Distributional regression for fitting heterogeneous longitudinal response

Statistiques au sommet de Rochebrune 2026

A. Barbieri[1], A. Alcaraz[2], M. Abed[1], H. De Courson[1,3] and H. Jacqmin-Gadda[1]

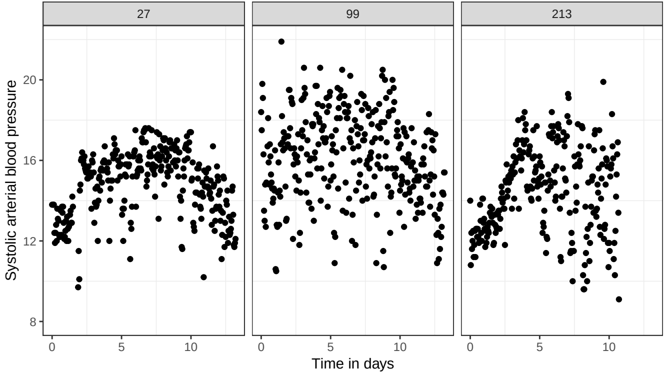
[1]Univ. Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France
[2]Univ Bretagne Sud, CNRS UMR 6205, LMBA, France
[3]CHU de Bordeaux, Anesthesiology and Critical Care Department, INSERM, U1219, Bordeaux, France

Clinical motivation

- Explain and predict a cardio and cerebro-vascular events
  - Joint modeling for longitudinal and time-to-event data
- Focus on blood pressure
  - A modifiable and known risk factor
  - Recent risk factor hypotheses
    - Individual variability of measurements
    - Probability of being in a given interval (e.g. under a specific threshold)
  - Fit individual conditional distribution over time
- Data collected from 201 patients hospitalized in the intensive care unit in Bordeaux Hospital after a subarachnoid hemorrhage
  - Time-independent covariates (e.g. sex, age...)
  - Intensive biomarker measures (e.g. blood pressure)

Blood pressure measurements

- Heterogeneous longitudinal response
  - Presence of heteroskedasticity, asymmetry, *outliers*



GAMLSS approach

Generalized Additive Models for Location Scale and Shape (Kneib et al. 2023)

- Fit the conditional distribution of a response variable
  - Assume a probability distribution for the response variable
  - Each distribution parameter is modeled by a function of covariates, time and subject-specific random effect
  - Find the best balance between flexibility and interpretation

- Location-scale mixed model (LSMM) for dealing with heteroskedasticity (Courcoul et al. 2025)
  - For individual $i = 1, \ldots, n$ and its measurement $j = 1, \ldots, n_i$

$$Y_{ij} | \mathbf{r}_i \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}) \quad \text{with} \quad \begin{cases} \mu_{ij} & = \mathbf{x}_{ij,\mu}^\top \boldsymbol{\beta} + \mathbf{z}_{ij,\mu}^\top \mathbf{b}_i \\ log(\sigma_{ij}) & = \mathbf{x}_{ij,\sigma}^\top \boldsymbol{\xi} + \mathbf{z}_{ij,\sigma}^\top \mathbf{u}_i \end{cases}$$

  - ⋆ $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ the fixed parameter vectors
  - ⋆ With vector of random effects $\mathbf{r}_i = \left(\mathbf{b}_i^\top, \mathbf{u}_i^\top\right)^\top \sim \mathcal{N}(\mathbf{0}, \Sigma_r)$

Illustration of LSMM

Individual conditional distributions over time

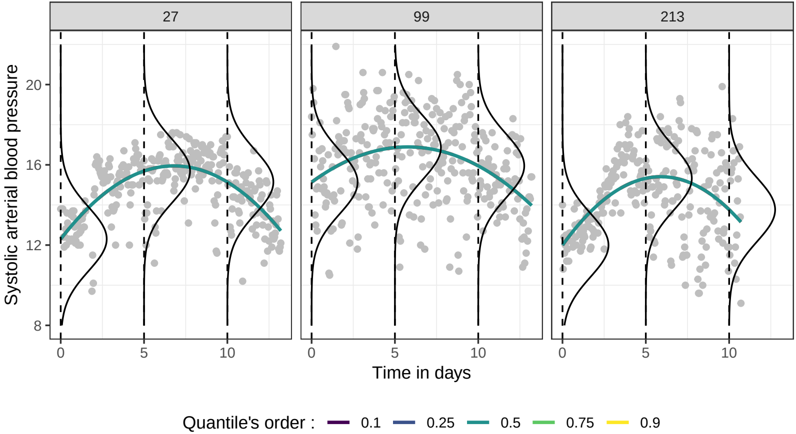


Illustration of LSMM

Individual conditional distributions over time

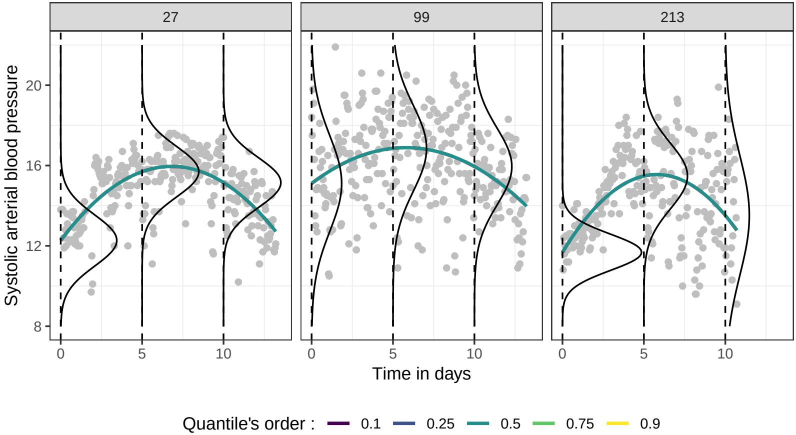
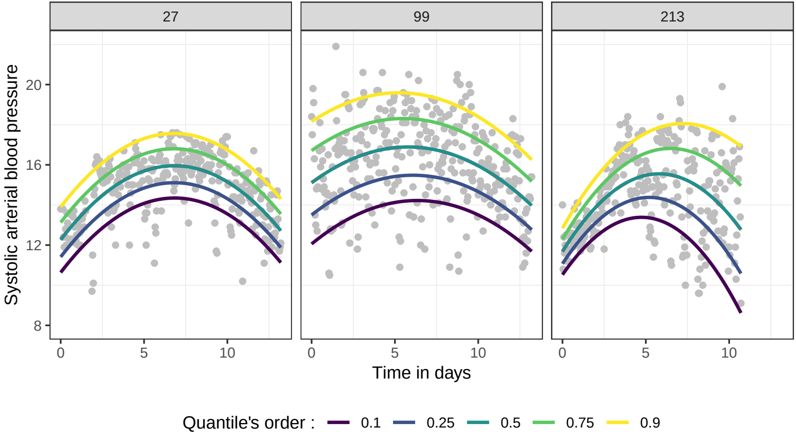


Illustration of LSMM

Individual conditional distributions over time



Quantile regression approach

(Koenker, 2005; Geraci, 2014)

- Focus on conditional quantiles the distribution of interest
  - A model for each fitted quantile
  - Explore the entire conditional distribution (asymmetry, heteroskedasticity)
  - Robust to outliers and flexible fit
- Linear quantile mixed model (LQMM)
  - Inference naturally based on Asymmetric Laplace ($\mathcal{AL}$) distribution

$$Y_{ij}|\mathbf{b}_i \sim \mathcal{AL}\left(\mu_{ij}, \sigma, \tau\right) \quad \text{with} \quad \mu_{ij} = \mathbf{x}_{ij,\mu}^\top \boldsymbol{\beta} + \mathbf{z}_{ij,\mu}^\top \mathbf{b}_i,$$

  - μ_{ij} the location parameter and the conditional quantile of order τ
  - $\sigma \in \mathbb{R}^+$ the scale parameter
  - $\tau \in]0, 1[$ a skewness parameter
    - Pre-fixed by user to the order of the quantile of interest

Illustration of LQMM

Conditional quantile trajectories over time

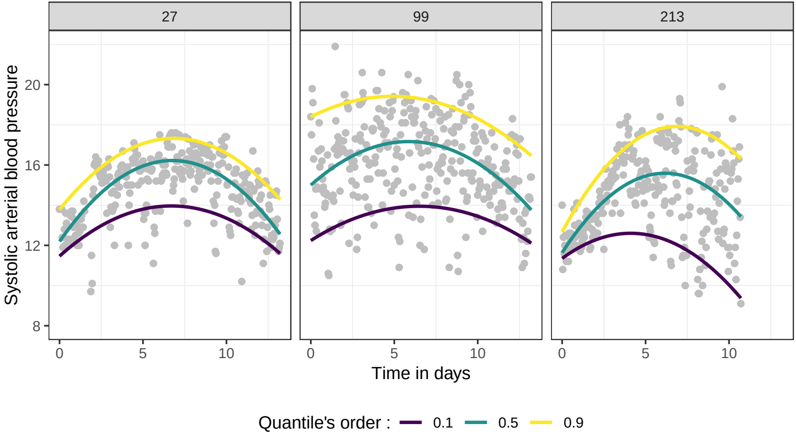
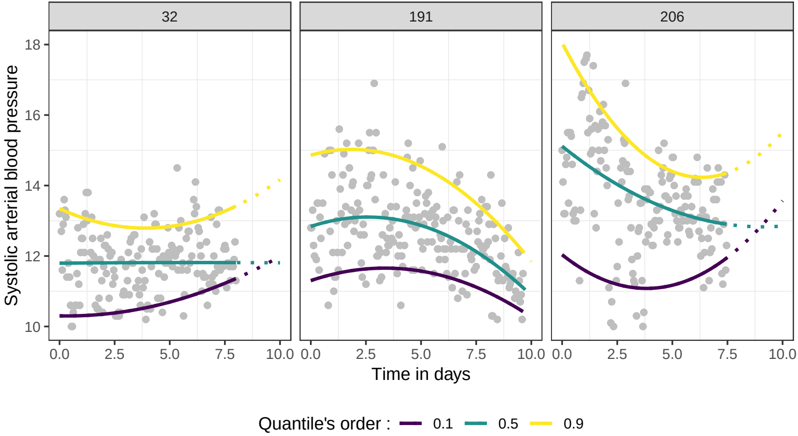


Illustration of LQMM

Conditional quantile trajectories over time



$\mathcal{AL}$ distribution regression model (ALDRM)

- GAMLSS approach using $\mathcal{AL}$ distribution
  - Each distribution parameter is modeled by a function of covariates, time and subject-specific random effects
  - Distribution driven by data
  - For individual $i = 1, \ldots, n$ and its measurement $j = 1, \ldots, n_i$

$$Y_{ij}|\mathbf{r}_i \sim \mathcal{AL}(\mu_{ij}, \sigma_{ij}, \tau_{ij}) \quad \text{with} \quad \begin{cases} \mu_{ij} & = \mathbf{x}_{ij,\mu}^{\top}\boldsymbol{\beta} + \mathbf{z}_{ij,\mu}^{\top}\mathbf{b}_i \\ log(\sigma_{ij}) & = \mathbf{x}_{ij,\sigma}^{\top}\boldsymbol{\xi} + \mathbf{z}_{ij,\sigma}^{\top}\mathbf{u}_i \\ logit(\tau_{ij}) & = \mathbf{x}_{ij,\tau}^{\top}\boldsymbol{\alpha} + \mathbf{z}_{ij,\tau}^{\top}\mathbf{a}_i \end{cases}$$

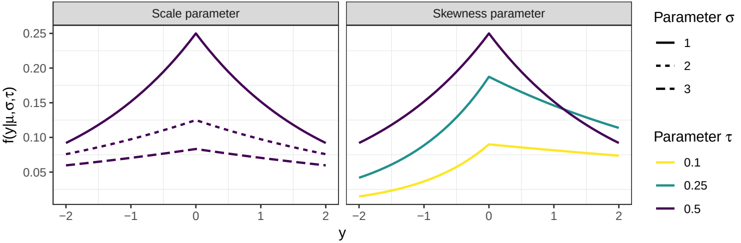
  - ⋆ $\boldsymbol{\beta}$, $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}$ the fixed parameter vectors
  - ⋆ $\mathbf{r}_i$ vector of subject-specific random effects

$$\mathbf{r}_i = \begin{pmatrix} \mathbf{b}_i \\ \mathbf{u}_i \\ \mathbf{a}_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \Sigma_r = \begin{pmatrix} \Sigma_b & 0 & 0 \\ 0 & \Sigma_u & 0 \\ 0 & 0 & \Sigma_a \end{pmatrix}\right).$$

Why $\mathcal{AL}$ distribution?

Denoted by $Y \sim \mathcal{AL}(\mu, \sigma, \tau)$ (Yu et al. 2005)

- Flexible distribution and intuitive interpretation of parameters
  - $\mu \in \mathbb{R}$ is the mode
  - $\sigma \in \mathbb{R}^+$ is the dispersion around the mode
  - $\tau \in]0, 1[$ is the skewness around the mode



- Explicit features given the distributional parameters
  - Cumulative distribution function $F_{Y_{ij}|\mathbf{r}_i}(y)$
  - Inverse cumulative distribution function $\mathcal{Q}_{Y_{ij}|\mathbf{r}_i}(\gamma)$ with $\gamma \in]0, 1[$
  - Expectation and variance

Estimation procedure

Some details

- Bayesian estimation

$$\pi\left(\boldsymbol{\theta}, \mathbf{r}, \boldsymbol{w} | \boldsymbol{y}\right) \quad \propto \quad \prod_{i=1}^{n} \prod_{j=1}^{n_i} f\left(y_{ij} | \mathbf{r}_i, w_{ij}, \boldsymbol{\theta}\right) f\left(w_{ij} | \boldsymbol{\theta}\right) \phi\left(\mathbf{r}_i | \Sigma_r\right) \pi\left(\boldsymbol{\theta}\right)$$

  - Use the rewriting of the asymmetrical Laplace distribution
  - θ the vector of all parameters and prior distribution $\pi\left(\boldsymbol{\theta}\right)$
    - Vague priors
    - Gaussian distributions for fixed effects
    - Inverse-Wishart distribution for covariance matrices of random effects (or Inverse-Gamma for variance parameter)
  - Use the JAGS software to generate posterior samples of parameters
  - Convergence assessed using Gelman-Rubin criterion and parameter trace plots

Application - ALDRM

- Data collected from 201 patients hospitalized in the intensive care unit in Bordeaux Hospital
  - Time-independent covariates (e.g. sex, age...)
  - Intensive biomarker measures (e.g. blood pressure)
    - Total of 42,450 observations
    - Number of measurements per patient ranges from 10 to 322 (median: 224)

- Model specification with $Y_{ij}|\mathbf{b}_i, \mathbf{u}_i, \mathbf{a}_i \sim \mathcal{AL}(\mu_{ij}, \sigma_{ij}, \tau_{ij})$
  - $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_b)$, $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_u)$ and $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_a)$

$$
\begin{aligned}
\mu_{ij} &= \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 Age_i + \beta_4 Sexe_i \\
& \quad b_{i0} + b_{i1} t_{ij} + b_{i2} t_{ij}^2 \\
log(\sigma_{ij}) &= \xi_0 + \xi_1 t_{ij} + \xi_2 Age_i \\
& \quad u_{i0} + u_{i1} t_{ij} \\
logit(\tau_{ij}) &= \alpha_1 Age_i + \alpha_2 Sexe_i \\
& \quad a_{i0} + a_{i1} t_{ij}
\end{aligned}
$$

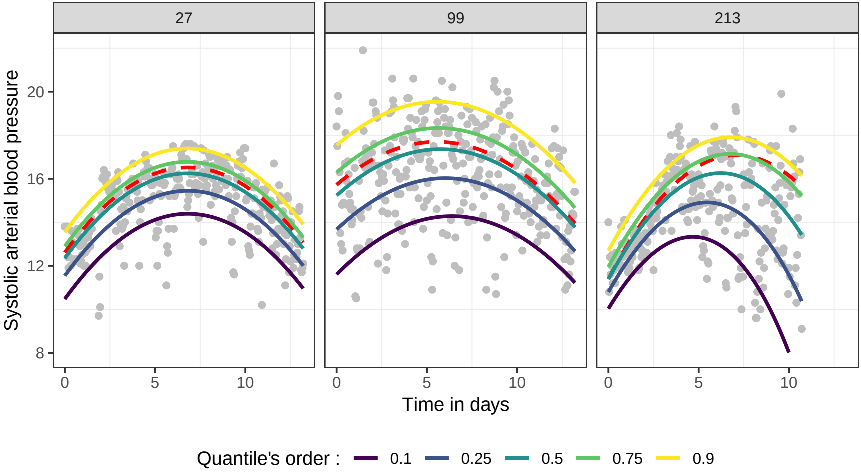
Application - Estimation results

Fixed effects (n.iter = 40000, n.burnin = 10000, n.thin = 10)

	$\widehat{\theta}$	sd	$CI_{95\%}(\theta)$	$\widehat{R}$
Location part				
$\beta_{Intercept}$	13.240	0.163	[12.922;13.559]	1.000
β_{time}	0.310	0.041	[0.231;0.392]	1.000
β_{time^2}	-0.030	0.004	[-0.038; -0.021]	1.000
β_{Age}	0.610	0.119	[0.380; 0.842]	1.000
β_{Sexe_m}	0.632	0.235	[0.17; 1.091]	1.000
Scale part				
$\xi_{Intercept}$	-0.555	0.019	[-0.592; -0.517]	1.022
ξ_{time}	-0.011	0.004	[-0.019; -0.003]	1.008
ξ_{Age}	0.084	0.016	[0.054; 0.114]	1.002
Skewness part				
α_{Age}	0.126	0.038	[0.052; 0.199]	1.003
α_{Sexe_m}	0.146	0.053	[0.041; 0.251]	1.003

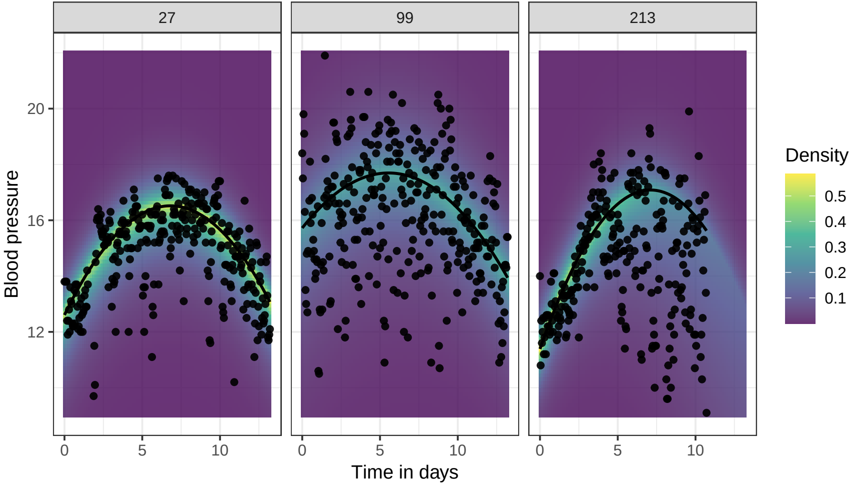
Individual conditional distributions

Quantile, density and cumulative distributions over time



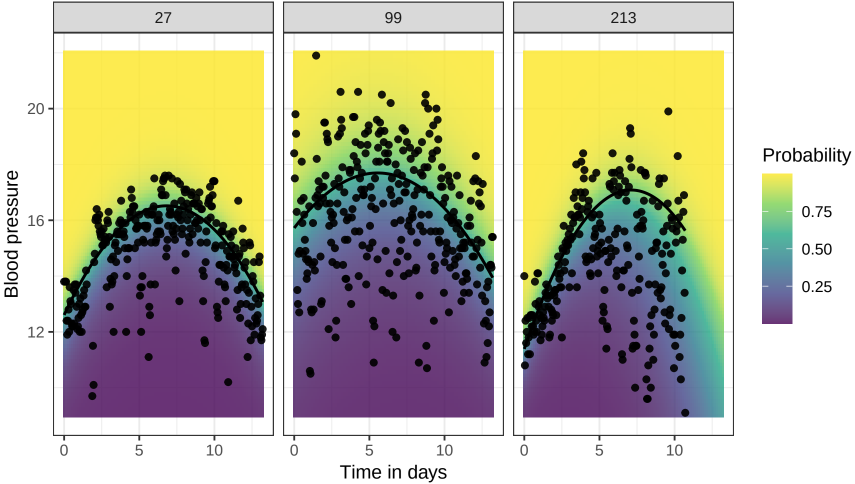
Individual conditional distributions

Quantile, density and cumulative distributions over time



Individual conditional distributions

Quantile, density and cumulative distributions over time



Model selection strategy

Choice of the probability distribution

- Which distribution to choose?
  - Use practical arguments (properties, parameter interpretation, flexibility, computational aspect)
  - Use criteria as AIC, DIC...

- New criteria for goodness-of-fit assessment
  - Focus on a set of quantiles
  - The best model is the one that minimizes

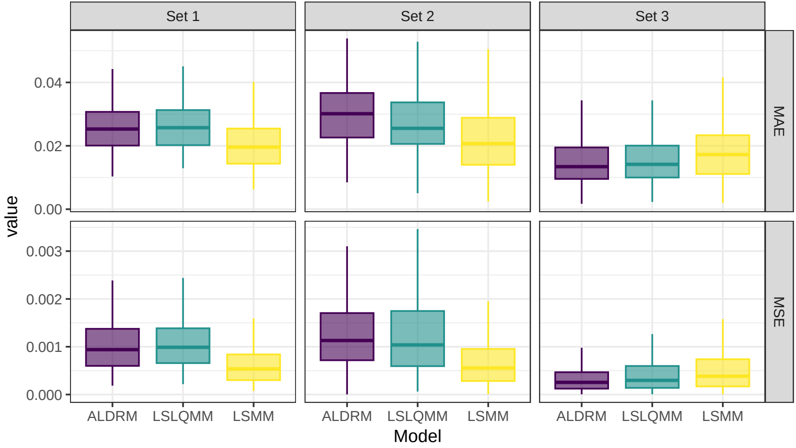
$$C_\Gamma = \frac{1}{n}\sum_{i=1}^{n} C_{i,\Gamma} \quad \text{with} \quad C_{i,\Gamma} = \frac{1}{\#\Gamma}\sum_{\gamma\in\Gamma} \ell\left(\widehat{\gamma}_i - \gamma\right)$$

  - ⋆ Γ a pre-fixed set of orders $\gamma \in]0,1[$ of the quantiles
  - ⋆ $\ell(.)$ a loss function
  - ⋆ $\widehat{\gamma}_i = \frac{n_i^\gamma}{n_i}$ with $n_i^\gamma = \sum_{j=1}^{n_i} \mathbb{1}\left\{y_{ij} < \widehat{\mathcal{Q}}_{Y_{ij}|\mathbf{r}_i}(\gamma)\right\}$

Application results

Distribution of $C_{i,\Gamma}$

- Sets: $\Gamma_1 = \{0.1, 0.2, \ldots, 0.9\}$, $\Gamma_2 = \{0.25, 0.5, 0.75\}$, $\Gamma_3 = \{0.1, 0.5, 0.9\}$
- Loss function $\ell(x) = x^2$ or $\ell(x) = |x|$ (MSE or MAE, resp.)



Discussion

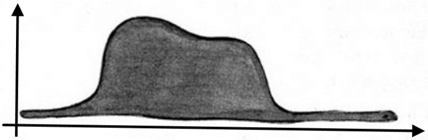
- ALDRM is a flexible distributional regression model
  - Deal with heteroskedasticity, asymmetry and *outliers*
  - Explore the entire distribution of interest
  - Provide various individual features (mode, dispersion, asymmetry, quantiles, cumulative probabilities...)

- All presented models are implemented in the R-package `BeQut`
  - Version in progress on `GitHub` and basic version on `CRAN`
  - Estimation procedure and model choice criterion validated by simulations
    - Conclusion: Everything is ok, trust me :) or see `https://arxiv.org/abs/2512.12362`

- Perspectives
  - Explore more the choice of the distribution based on data
  - Extension to join models for heterogeneous longitudinal and survival data
  - Improve computational aspects

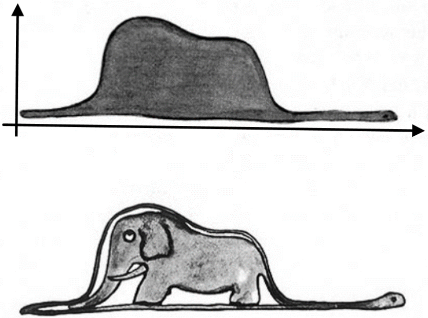
Why use R?

- What is it?



Why use R?

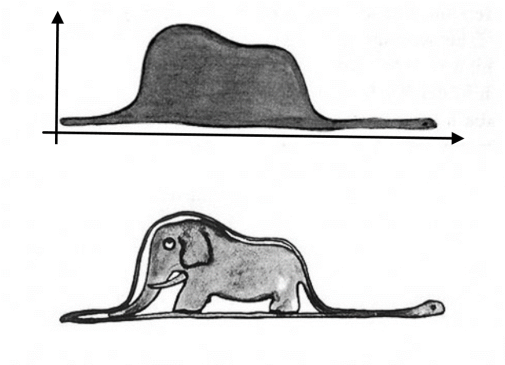
- What is it?



  - "*On ne voit bien qu'avec le coeur, l'essentiel est invisible pour les yeux*" (A de Saint-Exupéry, 1943)

Why use R?

- What is it?



  - "*On ne voit bien qu'avec le coeur, l'essentiel est invisible pour les yeux*" (A de Saint-Exupéry, 1943)

- "*Love is in the* R " (John Paul Young, 1977)

References:

- Antoine de Saint-Exupéry. Le petit prince. 1943.
- Courcoul L, Tzourio C, Woodward M, Barbieri A, and Jacqmin-Gadda H. A location-scale joint model for studying the link between the time-dependent subject-specific variability of blood pressure and competing events. *Statistics in Medicine*, 44(20-22):e70244, 2025.
- Geraci M. Linear quantile mixed models: The lqmm package for laplace quantile regression. *Journal of Statistical Software*, 57(13):1-29, 2014.
- Kneib T, Silbersdorff A, and Safken B. Rage against the mean: a review of distributional regression approaches. *Econometrics and Statistics*, 26:99-123, 2023.
- Koenker R. Quantile Regression. Cambridge University Press, 2005.
- Yu K. and Zhang J. A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics - Theory and Methods*, 34(9):1867-1879, 2005.
- Kozumi H and Kobayashi G. Gibbs sampling methods for bayesian quantile regression. 81(11):1565-1578, 2011.
- Plummer M. JAGS: Just Another Gibbs Sampler. Version 4.2.0, 2016.

Acknowledgments:

- MCIA of the Universié de Bordeaux and of the Université de Pau et des Pays de l'Adour
- French National Research Agency (grant ANR-21-CE36 for the project "Joint Models for Epidemiology and Clinical research")