Classification using a joint model of longitudinal data and binary outcomes based on the SAEM algorithm

Maritza Márquez [*], Cristian Meza, Claudio Fuentes, Rolando de la Cruz

[*] *Universidad Adolfo Ibañez, Chile*

June 14, 2023

Collaborators

Cristian Meza
cristian.meza@uv.cl
CIMFAV - Universidad de Valparaíso
Chile

Claudio Fuentes
fuentesc@oregonstate.edu
Oregon State University, Corvallis
USA

Rolando De la Cruz
rolando.delacruz@uai.cl
Universidad Adolfo Ibañez
Chile

Index

Motivation

Chilean women pregnancies data

(β-HCG hormone)

De la Cruz et al. (2011) and De la Cruz et al. (2016)

- They analyzed data from a clinical study on the risk of loss in a group of pregnant Chilean women.
- They model the association between a binary outcome (pregnancy outcome) and features of longitudinal measurements (hormone levels) through a common set of latent random effects in **173** women during the first trimester using different modeling strategies.
- These women were classified into two groups:
  - **Normal group** (**124** women who came to term with their pregnancy).
  - **Abnormal group** (**49** women who suffered a loss).
- They are unbalanced data that fluctuate between **1 to 6** observations, having a total of **375** observations.

Chilean women pregnancies data

(β-HCG hormone)

De la Cruz et al. (2011) and De la Cruz et al. (2016)

- They analyzed data from a clinical study on the risk of loss in a group of pregnant Chilean women.
- They model the association between a binary outcome (pregnancy outcome) and features of longitudinal measurements (hormone levels) through a common set of latent random effects in **173** women during the first trimester using different modeling strategies.
- These women were classified into two groups:
  - **Normal group** (**124** women who came to term with their pregnancy).
  - **Abnormal group** (**49** women who suffered a loss).
- They are unbalanced data that fluctuate between **1 to 6** observations, having a total of **375** observations.

Chilean women pregnancies data

(β-HCG hormone)

De la Cruz et al. (2011) and De la Cruz et al. (2016)

- They analyzed data from a clinical study on the risk of loss in a group of pregnant Chilean women.
- They model the association between a binary outcome (pregnancy outcome) and features of longitudinal measurements (hormone levels) through a common set of latent random effects in **173** women during the first trimester using different modeling strategies.
- These women were classified into two groups:
  - **Normal group** (**124** women who came to term with their pregnancy).
  - **Abnormal group** (**49** women who suffered a loss).
- They are unbalanced data that fluctuate between **1 to 6** observations, having a total of **375** observations.

Chilean women pregnancies data

(β-HCG hormone)

De la Cruz et al. (2011) and De la Cruz et al. (2016)

- They analyzed data from a clinical study on the risk of loss in a group of pregnant Chilean women.
- They model the association between a binary outcome (pregnancy outcome) and features of longitudinal measurements (hormone levels) through a common set of latent random effects in **173** women during the first trimester using different modeling strategies.
- These women were classified into two groups:
  - **Normal group** (**124** women who came to term with their pregnancy).
  - **Abnormal group** (**49** women who suffered a loss).
- They are unbalanced data that fluctuate between **1 to 6** observations, having a total of **375** observations.

Chilean women pregnancies data

(β-HCG hormone)

De la Cruz et al. (2011) and De la Cruz et al. (2016)

- They analyzed data from a clinical study on the risk of loss in a group of pregnant Chilean women.
- They model the association between a binary outcome (pregnancy outcome) and features of longitudinal measurements (hormone levels) through a common set of latent random effects in **173** women during the first trimester using different modeling strategies.
- These women were classified into two groups:
  - **Normal group** (**124** women who came to term with their pregnancy).
  - **Abnormal group** (**49** women who suffered a loss).
- They are unbalanced data that fluctuate between **1 to 6** observations, having a total of **375** observations.

Chilean women pregnancies data

(β-HCG hormone)

De la Cruz et al. (2011) and De la Cruz et al. (2016)

- They analyzed data from a clinical study on the risk of loss in a group of pregnant Chilean women.
- They model the association between a binary outcome (pregnancy outcome) and features of longitudinal measurements (hormone levels) through a common set of latent random effects in **173** women during the first trimester using different modeling strategies.
- These women were classified into two groups:
  - **Normal group** (**124** women who came to term with their pregnancy).
  - **Abnormal group** (**49** women who suffered a loss).
- They are unbalanced data that fluctuate between **1 to 6** observations, having a total of **375** observations.

Chilean women pregnancies data

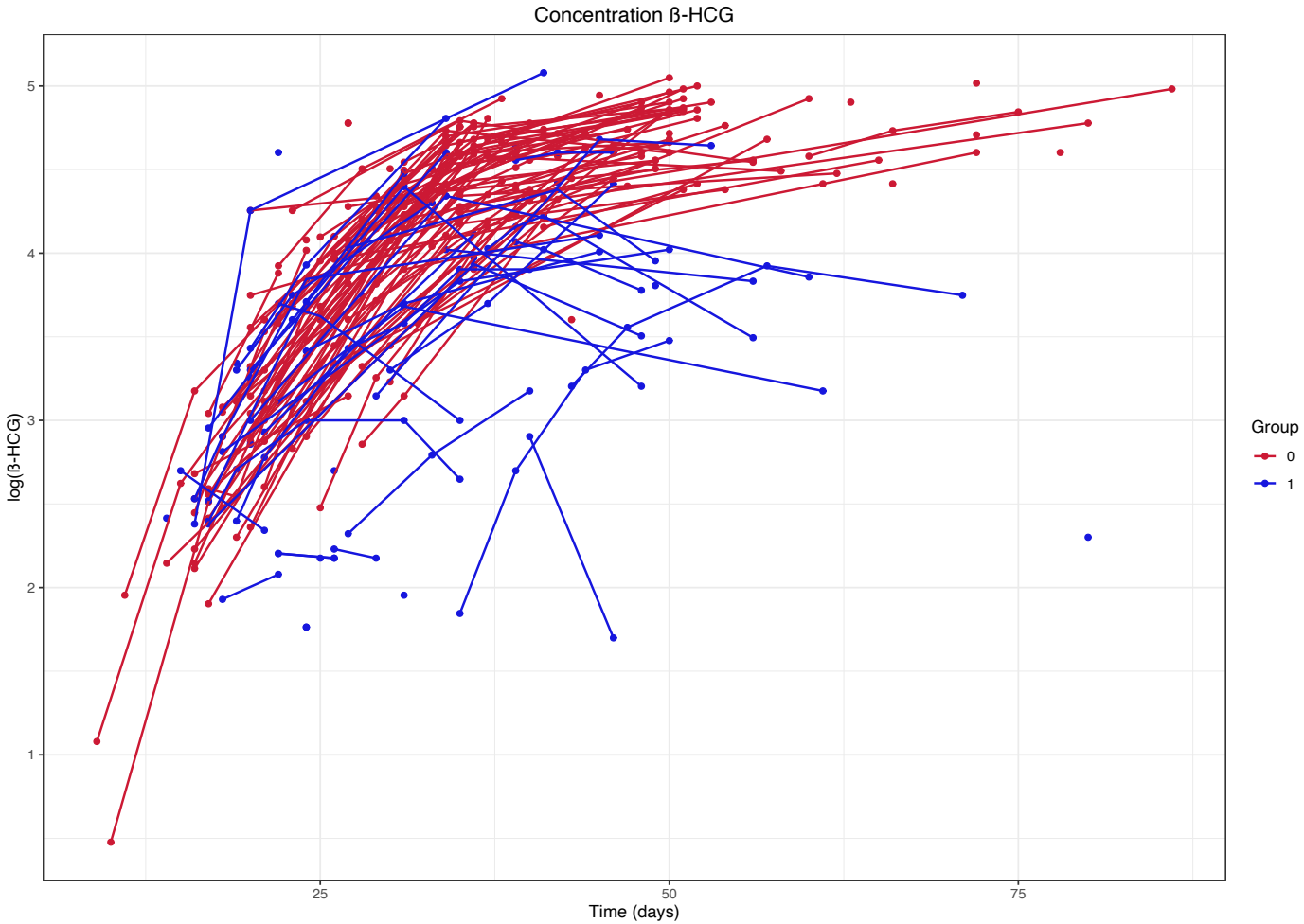

Figure 1: Observed profiles $\log_{10}(\beta - HCG)$.

Chilean women pregnancies data

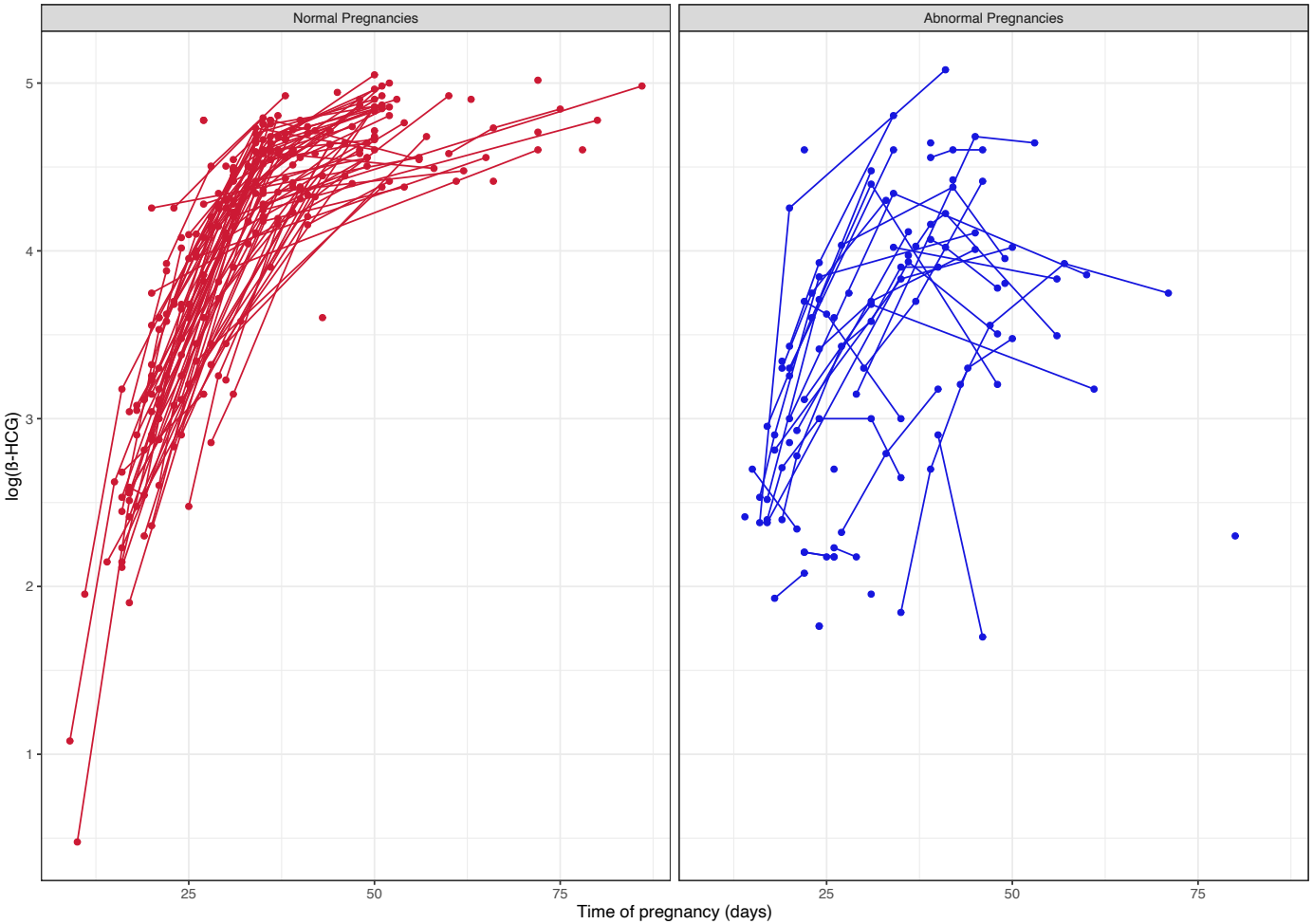

Figure 2: Observed profiles of $\log_{10}(\beta - HCG)$ for normal (left panel), and abnormal groups (right panel).

Radboudumc women pregnancies data

Gestational Trophoblastic Diseases (GTD)

Dandis et al. (2020)

- They analyzed data from the Dutch Central Registry for Hydatidiform Moles at the Radboud University Medical Center (Radboudumc) in Nijmegen.
- They propose four approaches (2SMLE, JMMLE, 2SB, JMB) to predict the risk of a future binary outcome (presence gestational trophoblastic neoplasia (GTN)) based on a repeatedly measured predictor (serum levels of human chorionic gonadotropin (hCG)) in **439** women in a period of two to seven weeks.
- These women were classified into two groups:
  - **Uneventful group** (**299** women).
  - **GTN group** (**140** women with chronic gestacional trophoblastic neoplasia).
- The data fluctuate between **1 to 6** observations, having a total of **1674** observations.

Radboudumc women pregnancies data

Gestational Trophoblastic Diseases (GTD)

Dandis et al. (2020)

- They analyzed data from the Dutch Central Registry for Hydatidiform Moles at the Radboud University Medical Center (Radboudumc) in Nijmegen.
- They propose four approaches (2SMLE, JMMLE, 2SB, JMB) to predict the risk of a future binary outcome (presence gestational trophoblastic neoplasia (GTN)) based on a repeatedly measured predictor (serum levels of human chorionic gonadotropin (hCG)) in **439** women in a period of two to seven weeks.
- These women were classified into two groups:
  - **Uneventful group** (**299** women).
  - **GTN group** (**140** women with chronic gestacional trophoblastic neoplasia).
- The data fluctuate between **1 to 6** observations, having a total of **1674** observations.

Radboudumc women pregnancies data

Gestational Trophoblastic Diseases (GTD)

Dandis et al. (2020)

- They analyzed data from the Dutch Central Registry for Hydatidiform Moles at the Radboud University Medical Center (Radboudumc) in Nijmegen.
- They propose four approaches (2SMLE, JMMLE, 2SB, JMB) to predict the risk of a future binary outcome (presence gestational trophoblastic neoplasia (GTN)) based on a repeatedly measured predictor (serum levels of human chorionic gonadotropin (hCG)) in **439** women in a period of two to seven weeks.
- These women were classified into two groups:
  - **Uneventful group** (**299** women).
  - **GTN group** (**140** women with chronic gestacional trophoblastic neoplasia).
- The data fluctuate between **1 to 6** observations, having a total of **1674** observations.

Radboudumc women pregnancies data

Gestational Trophoblastic Diseases (GTD)

Dandis et al. (2020)

- They analyzed data from the Dutch Central Registry for Hydatidiform Moles at the Radboud University Medical Center (Radboudumc) in Nijmegen.
- They propose four approaches (2SMLE, JMMLE, 2SB, JMB) to predict the risk of a future binary outcome (presence gestational trophoblastic neoplasia (GTN)) based on a repeatedly measured predictor (serum levels of human chorionic gonadotropin (hCG)) in **439** women in a period of two to seven weeks.
- These women were classified into two groups:
  - **Uneventful group** (**299** women).
  - **GTN group** (**140** women with chronic gestacional trophoblastic neoplasia).
- The data fluctuate between **1 to 6** observations, having a total of **1674** observations.

Radboudumc women pregnancies data

Gestational Trophoblastic Diseases (GTD)

Dandis et al. (2020)

- They analyzed data from the Dutch Central Registry for Hydatidiform Moles at the Radboud University Medical Center (Radboudumc) in Nijmegen.
- They propose four approaches (2SMLE, JMMLE, 2SB, JMB) to predict the risk of a future binary outcome (presence gestational trophoblastic neoplasia (GTN)) based on a repeatedly measured predictor (serum levels of human chorionic gonadotropin (hCG)) in **439** women in a period of two to seven weeks.
- These women were classified into two groups:
  - **Unevenful group** (**299** women).
  - **GTN group** (**140** women with chronic gestacional trophoblastic neoplasia).
- The data fluctuate between **1 to 6** observations, having a total of **1674** observations.

Radboudumc women pregnancies data

Gestational Trophoblastic Diseases (GTD)

Dandis et al. (2020)

- They analyzed data from the Dutch Central Registry for Hydatidiform Moles at the Radboud University Medical Center (Radboudumc) in Nijmegen.
- They propose four approaches (2SMLE, JMMLE, 2SB, JMB) to predict the risk of a future binary outcome (presence gestational trophoblastic neoplasia (GTN)) based on a repeatedly measured predictor (serum levels of human chorionic gonadotropin (hCG)) in **439** women in a period of two to seven weeks.
- These women were classified into two groups:
  - **Unevenful group** (**299** women).
  - **GTN group** (**140** women with chronic gestacional trophoblastic neoplasia).
- The data fluctuate between **1 to 6** observations, having a total of **1674** observations.

Radboudumc women pregnancies data

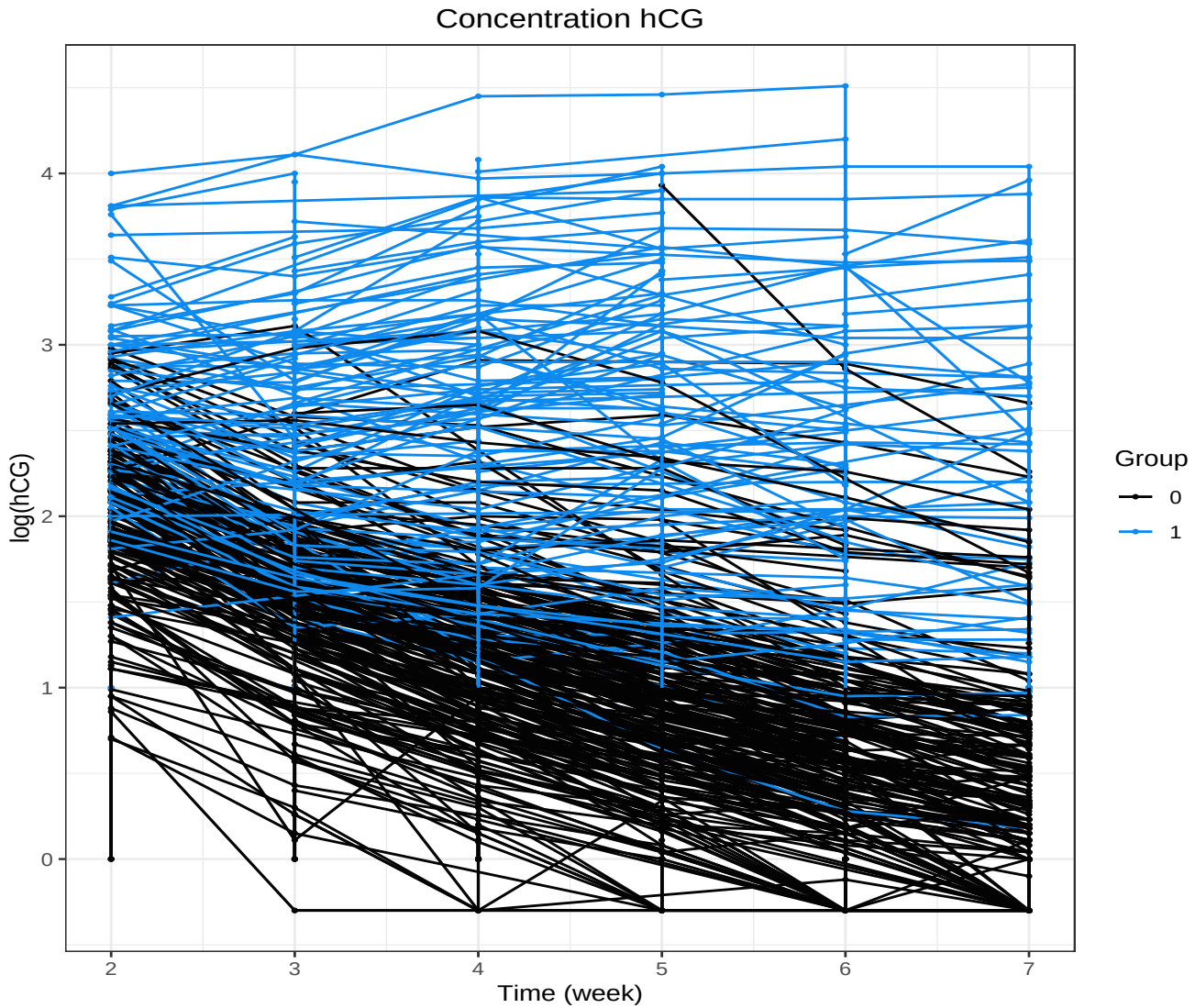

Figure 3: Observed profiles log $-$*transformed*(*hCG*).

Radboudumc women pregnancies data

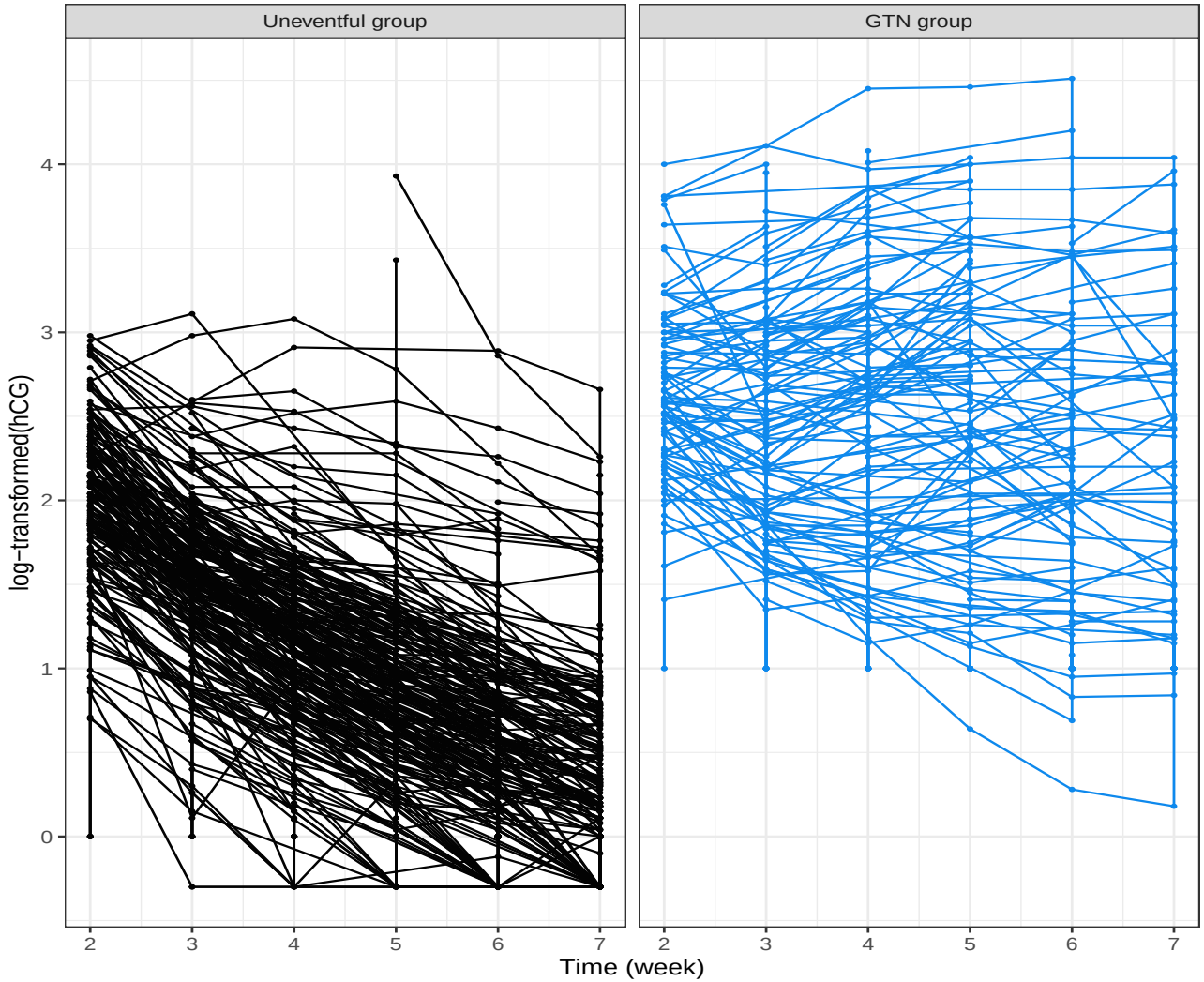

Figure 4: Observed profiles of log $-transformed(hCG)$: on the left panel is the uneventful group; and on the right panel, the group who experience GTN.

Introduction

We propose:

- A a joint model based on an NLME model for the longitudinal part taking several random effects as covariates in a submodel GLM for the primary response of interest (De la Cruz et al., 2016; Dandis et al., 2020).
- The resulting joint model (NLME/GLM) is estimated using a new estimation method based on the likelihood, employing a stochastic approximation version of the EM algorithm, the so-called SAEM algorithm (Delyon et al., 1999; Kuhn and Lavielle, 2005).
- We made classification into two groups.

Model Formulation

Joint Model (longitudinal part)

Let y_{ij}, the measured concentration of the hormone for the i-th woman at time t_{ij}.

NLME

$$y_{ij} = \mu(t_{ij}; \boldsymbol{\phi}_i) + \nu(t_{ij}, \boldsymbol{\phi}_i, \xi)\varepsilon_{ij}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq n_i \qquad (1)$$
$$\boldsymbol{\phi}_i = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{W}_{ij}\boldsymbol{\beta}_i, \quad \boldsymbol{\beta}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}),$$

- $\boldsymbol{\beta}$ is a vector unknown fixed effects parameters.
- $\boldsymbol{\beta}_i$ is a vector unobservable random effects.
- μ is a nonlinear function.
- $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\beta}_i$ and ε_{ij}'s are mutually independent.
- $\nu(\cdot)$ is a function that models the variability of the residual error which depends on some additional vector of parameters ξ.

Joint Model (longitudinal part)

Consider the case where the function ν is expressed as a function of the structural model μ, i.e.,

$$\nu(t_{ij}, \boldsymbol{\phi}_i, \xi) = \nu(\mu(t_{ij}, \boldsymbol{\phi}_i), \xi),$$

And so it is:

$$y_{ij} = \mu\left(t_{ij}; \boldsymbol{\phi}_i\right) + \nu(\mu(t_{ij}, \boldsymbol{\phi}_i), \xi)\epsilon_{ij}. \tag{2}$$

Joint Model (variability of the residual error)

- Residual Error Model I (**REM I**): $y = \mu + a\epsilon$. Where the function v is constant, and the additional parameter is $\xi = a$.
- Residual Error Model II (**REM II**): $y = \mu + b\mu^c\epsilon$. Such that, the function v is proportional to the structural model μ, and the additional parameters are $\xi = (b, c)$. By default, the parameter c is fixed at 1 and the additional parameter is $\xi = b$.
- Residual Error Model III (**REM III**): $y = \mu+(a+b\mu^c)\epsilon$. In the case, function v is a linear combination of a constant term and a term proportional to the structural model μ, and the additional parameters are $\xi = (a, b)$ (by default, the parameter c is fixed at 1).
- Residual Error Model IV (**REM IV**): $y = \mu+\sqrt{(a^2 + b^2\mu^{2c})}\epsilon$. The function v is a combination of a constant term and a term proportional to the structural model μ ($v = b\mu^c$), and the additional parameters are $\xi = (a, b)$ (by default, the parameter c is fixed at 1).

Joint Model (binary part)

We consider a primary response observed D_i for the i-th individual.

This primary response and the random effects are related through a GLM such that the distribution of D_i given $\boldsymbol{\beta}_i$ is:

$$P\left(D_i|\boldsymbol{\beta}_i;\boldsymbol{\theta}\right) = \exp\left\{\frac{D_i\left(\eta'\boldsymbol{\beta}_i\right) - \alpha_2\left(\eta'\boldsymbol{\beta}_i\right)}{\alpha_1\left(\tau\right)} + \alpha_3\left(D_i,\tau\right)\right\}, \qquad (3)$$

- $\boldsymbol{\theta} = (\eta', \tau)$ such that η' is the parameter of primary interest, τ is a dispersion parameter.
- $\alpha_1(\cdot), \alpha_2(\cdot)$ and $\alpha_3(\cdot)$ are known functions.

Joint Model (binary part)

As discussed Wang et al. (2000), we can further assume that y_{ij} and D_i are conditionally independent given $\boldsymbol{\beta}_i$,

$$\begin{aligned} P\left(y_{ij}, D_i, \boldsymbol{\beta}_i\right) &= P\left(y_{ij}, D_i|\boldsymbol{\beta}_i\right) P\left(\boldsymbol{\beta}_i\right) \\ &= P\left(y_{ij}|\boldsymbol{\beta}_i\right) P\left(D_i|\boldsymbol{\beta}_i\right) P\left(\boldsymbol{\beta}_i\right), \end{aligned} \tag{4}$$

- $P(y_{ij}|\boldsymbol{\beta}_i)$ is the normal density function of $y_{ij}|\beta i$.
- $P(D_i|\boldsymbol{\beta}_i)$ is the Bernoulli distribution of $D_i|\boldsymbol{\beta}_i$.
- $P(\boldsymbol{\beta}_i)$ is the normal density function of $\boldsymbol{\beta}_i$.

Joint Model (Likelihood)

The log-likelihood for the joint model $\left(y_{ij}, D_i\right)$ is given by

$$\mathcal{L}\left(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{D}\right) = \sum_{i=1}^{N} \log \int_{\mathbb{R}^q} P\left(y_{ij}|\boldsymbol{\beta}_i\right) P\left(D_i|\boldsymbol{\beta}_i\right) P\left(\boldsymbol{\beta}_i\right) d\boldsymbol{\beta}_i, \tag{5}$$

where $\boldsymbol{y} = \left(y_{1j}, \ldots, y_{Nj}\right)$ wich $1 \leq j \leq n_i$ and $\boldsymbol{D} = (D_1, \ldots, D_N)$.

Estimation via SAEM algorithm

For the non-observed data $\boldsymbol{\psi} = \boldsymbol{\beta}_i$ and the observed data $\boldsymbol{\mathcal{Y}} = (y_{ij}, D_i)$, the likelihood $(\boldsymbol{\mathcal{Y}}, \psi; \theta)$ was maximized with respect to $\boldsymbol{\theta}$ using the **SAEM algorithm** (Delyon et al., 1999; Kuhn and Lavielle, 2004). This algorithm replaces the usual E-step of EM by a stochastic procedure.

It is a robust alternative to Lindstrom and Bates (1990) algorithm (nlme library in R) and implementation can be found in the R package saemix or in the **Monolix software** (https://lixoft.com/).

Then, at iteration k, the SAEM algorithm proceeds as follows:

- **Simulation step:** draw $\psi^{(k)}$ from the conditional distribution $p(\cdot|\boldsymbol{\mathcal{Y}}, \theta^{(k)})$.
- **Stochastic approximation step:** update $Q_k(\theta)$ according to:

$$Q_k(\theta) = Q_{k-1}(\theta) + \lambda_k \left(\log \ell(\boldsymbol{\mathcal{Y}}, \psi; \theta) - Q_{k-1}(\theta) \right),$$

  where $Q_k(\theta) = \mathbb{E}\left[\log \ell(\boldsymbol{\mathcal{Y}}, \psi; \theta) | \boldsymbol{\mathcal{Y}}, \theta_{(k-1)}\right]$ and λ_k is a parameter used to accelerate convergence (Kuhn and Lavielle, 2005).
- **Maximization step:** updated $\theta^{(k)}$ according to

$$\theta_{(k+1)} = \arg\max_{\theta} Q_k(\theta).$$

Estimation via SAEM algorithm

For the non-observed data $\boldsymbol{\psi} = \boldsymbol{\beta}_i$ and the observed data $\boldsymbol{\mathcal{Y}} = (y_{ij}, D_i)$, the likelihood $(\boldsymbol{\mathcal{Y}}, \psi; \theta)$ was maximized with respect to $\boldsymbol{\theta}$ using the **SAEM algorithm** (Delyon et al., 1999; Kuhn and Lavielle, 2004). This algorithm replaces the usual E-step of EM by a stochastic procedure.

It is a robust alternative to Lindstrom and Bates (1990) algorithm (nlme library in R) and implementation can be found in the R package saemix or in the **Monolix software** (`https://lixoft.com/`).

Then, at iteration k, the SAEM algorithm proceeds as follows:

- **Simulation step:** draw $\psi^{(k)}$ from the conditional distribution $p(\cdot|\boldsymbol{\mathcal{Y}}, \theta^{(k)})$.
- **Stochastic approximation step:** update $Q_k(\theta)$ according to:

$$Q_k(\theta) = Q_{k-1}(\theta) + \lambda_k \left(\log \ell(\boldsymbol{\mathcal{Y}}, \psi; \boldsymbol{\theta}) - Q_k, (\boldsymbol{\theta})\right),$$

  where $Q_k(\theta) = \mathbb{E}\left[\log \ell(\boldsymbol{\mathcal{Y}}, \psi; \boldsymbol{\theta}) | \boldsymbol{\mathcal{Y}}, \theta_{(k-1)}\right]$ and λ_k is a parameter used to accelerate convergence (Kuhn and Lavielle, 2005).
- **Maximization step:** updated $\theta^{(k)}$ according to

$$\theta_{(k+1)} = arg \max_{\theta} Q_k(\theta).$$

Estimation via SAEM algorithm

For the non-observed data $\boldsymbol{\psi} = \boldsymbol{\beta}_i$ and the observed data $\boldsymbol{\mathcal{Y}} = (y_{ij}, D_i)$, the likelihood $(\boldsymbol{\mathcal{Y}}, \psi; \theta)$ was maximized with respect to $\boldsymbol{\theta}$ using the **SAEM algorithm** (Delyon et al., 1999; Kuhn and Lavielle, 2004). This algorithm replaces the usual E-step of EM by a stochastic procedure.

It is a robust alternative to Lindstrom and Bates (1990) algorithm (nlme library in R) and implementation can be found in the R package saemix or in the **Monolix software** (`https://lixoft.com/`).

Then, at iteration k, the SAEM algorithm proceeds as follows:

- **Simulation step:** draw $\psi^{(k)}$ from the conditional distribution $p(\cdot|\boldsymbol{\mathcal{Y}}, \theta^{(k)})$.
- **Stochastic approximation step:** update $Q_k(\theta)$ according to:

$$Q_k(\theta) = Q_{k-1}(\theta) + \lambda_k \left(\log \ell(\boldsymbol{\mathcal{Y}}, \boldsymbol{\psi}; \boldsymbol{\theta}) - Q_k, (\boldsymbol{\theta})\right),$$

  where $Q_k(\theta) = \mathbb{E}\left[\log \ell(\boldsymbol{\mathcal{Y}}, \boldsymbol{\psi}; \boldsymbol{\theta}) | \boldsymbol{\mathcal{Y}}, \boldsymbol{\theta}_{(k-1)}\right]$ and λ_k is a parameter used to accelerate convergence (Kuhn and Lavielle, 2005).
- **Maximization step:** updated $\theta^{(k)}$ according to

$$\theta_{(k+1)} = arg \max_{\theta} Q_k(\theta).$$

Estimation via SAEM algorithm

For the non-observed data $\boldsymbol{\psi} = \boldsymbol{\beta}_i$ and the observed data $\boldsymbol{\mathcal{Y}} = (y_{ij}, D_i)$, the likelihood $(\boldsymbol{\mathcal{Y}}, \psi; \theta)$ was maximized with respect to $\boldsymbol{\theta}$ using the **SAEM algorithm** (Delyon et al., 1999; Kuhn and Lavielle, 2004). This algorithm replaces the usual E-step of EM by a stochastic procedure.

It is a robust alternative to Lindstrom and Bates (1990) algorithm (nlme library in R) and implementation can be found in the R package saemix or in the **Monolix software** (`https://lixoft.com/`).

Then, at iteration k, the SAEM algorithm proceeds as follows:

- **Simulation step:** draw $\psi^{(k)}$ from the conditional distribution $p(\cdot|\boldsymbol{\mathcal{Y}}, \theta^{(k)})$.
- **Stochastic approximation step:** update $Q_k(\theta)$ according to:

$$Q_k(\theta) = Q_{k-1}(\theta) + \lambda_k \left(\log \ell(\boldsymbol{\mathcal{Y}}, \boldsymbol{\psi}; \boldsymbol{\theta}) - Q_k, (\boldsymbol{\theta})\right),$$

  where $Q_k(\theta) = \mathbb{E}\left[\log \ell(\boldsymbol{\mathcal{Y}}, \boldsymbol{\psi}; \boldsymbol{\theta}) | \boldsymbol{\mathcal{Y}}, \boldsymbol{\theta}_{(k-1)}\right]$ and λ_k is a parameter used to accelerate convergence (Kuhn and Lavielle, 2005).
- **Maximization step:** updated $\boldsymbol{\theta^{(k)}}$ according to

$$\theta_{(k+1)} = arg \max_{\theta} Q_k(\theta).$$

Estimation via SAEM algorithm

Kuhn and Lavielle (2005) propose to combine the SAEM with a Markov chain Monte Carlo (MCMC) procedure when the simulation step cannot be directly performed, as for instance in the NLME.

Application 1

Prediction of miscarriage in first trimester by serum β-HCG

The representation of the β-HCG levels for the i-th woman is:

$$y_{ij} = \frac{a_i}{1 + \exp\left[-(t_{ij} - b_i)/\theta\right]} + \nu(\mu(t_{ij}, \boldsymbol{\phi}_i), \xi)\epsilon_{ij}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq n_i, \tag{6}$$

$$y_{ij} = \frac{a_i}{1 + \exp\left[-(t_{ij} - b_i)/c_i\right]} + \nu(\mu(t_{ij}, \boldsymbol{\phi}_i), \xi)\epsilon_{ij}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq n_i, \tag{7}$$

We consider that the random effects $\boldsymbol{\phi}_i$ follow a normal distribution with mean $\boldsymbol{\mu} = (a_{pop}, b_{pop}, c_{pop})$ and variance-covariance matrix $\boldsymbol{\Gamma} = diag(\sigma_a^2, \sigma_b^2, \sigma_c^2)$.

Prediction of miscarriage in first trimester by serum β-HCG

We also consider the longitudinal model with log-normal random effects.

$$y_{ij} = \frac{a_i}{1 + \exp\left[-(t_{ij} - b_i)/c_i\right]} + \nu(\mu(t_{ij}, \boldsymbol{\phi}_i), \xi)\epsilon_{ij}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq n_i, \tag{8}$$

$$\begin{aligned} log(a_i) &= log(a_{pop}) + \eta_{i1}, \quad \text{where} \quad \eta_{i1} \sim N(0, \sigma_a^2) \\ log(b_i) &= log(b_{pop}) + \eta_{i2}, \quad \text{where} \quad \eta_{i2} \sim N(0, \sigma_b^2) \\ log(c_i) &= log(c_{pop}). \end{aligned}$$

And $\nu(\mu(t_{ij}, \boldsymbol{\phi}_i), \xi)$ denotes the error structure according to **REM I**, **REM II**, **REM III**, and **REM IV**.

Prediction of miscarriage in first trimester by serum β-HCG

We consider the problem of predicting a loss (abnormal pregnancy) in the set of pregnant women.

Let $\boldsymbol{D} = \{0, 1\}$ denote normal and abnormal pregnancy outcomes, respectively,

The relationship between pregnancy outcomes and the asymptotic levels of β-HCG, follow the primary logistic regression models:

$$\mathbb{P}(D_i = 1|a_i) = \frac{1}{1 + \exp\{-(\eta_0 + \eta_1 a_i)\}}. \tag{9}$$

And

$$\mathbb{P}(D_i = 1|a_i, b_i) = \frac{1}{1 + \exp\{-(\eta_0 + \eta_1 a_i + \eta_2 b_i)\}} \tag{10}$$

Results 1

Results 1

Joint Model		Model (7)-(9)			Model (6)-(10)			Model (8)-(10)	
Parameters	**Estimate**	**S.E**	**R.S.E (%)**	**Estimate**	**S.E**	**R.S.E (%)**	**Estimate**	**S.E**	**R.S.E (%)**
a_{pop}	4.5534	0.05412	1.19	4.5403	0.0512	1.13	4.5456	0.04916	1.08
b_{pop}	15.6772	0.527	3.36	15.6176	0.545	3.49	15.6176	0.5733	3.82
c_{pop}	7.2885	0.5171	7.09	6.9984	0.4153	5.93	7.1844	0.4638	6.45
η_{0pop}	32.0155	12.7417	39.8	28.2743	10.1912	36.00	46.9476	74.805	159
η_{1pop}	−7.3993	2.8773	38.9	−6.5697	2.2996	35	−11.0641	16.8676	152
η_{2pop}	–	–	–	$2.63E-07$	0.001601	$6.09E+05$	0.08916	0.108	121
SD of the Random Effects									
ω_a	0.4952	0,0679	13,7	0.4682	0.04066	8.68	0.07939	0.01423	17.9
ω_b	3,604	1,7918	49,7	4.354	0.4352	10	9.7835	0.02983	10.5
ω_c	1,884	0,7472	39,7	–	–	–	–	–	–
Error Model Parameters									
a	0.2537	0.03008	11.9	0.2659	0.01825	6.86	0.2999	0.02639	8.8
$-2 \times log - likelihood$	657.2478			660,902			669.4938		
AIC	675.2478			678.902			687,4938		
BIC	703,6274			707.2817			715,8734		
BICc	710,5453			715.3525			723,9443		

Table 1: Parameter estimates for the pregnant women data using the SAEM algorithm .

Results 1

Group	Model (7)-(9)		Model (6)-(10)		Model (8)-(10)		**Total**
	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Total (173)
	Within sample						
Normal	123	1	123	1	124	0	124
Abnormal	9	40	10	39	8	41	49
	Leave-one-out CV						
Normal	124	0	124	0	124	0	124
Abnormal	8	41	2	47	3	46	49

Table 2: Classification in two groups using the SAEM algorithm.

Results 1

Accuracy Metrics	
Metrics	Model (7)-(9)
Error rate	**0.058**
Sensitivity	**0.992**
Specificity	**0.816**
Precision	**0.932**
Accuracy	**0.942**

Table 3: Accuracy metrics for the joint model (7)-(9) estimated using the SAEM algorithm.

Application 2

Predictions of post–molar gestational trophoblastic neoplasia

Let $\log(hCG)_{ij}$ represent the log-transformed hCG longitudinal measurements for patient i, $i = 1, \ldots, 439$, at week $t_{ij} = 2, \ldots, 7$ and at the age AGE_i. The model for the first part can be written as follows:

$$\log(hCG)_{ij} = \mu(b_i, t_{ij}) + \nu(\mu(t_{ij}, \boldsymbol{\phi}_i), \xi)\varepsilon_{ij} \tag{11}$$

where

$$\begin{aligned} \mu(t_{ij}, \boldsymbol{\phi}_i) &= a_i + b_i \times t_{ij} \\ \boldsymbol{\phi}_i = (a_i, b_i)^T &\sim \mathcal{N}(\mu_\phi, \Gamma) \text{ with } \mu_\phi = \begin{pmatrix} a_{pop} \\ b_{pop} \end{pmatrix} \text{ and } \Gamma = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

and $\nu(\mu(t_{ij}, \boldsymbol{\phi}_i), \xi)$ denotes the error structure according to **REM I**, **REM II**, **REM III**, and **REM IV**.

Predictions of post–molar gestational trophoblastic neoplasia

The second model considers here use as predictors in a logistic regression model with the status of GTN as the outcome:

$$\text{logit}\left(\text{P}\left(\text{GTN}_i = 1\right)\right) = \alpha_0 + \alpha_1 \times a_i + \alpha_2 \times b_i + \alpha_3 \times AGE_i, \quad (12)$$

where GTN_i reflects the GTN status of the i-th patient, and $\alpha = [\alpha_0, \alpha_1, \alpha_2, \alpha_3]$ is the vector of the logistic regression coefficients. The coefficients α_1 and α_2 reflect the strength of association between the two models.

Results 2

Results 2

Joint Model	Model 11-12 Residual Error Model **REM I**			Model 11-12 Residual Error Model **REM IV**		
Parameters	**Estimate**	**S.E**	**R.S.E (%)**	**Estimate**	**S.E**	**R.S.E (%)**
a_{pop}	2.5	0.034	1.34	2.50	0.034	1.36
b_{pop}	−0.22	0.0096	4.47	−0.22	0.0094	4.36
α_0	−1.66	1.51	90.8	−1.46	1.58	108
α_1	1.77	0.43	24.0	1.78	0.44	24.6
α_2	23.96	3.22	13.4	25.36	3.61	14.20
α_3	0.025	0.028	110	0.026	0.028	108
Variance components						
σ_a	0.59	0.03	5.03	0.6	0.03	4.99
σ_b	0.18	0.0078	4.28	0.18	0.0081	4.55
σ_{ab}	−0.091	0.061	67.2	−0.075	0.063	83.7
Error Model Parameters						
a	0.19	0.0045	2.43	0.16	0.0064	3.88
b	-	-	-	0.052	0.0067	12.9
$-2 \times log - likelihood$	1838.78			1818.13		
AIC	1858.78			1840.13		
BIC	1899.63			1885.06		
BICc	1910.63			1897.63		

Table 4: Parameter estimates of the models predicting GTN status using the SAEM algorithm.

Results 2

Group	REM I		REM IV		Total
	Within sample				
	GTN	No GTN	GTN	No GTN	Total (439)
GTN	121	19	122	18	140
No GTN	12	287	10	289	299
	Leave-one-out CV				
	GTN	No GTN	GTN	No GTN	Total (439)
GTN	121	19	122	18	140
No GTN	13	286	12	287	299

Table 5: Classification of the patients based on the available hCG measurements using the SAEM algorithm.

Results 2

Accuracy Metrics		
Metrics	Model **REM I**	Model **REM IV**
Error rate	0.0729	0.0683
Sensitivity	0.9377	0.9410
Specificity	0.903	0.9104
Precision	0.8643	0.8714
Accuracy	0.9271	0.9317

Table 6: Accuracy metrics for the joint model (11)-(12) with error structure **REM I**, and **REM IV** estimated using the SAEM algorithm.

Final Comments

Final Comments

- We proposed joint models (NLME/GLM) with several random effects and different distributions. Modeling different error structures.
- These models were estimated using the SAEM algorithm and we have classified them into two groups.

Acknowledgements

Acknowledgements

This work was funded by the Data Observatory Foundation, ANID Technology Center No. DO210001. This work was partially funded by grant ANID/PIA/ANILLOS ACT210096.

References

References

Dandis, R., Teerenstra, S., Massuger, L., Sweep, F., Eysbouts, Y., and IntHout, J. (2020). A tutorial on dynamic risk prediction of a binary outcome based on a longitudinal biomarker. *Biometrical Journal*, 62(2):398–413.

De la Cruz, R., Marshall, G., and Quintana, F. A. (2011). Logistic regression when covariates are random effects from a non-linear mixed model. *Biometrical journal*, 53(5):735–749.

De la Cruz, R., Meza, C., Arribas-Gil, A., and Carroll, R. J. (2016). Bayesian regression analysis of data with random effects covariates from nonlinear longitudinal measurements. *Journal of multivariate analysis*, 143:94–106.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM: Probability and Statistics*, 8:115–131.

Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, 49(4):1020–1038.

References

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687.

Wang, C., Wang, N., and Wang, S. (2000). Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*, 56(2):487–495.

The end

Thank you!

”Nothing in life is to be feared.
It is only to be understood.
Now is the time to understand more,
so that we may fear less.”

Marie Curie