SAEM FOR LOGISTIC REGRESSION WITH MISSING VALUES

{ WEI JIANG, JULIE JOSSE, MARC LAVIELLE }
CMAP, ECOLE POLYTECHNIQUE & INRIA XPOP

INTRODUCTION

Traumabase data (`traumabase.eu`) include **7495** major trauma patients distributed across **8** hospitals in Paris and **244** pre and post-hospital measurements. Major trauma is a public health challenge and a major source of mortality and handicap around the world. The aims of project include:

- Develop models to help the decisions taken by the emergency doctors;
- Provide real-time decision support.

Type of Accident	Age	Sex	Blood Pressure	Lactate	Hemorrhagic Shock
Falling	50	M	140	NM	Yes
Fire	28	F	NR	4.8	No
Knife	30	M	120	1.2	No
Traffic	23	M	110	3.6	No
Knife	33	M	106	NM	No
Traffic	58	F	150	NM	Yes

NA = Not Applicable, NM = Not Made, NR = Not Recorded.

Objectives

(Medical) Measurements $\xrightarrow{\text{Predict}}$ Hemorrhagic shock

(Statistical) X mixed $\xrightarrow[\text{Regression}]{\text{Logistic}}$ $Y = 0, 1$

Challenge: Modeling with missing data.

Proposal

- **Logistic regression** model with missing data;
- Parameter estimation based on **MLE** via **SAEM** [3];
- Variance estimation via Fisher information;
- Model selection with missing data based on likelihood;
- Assessment by simulation study;
- Application to the Traumabase data set;
- Comparison to existing imputation methods as *mice* [5], *imputePCA* [2].

MODEL

$x = (x_{ij})$ a $n \times p$ matrix of quantitative covariates
$y = (y_i)$ an n-vector of binary responses $\{0, 1\}$
Logistic regression model

$$\mathbb{P}(y_i = 1|x_i; \beta) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}$$

Covariables

$$x_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$$

Log-likelihood for complete-data with the set of parameters $\theta = (\mu, \Sigma, \beta)$

$$\mathcal{LL}(\theta; x, y) = \sum_{i=1}^{n} \Big(\log(\mathrm{p}(y_i|x_i; \beta)) + \log(\mathrm{p}(x_i; \mu, \Sigma)) \Big).$$

Missing data mechanism: Missing at random [4]
Decomposition: $x_i = (x_{i,\text{mis}}, x_{i,\text{obs}})$

ALGORITHM

Parameter estimation: Stochastic approximation EM

Starting from an initial guess θ_0, the kth iteration consists of three steps:

- **Simulation:** For $i = 1, 2, \cdots, n$, draw one sample $x_{i,\text{mis}}^{(k)}$ from $\mathrm{p}(x_{i,\text{mis}}|x_{i,\text{obs}}, y_i; \theta_{k-1})$.
- **Stochastic approximation:** Update the function Q
  $Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\mathcal{LL}(\theta; x_{\text{obs}}, x_{\text{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right)$,
  where (γ_k) is a decreasing sequence of positive numbers.
- **Maximization:** Update the estimation of θ: $\theta_k = \arg\max_\theta Q_k(\theta)$.

Metropolis

- $\boldsymbol{z}_{im}^{(k)} \sim g_i(\boldsymbol{x}_{i,mis}) = \mathrm{p}(x_{i,\text{mis}}|x_{i,\text{obs}}; \mu, \Sigma) \sim \mathcal{N}_p(\mu_i, \Sigma_i)$, $u \sim \mathcal{U}[0, 1]$
- $r = \dfrac{f_i(\boldsymbol{z}_{im}^{(k)})/g_i(\boldsymbol{z}_{im}^{(k)})}{f_i(\boldsymbol{z}_{i,m-1}^{(k)})/g_i(\boldsymbol{z}_{i,m-1}^{(k)})}$; If $u < r$, accept $\boldsymbol{z}_{im}^{(k)}$

Variance estimation: Louis formula [?]

Observed information = Complete data information - Missing information [1].

$$\begin{aligned}\mathcal{I}(\theta) = &- \mathbb{E}\left(\frac{\partial^2 \mathcal{LL}(\theta; x, y)}{\partial\theta\partial\theta^T} | x_{\text{obs}}, y; \theta \right) \\ &- \mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta} \frac{\partial \mathcal{LL}(\theta; x, y)^T}{\partial\theta} | x_{\text{obs}}, y; \theta \right) \\ &+ \mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta} | x_{\text{obs}}, y; \theta \right) \mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta} | x_{\text{obs}}, y; \theta \right)^T.\end{aligned}$$

Model selection based on observed likelihood

$$\begin{aligned}\mathrm{p}(y_i, x_{i,\text{obs}}; \theta) &= \int \mathrm{p}(y_i, x_{i,\text{obs}}|x_{i,\text{mis}}; \theta)\mathrm{p}(x_{i,\text{mis}}; \theta) dx_{i,\text{mis}} \\ &= \mathbb{E}_{g_i}\left(\mathrm{p}(y_i, x_{i,\text{obs}}|x_{i,\text{mis}}; \theta) \frac{\mathrm{p}(x_{i,\text{mis}}; \theta)}{g_i(x_{i,\text{mis}})} \right).\end{aligned}$$

SIMULATION STUDY

1000 times of data simulation with NA.

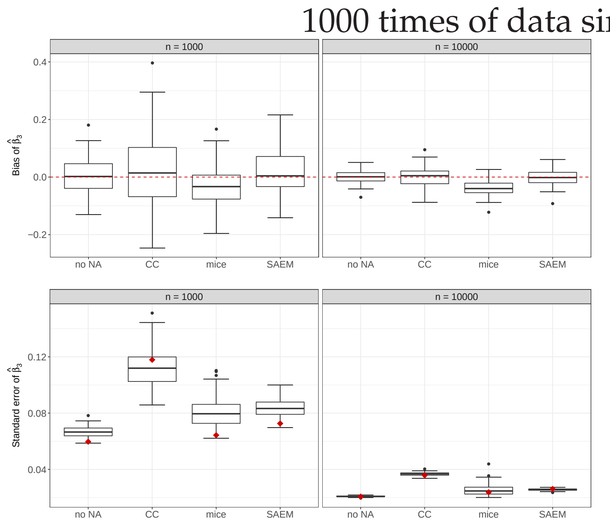


Figure 1: Top: Empirical distribution of the **bias** of $\hat{\beta}_3$; Bottom: Distribution of the estimated **standard errors** of $\hat{\beta}_3$.

Table 1: Comparison of execution time.

Time	no NA	MCEM	mice	SAEM
mean	4.65×10^{-3}	773	0.70	13.50

Table 2: Coverage (%) for $n = 10\,000$.

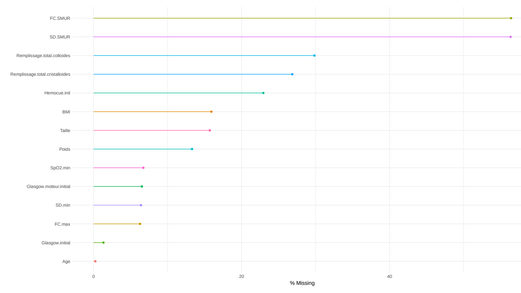
	no NA	CC	mice	SAEM
β_0	95.2	94.4	95.2	94.9
β_1	96.0	94.7	93.9	95.1
β_2	95.5	94.6	94.0	94.3
β_3	94.9	94.3	**86.5**	94.7
β_4	94.6	94.2	96.2	95.4
β_5	95.9	94.4	**89.6**	94.7

Table 3: Percentage of times that different criterion select the correct true model (C), overfit (O) and underfit (U).

	Non-Correlated			Correlated		
Criterion	C	O	U	C	O	U
BIC_{obs}	**92**	3	5	**94**	2	4
BIC_{orig}	96	2	2	93	0	7
BIC_{cc}	79	1	20	91	0	9
AIC_{obs}	60	40	0	65	32	3
AIC_{orig}	73	27	0	75	20	5
AIC_{cc}	67	32	1	77	16	7

APPLICATION ON TRAUMABASE

- Data preprocessing ⇒ **6384 patients**.
- Clinical experience ⇒ **14 influential quantitative measurements**
- Full model estimation ⇒ Find the **outliers**: For 1144 th patient, Weight (200 kg) and Height (100 cm).



- 15 times of random data split ⇒ 80% training set + 20 % test set
- Estimation & model selection via BIC and forward method
  ⇒ Available interpretation of results. Ex: The more one bleed, the lower the HemoCu is, and the more the blood will be transfused. Then the more likely one will end up in a hemorrhagic shock.

Metrics	SAEM	impPCA	impMean	*mice*
Brier score	**6.316**	6.351	6.324	6.400
Logarithmic score	**-22.04**	-22.21	-22.08	-22.12
AUC	**86.70**	86.67	86.62	86.62
Accuracy	**83.23**	81.96	82.74	**83.54**
Sensitivity	**78.26**	77.59	76.86	76.29
Specificity	83.70	82.21	83.15	**84.58**
Precision	30.56	30.68	30.97	**32.88**

CONCLUSION & FUTURE RESEARCH

- SAEM for logistic regression with missingness can be computationally efficient;
- Unbiased estimation and a more reasonable coverage of confidence interval than competitors;
- Model selection by criterion BIC with missing data can be well performed.
- Transformation of variables to meet the hypothesis of normal distribution;
- Deal with both quantitative and categorical data;
- Deal with both MAR and MNAR missing values.
- Causal inference and especially for the propensy score analysis.

REFERENCES

[1] J. G. Ibrahim, M.-H. Chen, and S. R. Lipsitz. Monte carlo em for missing covariates in parametric regression models. *BIOMETRICS*, 55:591–596, 1999.

[2] J. Josse and F. Husson. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

[3] M. Lavielle. *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools*. Chapman and Hall/CRC, 2014.

[4] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., 2002.

[5] S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.