V is continuous should be understood with respect to the problem of:

Mixture of c logistic regressions (Follmann and Lambert, 1991) with constant mixing probabilities.

The model is identifiable if the number of distinct covariate combinations values is "sufficiently large". Specifically, FL show that c has to be constrained by

$$c \leq \sqrt{N+2} - 1$$

with N the number of distinct observed values of the covariate vector.

$\hookrightarrow$ a single $0/1$ covariate will identify only one component,

$\hookrightarrow$ a mixture of two Bernoulli distributions is identifiable if the number of unique combinations of the covariate vector is at least 7.

1. Logistic regression: introduction and the problem of immunes
2. Estimation in logistic regression with immunes
3. Simulation results
4. Discussion

A numerical investigation of identifiability

We simulate data from the model defined by

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

and

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

where $X_{i2} \sim \mathcal{N}(0,1)$, $Z_{i2} \sim \mathcal{N}(1,1)$, and Z_{i3} and Z_{i4} are indicator variables built from a categorical variable with 3 categories.

A numerical investigation of identifiability (ctd)

Results for $\beta = (-1.7, -2, -3.4, 5, .3)$ and $\theta = (.71, 1, 2, -3)$ (25% of immunes).

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
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	-1.709	-2.513	-3.843	5.540	0.301	0.824	0.976	2.558	-3.576
	(1.819)	(1.015)	(1.667)	(2.503)	(3.132)	(2.319)	(3.073)	(2.691)	(2.941)
	[1.348]	[0.715]	[1.204]	[1.845]	[2.296]	[1.838]	[2.118]	[2.084]	[2.376]
500	-1.695	-2.093	-3.286	4.954	0.301	0.761	0.988	2.316	-2.745
	(0.999)	(0.543)	(1.063)	(1.265)	(1.848)	(1.203)	(1.566)	(1.485)	(2.489)
	[0.741]	[0.395]	[0.760]	[0.953]	[1.481]	[0.986]	[1.060]	[1.135]	[1.830]

Note: $(\cdot)$: root mean square error. $[\cdot]$: mean absolute error. The percentage of infected among the susceptibles is 30%. All results are based on 1000 replicates.

A numerical investigation of identifiability (ctd)

Results for $\beta = (-1.7, -2, -3.4, 5, .3)$ and $\theta = (-.3, -1, 2.1, 1)$ (50% of immunes).

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
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	-1.716	-2.641	-3.816	5.866	0.302	-0.279	-1.537	2.616	1.352
	(2.455)	(1.491)	(1.866)	(3.121)	(3.133)	(1.942)	(1.909)	(2.749)	(3.155)
	[1.830]	[1.127]	[1.467]	[2.531]	[2.296]	[1.484]	[1.334]	[2.143]	[2.469]
500	-1.714	-2.281	-3.764	5.295	0.301	-0.313	-1.317	2.364	1.211
	(1.341)	(0.794)	(1.257)	(1.929)	(1.907)	(1.071)	(1.222)	(1.689)	(1.881)
	[1.053]	[0.597]	[0.951]	[1.554]	[1.431]	[0.858]	[0.760]	[1.263]	[1.474]

A numerical investigation of identifiability (ctd)

Results for $\beta = (-1.7, -2, -3.4, 5, .3)$ and $\theta = (.4, -1, -.6, -2)$ (75% of immunes).

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
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	-1.581 (2.951) [2.157]	-2.792 (2.412) [1.897]	-3.847 (3.687) [2.912]	5.502 (5.192) [4.434]	0.248 (3.214) [2.488]	0.469 (2.256) [1.803]	-1.571 (2.042) [1.287]	-0.501 (2.356) [1.896]	-1.846 (3.517) [2.834]
500	-1.530 (1.446) [1.022]	-2.435 (1.466) [1.142]	-3.714 (1.934) [1.563]	5.331 (3.221) [2.659]	0.292 (2.110) [1.700]	0.464 (1.335) [1.076]	-1.323 (0.976) [0.611]	-0.562 (1.678) [1.307]	-1.901 (1.978) [1.509]

Simulation study: the normal approximation

We simulate data from the model defined by

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

and

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

where $X_{i2} \sim \mathcal{N}(0,1)$ and $Z_{i2} \sim \mathcal{N}(1,1)$.

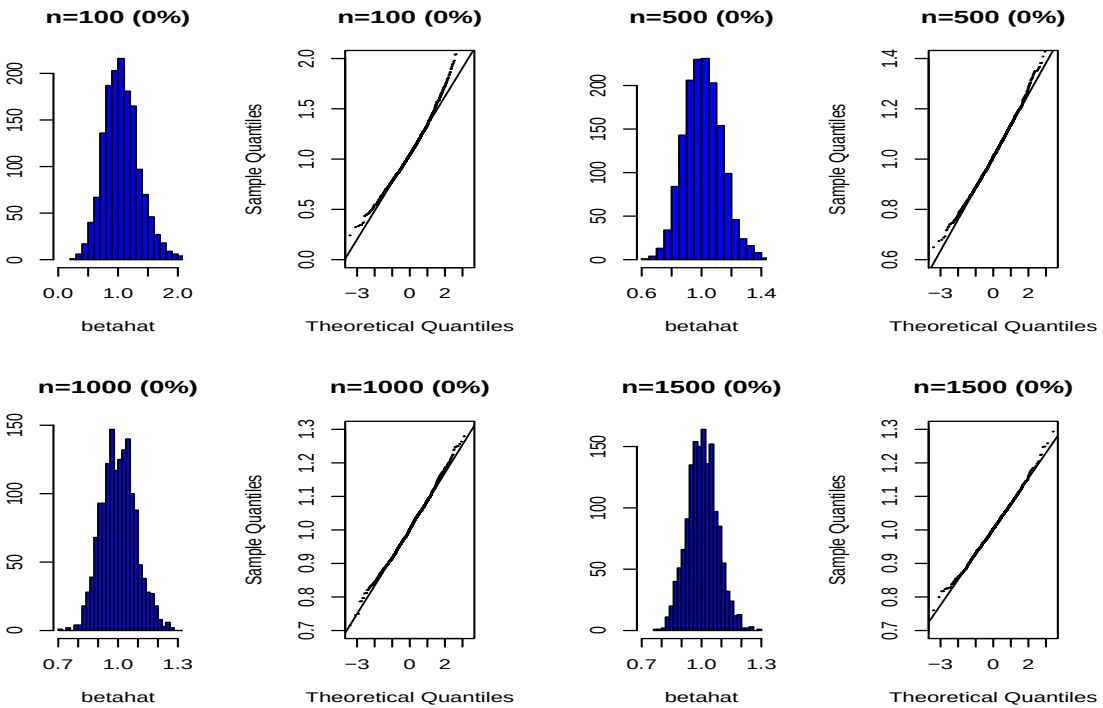
The sample size is taken: $n = 100, 500, 1000, 1500$ and the percentage of immunes in the sample: 25%, 50%, and 75%.

Simulation results for $\beta = (-.8, 1)$

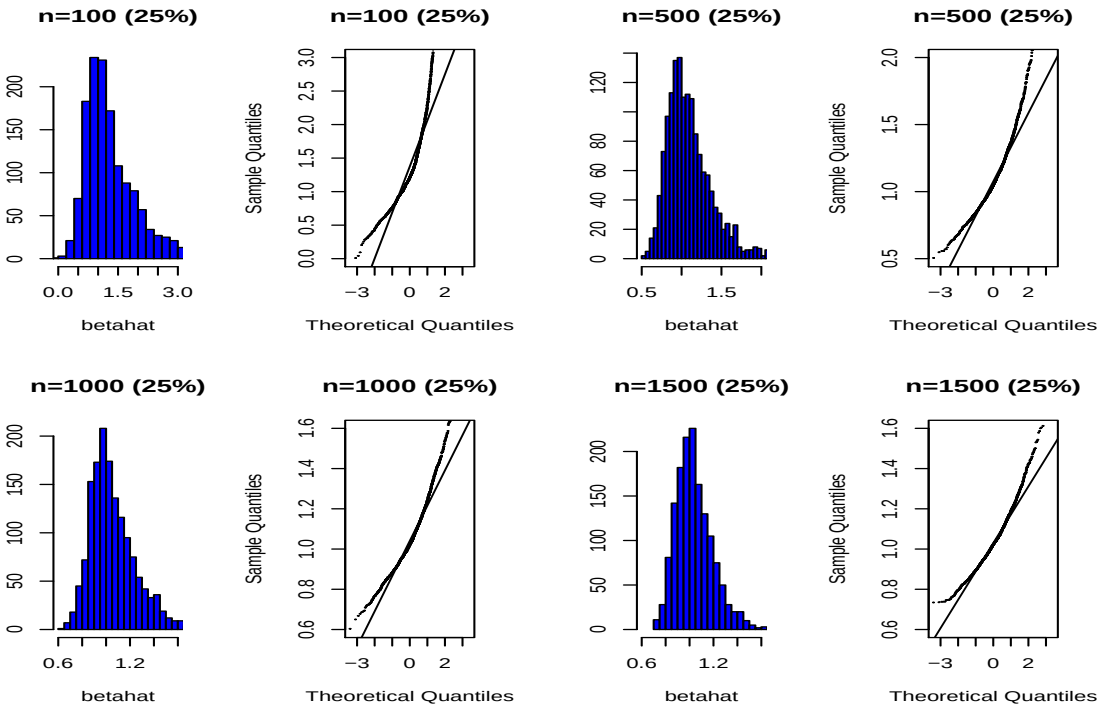
	percentage of immunes in the sample							
	0%		25%		50%		75%	
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$
100	-0.834	1.064	-0.773	1.114	-0.787	1.137	-0.750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

Note: $(\cdot)$: root mean square error. $[\cdot]$: mean absolute error. For each % of immunes, the % of infected among the susceptibles is 30%.

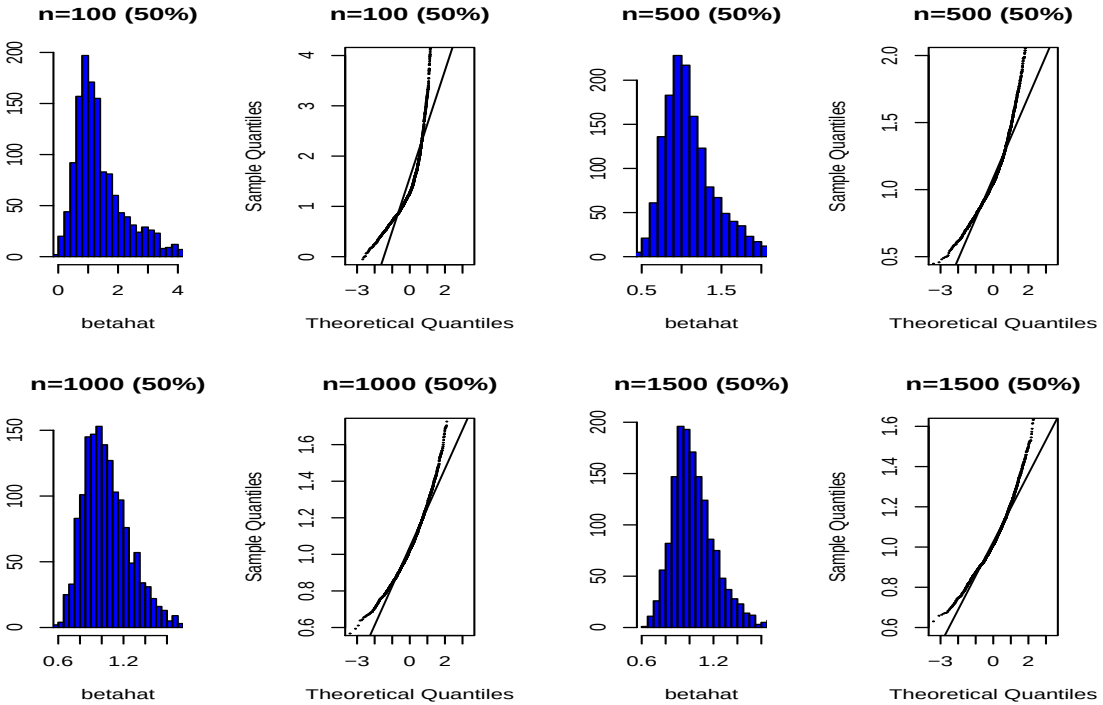
Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ (no immunes)



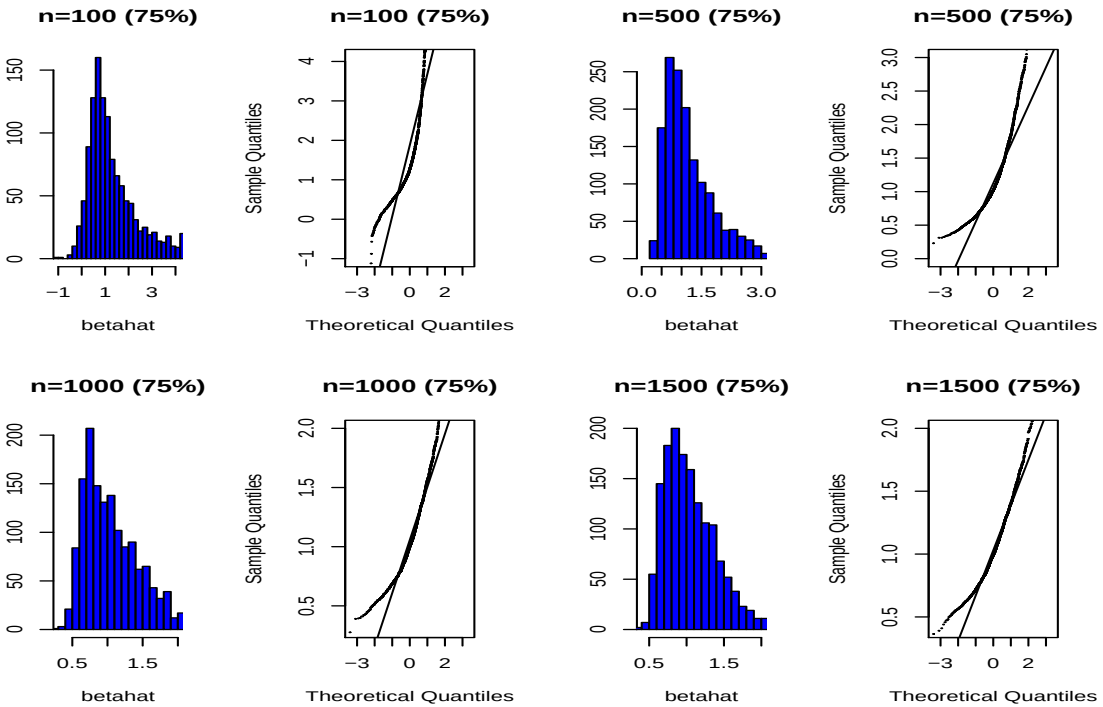
Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ (25% of immunes)



Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ (50% of immunes)



Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ (75% of immunes)



Simulation results for $\beta = (-.8, 0)$

	percentage of immunes in the sample							
	0%		25%		50%		75%	
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$
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

1. Logistic regression: introduction and the problem of immunes
2. Estimation in logistic regression with immunes
3. Simulation results
4. Discussion

Discussion and perspectives

- Logistic regression with a cure fraction can be viewed as a zero-inflated Bernoulli regression problem. The proposed model extends the ones previously investigated in the domain.
- Confidence bands for the probability of infection are under investigation.
- Robustness to misspecification of the model for the cure fraction.
- Further generalization: random-effects logistic regression for clustered data (family, treatment).