Multivariate Lehmann shared frailty models for survival and binary outcomes

Marco Bonetti[1], Edoardo Ratti[2] and Maria Veronica Vinattieri[3]

[1]Bocconi University, Milan, Italy [2]University of Milano - Bicocca, Milan, Italy
[3]Karolinska Institutet, Stockholm, Sweden

July 25th, 2025

19th International Summer Conference on
Probability and Statistics
Sofia, Bulgaria

Supported in part by the MUSA – Multilayered Urban Sustainability Action – project, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5.

Summary

1. Introduction
   - Survival data
   - Cure rate models
   - Multivariate survival data
2. Multivariate Lehmann shared frailty cure rate models
   - A two-group shared frailty cure rate model
   - A shared gamma frailty cure rate model
3. A Lehmann mixed model for longitudinal binary outcomes
   - The model
   - Estimation
   - Risk Prediction
   - Results
4. Conclusions

Survival data

- $T \in [0, \infty]$ is the (non-negative) absolutely continuous random variable representing the individual *time-to-event*. T has:

  - Probability density function $f(t)$;
  - Cumulative distribution function $F(t)$;
  - Survival function $S(t) = 1 - F(t) = P(T \geq t)$;
  - Hazard function $h(t) = f(t)/S(t)$.

  C is the (right) censoring time of the subject.

  Observed data: (X, δ), with $X = \min(T, C)$ and $\delta = \mathbf{I}(T \leq C)$ the indicator of having observed the event.

- The **Lehmann family of distributions** defines the proportional hazards (PH) structure for survival distributions:

$$\{S_\alpha(t) = [S_0(t)]^\alpha \,, \quad \alpha > 0\} \,,$$

  It is easy to check that $\lambda_\alpha(t) = \alpha \, \lambda_0(t)$.

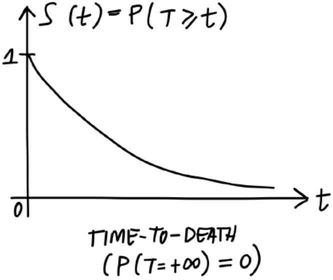
- When α is modelled through a regression structure wrt **observed** covariates ($\alpha(\mathbf{z}) = \exp(\beta' \mathbf{z})$) one obtains the celebrated Cox proportional hazards survival model; if α has a distribution, one has **frailty** (latent) models.

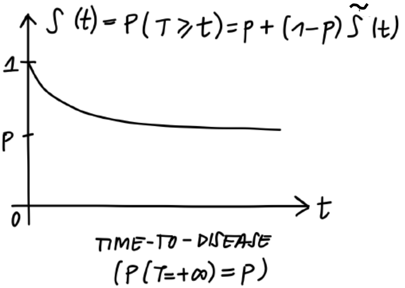
Cure rate (CR) survival models

- Now, suppose that a fraction p of all subjects ("non-susceptible" proportion) will never experience the event – think of the onsert of breast cancer – no matter how long they live.

  ⇒ "Cure rate" (CR) mixture survival models:

$$S(t) = p + (1 - p)\widetilde{S}(t).$$





Lehmann CR models

- **Idea:** extend the PH model to the more general **Lehmann cure rate** model obtained by applying the Lehmann power transformation to a baseline cure rate model:
$$\left\{ S_\alpha(t) = \left[p + (1-p)\widetilde{S}(t) \right]^\alpha, \quad \alpha > 0 \right\}.$$
- Immediately: For a fixed value α, the survival function $S_\alpha(t)$ also defines a cure rate model. Indeed, $\lim_{t\to\infty} S_\alpha(t) = p^\alpha$, and $S_\alpha(t)$ can be written as

$$S_\alpha(t) = p^\alpha + (1-p^\alpha)\,\widetilde{S}_\alpha(t)$$

with conditional (proper) survival function for the cases equal to

$$\widetilde{S}_\alpha(t) = \frac{\left[p + (1-p)\widetilde{S}(t) \right]^\alpha - p^\alpha}{1-p^\alpha}$$

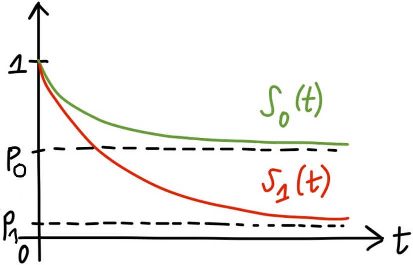
and conditional density function

$$\widetilde{f}_\alpha(t) = -\frac{d}{dt}\widetilde{S}_\alpha(t) = \frac{1-p}{1-p^\alpha}\alpha\left[p + (1-p)\widetilde{S}(t) \right]^{(\alpha-1)}\widetilde{f}(t).$$

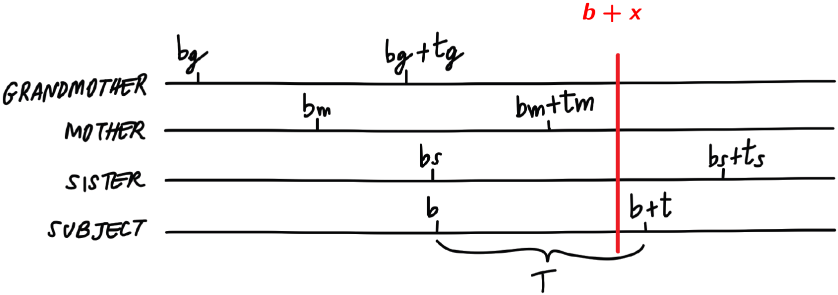
- Note that if all densities are positive $\forall t > 0$, then

$$\alpha > 1 \iff p^{\alpha} < p \iff S_{\alpha}(t) < (p + (1-p)\widetilde{S}(t)) \ \forall t \geq 0.$$

- Here, too, one may work with observed covariates (through a regression structure $\alpha = \alpha(\mathbf{z})$) or through a latent structure for α.
- In particular, let us allow for two **latent** risk classes: low (or “general”) risk (R=0) and “high” risk (R=1). Let $h = P(R = 1)$. Thus for the two risk classes one has $S_r(t) = p_r + (1-p_r)\widetilde{S}_r(t)$, with $r \in \{0, 1\}$.



Multivariate (“family”) time-to-event data



- The data generating process produces $(B, Bg, Bm, Bs, T, Tg, Tm, Ts)^T$.
- At calendar time $b + x$, we administratively censor the observation of the survival times, and thus observe a realization of

$$(B, Bg, Bm, Bs, (X, \delta), (Xg, \delta g), (Xm, \delta m), (Xs, \delta s))^T,$$

with, e.g., $Xm = min(Tm, Cm = B + x - Bm)$, $\delta m = \mathbb{I}(Tm \leq B + x - Bm)$.

- The CR model allows for the times t, ts, tm, or tg to be equal to $+\infty$.
- For each family we identify what we call the “main subject.” For the remaining $(n_i - 1)$ members we assume that their survival distributions are all equal.

A two-group shared frailty Lehmann cure rate model

Putting this together, we can extend the univariate Lehmann CR model to a multivariate shared frailty Lehmann CR model with two latent risk classes. In particular, we assume:

- **Conditional independence**, and
- **Shared frailty** or risk class membership within families.

The (parametric) survival distribution is assumed to be the same for all members of the same family, but that can be relaxed to allow for, e.g. birth cohort effects. The form of the observed data likelihood function for this model takes into account common family memberships by grouping their contributions to the likelihood within each risk group. Indeed, let θ be the whole parameter vector of the model. The observed data likelihood is (with simplified notation):

$$L^*(\theta; \text{all data}) = \prod_{i=1}^{n} \left[f_{\underline{\mathbf{x}}}(\underline{\mathbf{x}}_i | R_i = 0; \theta)(1 - h) + f_{\underline{\mathbf{x}}}(\underline{\mathbf{x}}_i | R_i = 1; \theta)\, h \right],$$

where
$\underline{\mathbf{x}} = (\underline{x} = (x, \delta)^T, \underline{x}s = (xs, \delta s)^T, \underline{x}m = (xm, \delta m)^T, \underline{x}g = (xg, \delta g)^T)^T$, and
$h = P(R = 1)$.

More

Data generation

Example of simulated family data.

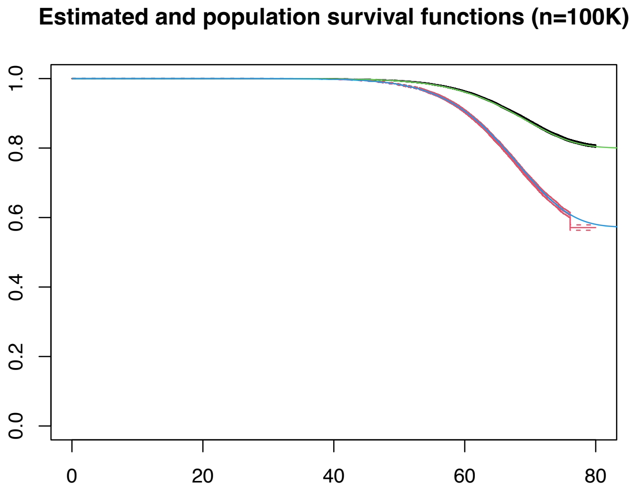
- Families (subject, sister, mother, grandmother) are generated on calendar time:

```
Bg <- runif(n,min=1880, max=1910)
Bm <- Bg + runif(n,min=25,max=35)
Bs <- Bm + runif(n,min=25,max=35)
Bval <- Bm + runif(n,min=25,max=35) # births as late as 2000
```

- We use a Weibull survival model for the cases in the low risk group: mean=66.6; sd=8.0; $p_0 = 0.80$. In high risk group (with $\alpha = 2.5$): mean=65.7 sd=7.8; $p_1 = 0.57$.
- Data are right censored by end of follow up or death:

```
Deathg <- Bg + runif(n,min=60,max=105)
Deathm <- Bm + runif(n,min=60,max=105)
Deaths <- Bs + runif(n,min=60,max=105)
Death <- Bval + runif(n,min=60,max=105)
```

- n=100K (or more) families of exactly 4 members each.
- Reparametrization for constraints, and parallel computing in R.



Sample from the model with Weibull baseline distribution.
($n = 1E06$; $nsims = 1000$; $p0 = 0.8$; $shape0 = 10$; $scale0 = 70$; $\alpha = 2.5$; $h = 0.2$.)

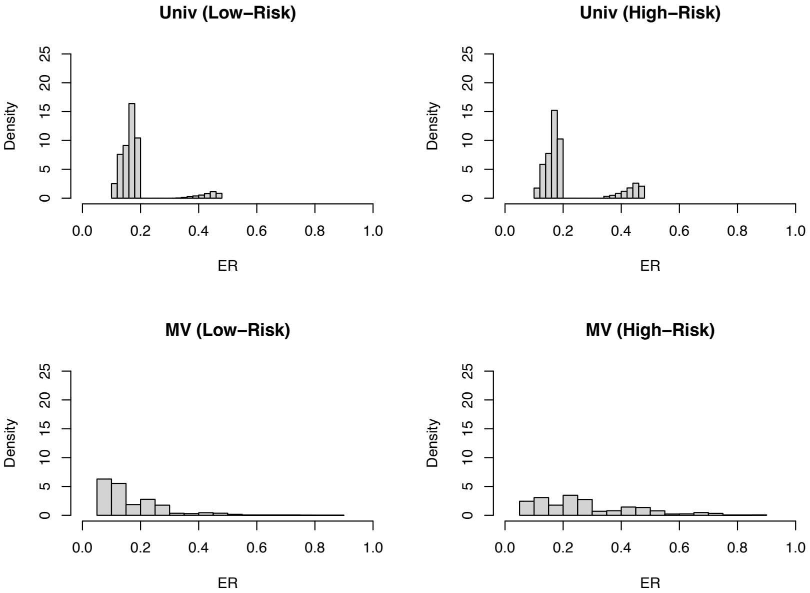
Risk group prediction

- To highlight the advantages of using a MV model one may compare its performance with that of a univariate model that only uses one subject (the "main subject").
- **Individual** risk group prediction for such univariate model is based on the conditional probability

$$P(R_i = 1 \mid (x_i, \delta_i); \widehat{\theta}) = \frac{\widehat{h}\,\widetilde{f}_1(x_i)^{\delta_i}\,\widetilde{S}_1(x_i)^{1-\delta_i}}{\widehat{h}\,\widetilde{f}_1(x_i)^{\delta_i}\,\widetilde{S}_1(x_i)^{1-\delta_i} + (1-\widehat{h})\,\widetilde{f}_0(x_i)^{\delta_i}\,\widetilde{S}_0(x_i)^{1-\delta_i}},$$

  where $\widehat{\theta}$ is the vector of the estimated model parameters.
- For the MV model this expression needs to be augmented to account for the shared frailty multivariate structure.
- The estimated conditional probabilities can be used to classify the main subjects to the high-risk or the low-risk group.

Some results



- The estimated probabilities computed on the same (main) subjects from the MV and a univariate model a Spearman's rank correlation index of 0.3978 (0.4093 within the $R = 0$ group and 0.3334 within the R=1 group).
- We can compare the classification errors as one chooses different percentiles q_p of the predicted probabilities $P(R_i = 1 \, |(x_i, \delta_i); \widehat{\theta})$ or $P(R_i = 1 \, |(\boldsymbol{x}_i, \boldsymbol{\delta}_i); \widehat{\theta})$ to classify subjects to the high-risk (when $> q_p$) vs. the low-risk group (when $\leq q_p$).

  Indeed, for each estimation procedure the probabilities of the two classification errors are $P(\mathrm{low}|R = 1)$ (the false negative rate, or 1-sensitivity) and $P(\mathrm{high}|R = 0)$ (the false positive rate, or 1-specificity), and they can be estimated by the corresponding relative frequencies for different choices of p.

- Estimated probabilities of false negative and false positive for a selection of values of the threshold probability *p*, separately for the two estimation procedures:

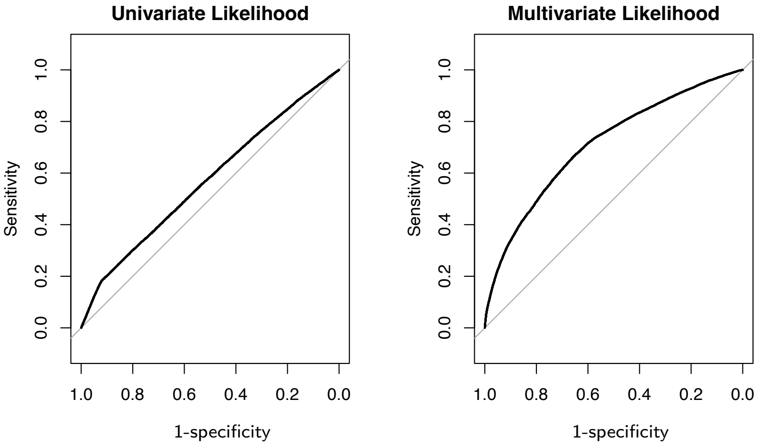
```
[1] "p0 = 0.8; shape0 = 10; scale0 = 70; alpha1 = 0.4; h = 0.2"

> print(errorsUNIVARIATE)
                  0.8      0.85        0.9       0.95
        FNR 0.7184863 0.7680744 0.81955663 0.90574953
        FPR 0.1796771 0.1295744 0.07913437 0.03896751

> print(errorsMV)
                  0.8      0.85        0.9       0.95
        FNR 0.5808513 0.6531839 0.73914133 0.85268886
        FPR 0.1453620 0.1009300 0.05989486 0.02573847
```

  (one may also look at the cross-classification rates).

- Also, the AUC measure obtained from the univariate model is estimated as 0.5725 (± .0014), while the same measure based on the multivariate model is estimated as 0.7126 (± .0012).
  [Based on 1,000 simulated samples]

▸ ROC curve shinyapp

ROC and AUC comparison of univariate vs. MV estimation



ROC curve: the plot of the points (1-specificity, sensitivity) for p in $[0, 1]$.

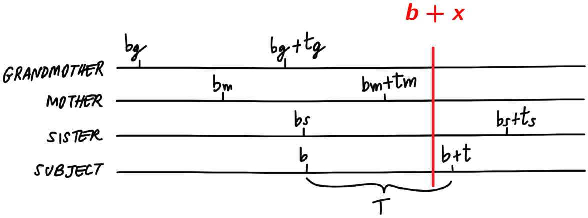
[Sens. = P(Class. High | R=1); Spec. = P(Class. Low | R=0)]

FNR = P(Class. Low | R=1); FPR = P(Class High | R=0)

A note on Family History

- An often-used alternative model to the multivariate model for disease onset is a univariate model with the covariate:

$$FH(x) = 1(one\ or\ more\ cases\ among\ relatives\ by\ calendar\ time\ b + x).$$



⇒ Dangerous: FH(x) is **poorly defined**.

- Ex. from the MV model: misclassification of R vs. $FH(t)$ at observation time:

	R	
FH(x)	0	1
0	0.5197	0.0691
1	0.2807	0.1305

⇒ MV models – when applicable – provide much more information.

A Lehmann shared gamma frailty cure rate model

- Last step: A multivariate shared **gamma** frailty Lehmann cure rate model.
- For a generic subject, the survival data is still the pair $\underline{x} = (x = \min(t, c),\ \delta = \mathbb{I}(t \leq c))^T$ where t indicates the survival time, and c indicates the administrative (independent) censoring time, both measured from the same origin (here, birth). Families are still identified with $i = 1, \ldots, n$. The observed survival data is $\underline{X} = (\underline{X}_1, \ldots, \underline{X}_n)^T$ where, for the ith family, $\underline{X}_i = (\underline{x}_{i1}, \ldots, \underline{x}_{in_i})^T$.
- We again let

$$S_r(t) = \left[p + (1-p)\widetilde{S}(t)\right]^r,$$

but now we assume that the frailty latent random variable R follows a Gamma(θ, θ) distribution:

$$g_R(r; \theta) = \frac{\theta^\theta}{\Gamma(\theta)} r^{\theta - 1} \mathrm{e}^{-r\theta},\ \theta > 0,\ r > 0. \qquad (1)$$

- We still assume conditional independence within a family given the (shared) frailty R.

Detail of likelihood function

The closed form of the multivariate likelihood $L(\theta, p, \underline{\gamma}; \underline{X})$ for $i = 1, \dots, n$ families of varying size n_i is

$$L(\theta, p, \underline{\gamma}; \underline{X}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \int_{\mathbb{R}^+} f_r(x_{ij})^{\delta_{ij}} S_r(x_{ij})^{1-\delta_{ij}} g_R(r; \theta) \mathrm{d}r$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\frac{(1-p)\widetilde{f}(x_{ij})}{p + (1-p)\widetilde{S}(x_{ij})} \right]^{\delta_{ij}} \int_{\mathbb{R}^+} \prod_{j=1}^{n_i} r^{\delta_{ij}} S_r(x_{ij}) g_R(r; \theta) \mathrm{d}r$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\frac{(1-p)\widetilde{f}(x_{ij})}{p + (1-p)\widetilde{S}(x_{ij})} \right]^{\delta_{ij}} \int_{\mathbb{R}^+} r^{\sum_{j=1}^{n_i} \delta_{ij}} \prod_{j=1}^{n_i} S_r(x_{ij}) g_R(r; \theta) \mathrm{d}r$$

Thus, given the general distribution R ∼Gamma(shape $= \alpha,$ rate $= \beta$), the internal component is given by

$$\int_{\mathbb{R}^+} r^{\sum_{j=1}^{n_i} \delta_{ij}} \prod_{j=1}^{n_i} S_r(x_{ij}) g_R(r;\alpha,\beta)\mathrm{d}r = \int_{\mathbb{R}^+} \prod_{j=1}^{n_i} S_r(x_{ij}) r^{\sum_{j=1}^{n_i} \delta_{ij}} \frac{\beta^{\alpha}}{\Gamma(\alpha)} r^{\alpha-1} \mathrm{e}^{-\beta r} \mathrm{d}r$$

$$= \int_{\mathbb{R}^+} \prod_{j=1}^{n_i} S_r(x_{ij}) \frac{\beta^{v_2(\alpha,\boldsymbol{\delta}_i)}}{\Gamma(v_2(\alpha,\boldsymbol{\delta}_i))} \frac{\Gamma(v_2(\alpha,\boldsymbol{\delta}_i))}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} r^{(v_2(\alpha,\boldsymbol{\delta}_i)-1)} \mathrm{e}^{-\beta r} \mathrm{d}r$$

$$= \prod_{j=1}^{n_i} \frac{\Gamma(v_2(\alpha,\boldsymbol{\delta}_i))}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} \int_{\mathbb{R}^+} H(x_{ij}; p, \underline{\gamma})^r g_{R^*}(r; \alpha, \sum_{j=1}^{n_i} \delta_{ij}, \beta) \mathrm{d}r$$

$$= \prod_{j=1}^{n_i} \frac{\Gamma(v_2(\alpha,\boldsymbol{\delta}_i))}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} \mathbb{E}_{R^*}[\mathrm{e}^{r\log(H(x_{ij};p,\underline{\gamma}))}]$$

$$= \prod_{j=1}^{n_i} \frac{\Gamma(v_2(\alpha,\boldsymbol{\delta}_i))}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} \mathrm{MGF}(R^*; \log(H(x_{ij}; p, \underline{\gamma})))$$

$$= \prod_{j=1}^{n_i} \frac{\Gamma(v_2(\alpha,\boldsymbol{\delta}_i))}{\Gamma(\alpha)\beta^{\sum_{j=1}^{n_i}\delta_{ij}}} \left(1 - \frac{\log(H(x_{ij}; p, \underline{\gamma}))}{\beta}\right)^{-v_2(\alpha,\boldsymbol{\delta}_i)}$$

where we define the quantity $H(x_{ij}; p, \underline{\gamma}) = \prod_{j=1}^{n_i} S(x_{ij}) = \prod_{j=1}^{n_i} \left(p + (1-p)\widetilde{S}(x_{ij})\right)$, $v_2(\alpha, \boldsymbol{\delta}_i) = \alpha + \sum_{j=1}^{n_i} \delta_{ij}$ with $\boldsymbol{\delta}_i = (\delta_{i1}, \dots, \delta_{in_i})$.

Important: The computation of the likelihood exploits the form of the MGF of R^* at $\log(H(x_{ij}; p, \underline{\gamma}))$. Indeed, $MGF_R(y) = \mathbb{E}_R[e^{Ry}]$ and for $R \sim \text{Gamma}(\alpha, \beta)$ it is

$$MGF_R(y) = \left(1 - \frac{y}{\beta}\right)^{-\alpha}.$$

The multivariate likelihood with $\alpha = \beta = \theta$ is

$$L(\underline{\pi}; \underline{X}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\frac{(1-p)\widetilde{f}(x_{ij})}{p + (1-p)\widetilde{S}(x_{ij})}\right]^{\delta_{ij}} v_1(\theta, \boldsymbol{\delta}_i) \left(1 - \frac{\log\left(\prod_{j=1}^{n_i}\left(p + (1-p)\widetilde{S}(x_{ij})\right)\right)}{\theta}\right)^{-v_2(\theta, \boldsymbol{\delta}_i)}$$

with $\underline{\pi} = (\theta, p, \underline{\gamma})$ and $v_1(\theta, \boldsymbol{\delta}_i) = \Gamma(v_2(\theta, \boldsymbol{\delta}_i))/\Gamma(\theta)\theta^{\sum_{j=1}^{n_i} \delta_{ij}}$.

- Also, the **marginal proportion of non-susceptibles** $p_{\text{marg}} = P(\text{non-susceptible})$ is

$$p_{\text{marg}} = \mathbb{E}_R(P(\text{non-susceptible} \mid R)) = \mathbb{E}_R(p^R) = \mathbb{E}_R(e^{\log(p^R)}) = \mathbb{E}_R(e^{R\log(p)})$$
$$= \text{MGF}_R(\log(p)) = \left(1 - \frac{\log(p)}{\theta}\right)^{-\theta} = \left(\frac{\theta}{\theta - \log(p)}\right)^{\theta},$$

  again exploiting the Moment Generating Function of the Gamma(θ, θ).

  (Note that, as $\theta \to \infty$, $\lim_{\theta\to\infty} p_{\text{marg}} = p$, as expected since increasing θ corresponds to decreasing heterogeneity ($\text{var}(R) \to 0$).)

- This multivariate model can also be used for risk prediction to help identify high risk families. ▸ (Not shown)

- Note: As an alternative to the Multivariate Shared Frailty Cure-Rate model, one can also implement the semiparametric Cox PH model with multiplicative shared frailty structure (note that the proportion of susceptible subjects cannot be estimated directly).

Some insights

- Simplified family history summaries, such as a binary family history indicator used in univariate models, worsen predictive performance.
- The Multivariate Shared Frailty Cure-Rate model enlarges the set of available models, within which the traditional (proper) PH survival model is nested through the constraint that $p_0 = 0$.
- These models may help target high-risk families more effectively, enabling better screening and prevention strategies.
- Challenge: fitting and assessing the goodness-of-fit of CR models with many right-censored observations. Ongoing work exploits the size of the Swedish multi-generational breast cancer registry data.
- Some “To do” items: (i) incorporate individual risk factor covariates that may affect disease onset beyond the family (residual) effect, in a mixed model; (ii) relax the assumption of identical survival distributions within a family by introducing cohort effects.
- Also: Gamma frailties are nice to work with also beyond the standard setting.

A Lehmann mixed model for longitudinal binary outcomes

The MV shared frailty (gamma) CR survival models suggest a way to also develop novel mixed effects models for binary outcomes, e.g. for repeated binary measurements on the same individual.

Consider n independent clusters (individuals) $i = 1, \ldots, n$ each with observed binary outcomes $Y_{i\bullet} = (Y_{i1}, Y_{i2}, \ldots, Y_{ik})^T$ and covariate matrix $X_{i\bullet} = [x_{i1}, x_{i2}, \ldots, x_{ik}]$, with x_{ij} is a vector of regression covariates of length r, with $x_{i1} = 1$. Let β be a vector parameter of length $r - 1$.

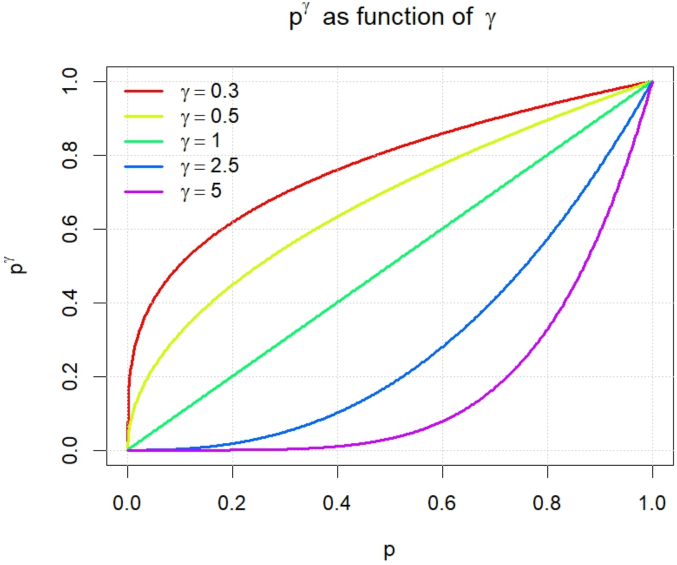
(i) $Y_{ij}|\gamma_i, x_{ij} \overset{ind}{\sim} \mathrm{Bernoulli}\left(p_{ij}^{\gamma_i}\right)$ with $p_{ij} = e^{\beta' x_{ij}} / \left(1 + e^{\beta' x_{ij}}\right)$.

(ii) $\gamma_i \overset{iid}{\sim} \text{Gamma}(\theta, \theta)$. Note that $E(\gamma_i) = 1$ and $\mathrm{var}(\gamma_i) = \theta^{-1}$.

(iii) The outcome random variables Y_{ij} are conditionally independent given γ_i.

Notes:

- Assumption (i) above is the Lehmann transformation applied to a probability (and that we have used in the shared frailty CR models).
- All probabilities $p_{ij}^{\gamma_i}$ corresponding to the same cluster (individual) are modified by the *same* frailty term $\gamma_i \Rightarrow$ *shared* frailty.
- The frailty effect γ_i acts multiplicatively on $\log(p_{ij})$.
- A *large* frailty term produces a *small* probability $P(Y_{ij} = 1|\gamma_i)$.


p^{γ} as function of γ
$\gamma = 0.3$
$\gamma = 0.5$
$\gamma = 1$
$\gamma = 2.5$
$\gamma = 5$
p^{γ}
0.0
0.2
0.4
0.6
0.8
1.0
p
0.0
0.2
0.4
0.6
0.8
1.0

Maximum Likelihood estimation

- The marginal distribution of all outcomes Y after integrating with respect to the n frailty terms is as follows:

$$
\begin{aligned}
f(Y;X,\beta,\theta) &= \int_{(\mathcal{R}^+)^n} f(Y,\gamma;X,\beta,\theta)\ f_{\Gamma}(\gamma)d\gamma \\
&= \int_{(\mathcal{R}^+)^n} \prod_{i=1}^{n} \left[f(y_{i\bullet}|\gamma_i;X_{i\bullet},\beta,\theta)\ f_{\Gamma_i}(\gamma_i) \right] d\gamma_1 \ldots d\gamma_n \\
&= \prod_{i=1}^{n} \left[\int_{\mathcal{R}^+} f(y_{i\bullet}|\gamma_i;X_{i\bullet},\beta,\theta)\ f_{\Gamma_i}(\gamma_i) d\gamma_i \right] \\
&= \prod_{i=1}^{n} \left[\int_{\mathcal{R}^+} \left(\prod_{j=1}^{k} \left(p_{ij}^{\gamma_i}\right)^{y_{ij}} \left(1 - p_{ij}^{\gamma_i}\right)^{1-y_{ij}} \right) \frac{1}{\Gamma(\theta)} \theta^{\theta} \gamma_i^{\theta-1} \mathrm{e}^{-\ \theta\gamma_i} d\gamma_i \right],
\end{aligned}
$$

which can be obtained in closed form and computed exactly (without numerical integration):

$$
\begin{aligned}
f(Y;X,\beta,\theta) = \prod_{i=1}^{n} \Bigg[& \left(\frac{\theta}{\theta - \log p_i^{(1)}}\right)^{\theta} - \sum_{(r),y_{ij}=0} \left(\frac{\theta}{\theta - \log p_i^{(1)} - \log p_{ir}}\right)^{\theta} + \\
& + \sum_{(rs),y_{ij}=0} \left(\frac{\theta}{\theta - \log p_i^{(1)} - \log p_{ir} - \log p_{is}}\right)^{\theta} - \\
& - \sum_{(rst),y_{ij}=0} \left(\frac{\theta}{\theta - \log p_i^{(1)} - \log p_{ir} - \log p_{is} - \log p_{it}}\right)^{\theta} + \\
& \vdots \\
& + (-1)^{n_{i0}} \left(\frac{\theta}{\theta - \log p_i^{(1)} - \log p_i^{(0)}}\right)^{\theta} \Bigg]
\end{aligned}
$$

where we have defined $p_i^{(1)} = \prod_{j=1, y_{ij}=1} p_{ij}$, $n_{i0} = \sum_{j=1}^{k} 1(y_{ij} = 0)$, and $p_i^{(0)} = p_{i\,r_1}, p_{i\,r2} \ldots, p_{i\,r_{n_{i0}}}$ is the product $\prod_{j=1}^{k} [1(y_{ij} = 0)\, p_{ij}]$, or the product of all probabilities p_{ij} such that their corresponding y_{ij} terms are equal to zero.

- The likelihood function, as well as the conditional distribution and expected value of each individual frailty, can be computed through exact algorithms that are however quite slow.
- We have implemented an improved R algorithm for the calculation of these quantities, and have performed the most time-consuming loops in C++ to allow for fast computation.
- The likelihood function can then be maximized numerically (we used R).
- Calculation of the mles was quite fast, even for $n = 500$K subjects observed over k=10 occasions.
- The delta method can be used to ensure that the proper standar errors for $(\widehat{\beta, \theta})^T$ are computed even though the numerical optimization function can be based on the re-parametrized parameters $(\beta, \psi)^T$ with $\theta = \exp(\psi)$ to ensure that the positivity constraint of θ is satisfied. Immediately:

$$\widehat{\mathrm{varcov}}(\widehat{\beta}, \widehat{\theta}) = \widehat{\nabla}\, \widehat{I}^{-1}\, \widehat{\nabla},$$

with $\widehat{\nabla} = \mathrm{diag}\left(1, 1, 1, \mathrm{e}^{\widehat{\psi}}\right)$ is the 4×4 diagonal matrix of the partial derivatives of the transformation, and $\widehat{I}$ is the the estimated Hessian matrix obtained numerically as a by-product of the maximization algorithm.

MLEs - Example

We simulated 1000 times from the model with the two uncorrelated covariates

$x_{ij} = (x_{ij}^{(1)}, x_{ij}^{(2)})'$:

$X^{(1)} \sim Ber(0.7)$ with coefficient $\beta_1 = 0.2$.
$X^{(2)} \sim N(0, 1)$ with coefficient $\beta_2 =$ -0.7.

We used $\beta_0 = -0.25$, $\theta = 3$, and $k = 5$ occasions.
Thus, we simulated data from the model:

$$Y_{ij} | \gamma_i, x_{ij} \overset{ind}{\sim} \mathrm{Bernoulli}\left(p_{ij}^{\gamma_i}\right)$$

with

$$p_{ij} = \frac{exp(-0.25 + 0.2x_{ij}^{(1)} - 0.7x_{ij}^{(2)})}{1 + exp(-0.25 + 0.2x_{ij}^{(1)} - 0.7x_{ij}^{(2)})}$$

and γ_i *i.i.d.* and Gamma$(3, 3)$ distributed.

▸ Example

Prediction

Risk prediction is based on the conditional distribution $F_{\gamma_i}(t|Y_{i\bullet}, X_{i\bullet}, \beta, \theta)$:

$$
\begin{aligned}
F_{\gamma_i}(t|Y_{i\bullet}, X_{i\bullet}, \beta, \theta) = \frac{1}{f(Y_{i\bullet}; X_{i\bullet}, \beta, \theta)} \Bigg[& \left(\frac{\theta}{\theta - \log p_i^{(1)}} \right)^{\theta} F(t; \theta, \theta - \log p_i^{(1)}) \\
& - \sum_{(r), y_{ij}=0} \left(\frac{\theta}{\theta - \log p_i^{(1)} - \log p_{ir}} \right)^{\theta} F(t; \theta, \theta - \log p_i^{(1)} - \log p_{ir}) + \\
& + \sum_{(rs), y_{ij}=0} \left(\frac{\theta}{\theta - \log p_i^{(1)} - \log p_{ir} - \log p_{is}} \right)^{\theta} \\
& \quad F(t; \theta, \theta - \log p_i^{(1)} - \log p_{ir} - \log p_{is}) - \\
& - \sum_{(rst), y_{ij}=0} \left(\frac{\theta}{\theta - \log p_i^{(1)} - \log p_{ir} - \log p_{is} - \log p_{it}} \right)^{\theta} \\
& \quad F(t; \theta, \theta - \log p_i^{(1)} - \log p_{ir} - \log p_{is} - \log p_{it}) + \\
& \vdots \\
& + (-1)^{n_{i0}} \left(\frac{\theta}{\theta - \log p_i^{(1)} - \log p_i^{(0)}} \right)^{\theta} F(t; \theta, \theta - \log p_i^{(1)} - \log p_i^{(0)}) \Bigg]
\end{aligned}
$$

where $F(t; a, b)$ is the cumulative distribution function of the $Gamma(a, b)$ random variable, and where we plug in the mles for the unknown parameters β and θ.

For **risk prediction** we can use the following quantites:

(a) Conditional Means: $E(\gamma_i | Y_{i\bullet}, X_{i\bullet}, \widehat{\beta}, \widehat{\theta})$

(b) Conditional Medians: η_i: $P(\gamma_i \leq \eta_i | Y_{i\bullet}, X_{i\bullet}, \widehat{\beta}, \widehat{\theta})$

(c) Prediction Intervals: $[L_i, U_i]$ such that, say, $P(\gamma_i \leq L_i | Y_{i\bullet}, X_{i\bullet}, \widehat{\beta}, \widehat{\theta}) = \alpha/2$ and $P(\gamma_i \leq U_i | Y_{i\bullet}, X_{i\bullet}, \widehat{\beta}, \widehat{\theta}) = 1 - \alpha/2$, so that $P(L_i \leq \gamma_i \leq U_i | Y_{i\bullet}, X_{i\bullet}, \widehat{\beta}, \widehat{\theta}) = 1 - \alpha$ follows. We used $\alpha = 0.2$.

The quantities (b) and (c) require the (careful) calculation of the quantile function from the conditional distribution of $\gamma_i | Y_{i\bullet}, X_{i\bullet}, \widehat{\beta}, \widehat{\theta}$.

In particular, the distribution function $F_{\gamma_i}(t)$ can be inverted numerically for any probability p, for example by minimizing the quantity $(F_{\gamma_i}(t) - p)^2$ with respect to t, so that the percentile t_p can be obtained as the argmin of the optimization.

Predictions can then be produced also for the individual probabilities of success $p_{ij}^{\gamma_i}$.

Metrics for Prediction Performance (preliminary results)

- Coverage of prediction intervals: $E\left[P(L \leq \theta \leq U)\right]$
- Average absolute (relative) prediction error: $E\left[\frac{\left|\widehat{\gamma}-\gamma\right|}{\gamma}\right]$
- Average (absolute) width of prediction intervals: $E(U - L)$
- Average relative width of prediction intervals: $E\left[\frac{U-L}{\gamma}\right]$
- Another possibility: Root MSE of Prediction $\left[E\left(\widehat{\gamma}-\gamma\right)^2\right]^{1/2}$

These quantities are of interest both pre- and post-calibration, and both marginally and locally across the range of values of the true frailties γ_i.

- Conditional probabilities of correct classification in lower and upper 20% (say) of frailty distribution:
  (i) $P(\gamma < 20th\ perc \mid \widehat{\gamma} < 20th\ perc) = P(\widehat{\gamma} < 20th\ perc \mid \gamma < 20th\ perc)$
  (ii) $P(\gamma > 80th\ perc \mid \widehat{\gamma} > 80th\ perc) = P(\widehat{\gamma} > 80th\ perc \mid \gamma > 80th\ perc)$

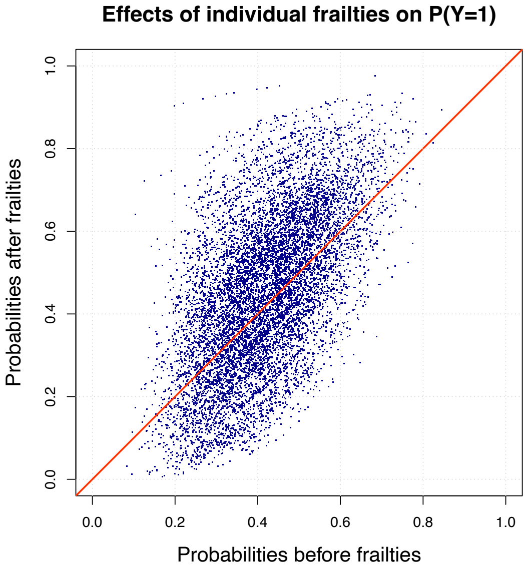
Calibration

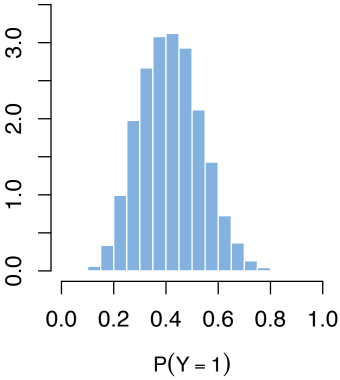
- The conditional means and medians (as well as the prediction intervals) show some bias in predicting the true individual frailties.
- **Idea:** since the model assumes that $\Gamma \sim \mathrm{Gamma}(\theta, \theta)$, we can calibrate the predictions to match that marginal distribution. This can be done through the quantile transformation, implemented on the $\mathrm{Gamma}(\widehat{\theta}, \widehat{\theta})$ distribution to obtain the new calibrated predictions

$$\widetilde{\gamma}_i = F^{-1}_{Gamma(\widehat{\theta},\widehat{\theta})} \left(\mathrm{Rank}(\widehat{\gamma}_i)/n - \frac{1}{2n} \right).$$

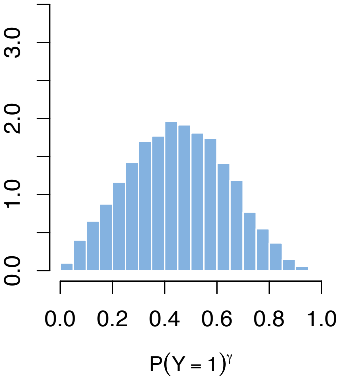
  (where $\mathrm{Rank}(\text{smallest } \widehat{\gamma}_i) = 1$ and $\mathrm{Rank}(\text{largest } \widehat{\gamma}_i) = n$).
- The prediction intervals $[L_i, U_i]$ can then be shifted by $[\ \widetilde{\gamma}_i - \widehat{\gamma}_i\ ]$ (why?). The interval's width thus remains unchanged.
- Our experiments suggest that calibration is **not** necessarily useful, but also that in some cases it can be beneficial in the prediction of *very small* and *very large* frailties.

The following are **some results** from simulated data with $\beta = (-0.5, 0.2, -0.5)^T$, $\theta = 5$, $k = 10$, and $n = 1,000$. The two independent covariates were $X_1 \sim \mathrm{Bernoulli}(0.7)$ and $X_2 \sim \mathrm{N}(0, 1)$.

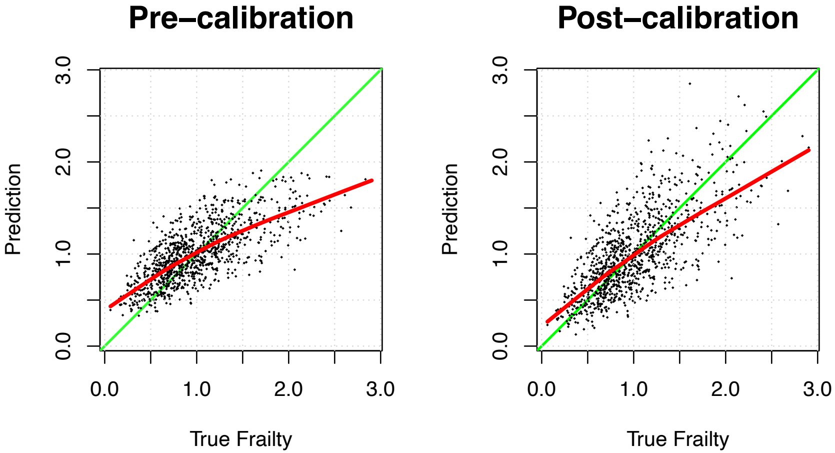

Effects of individual frailties on P(Y=1)
Probabilities after frailties
0.0
0.2
0.4
0.6
0.8
1.0
Probabilities before frailties
0.0
0.2
0.4
0.6
0.8
1.0

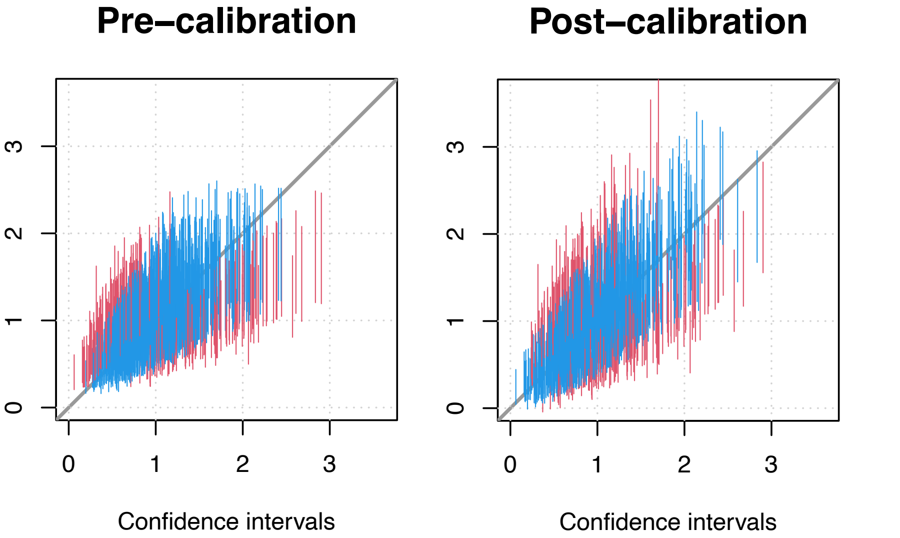


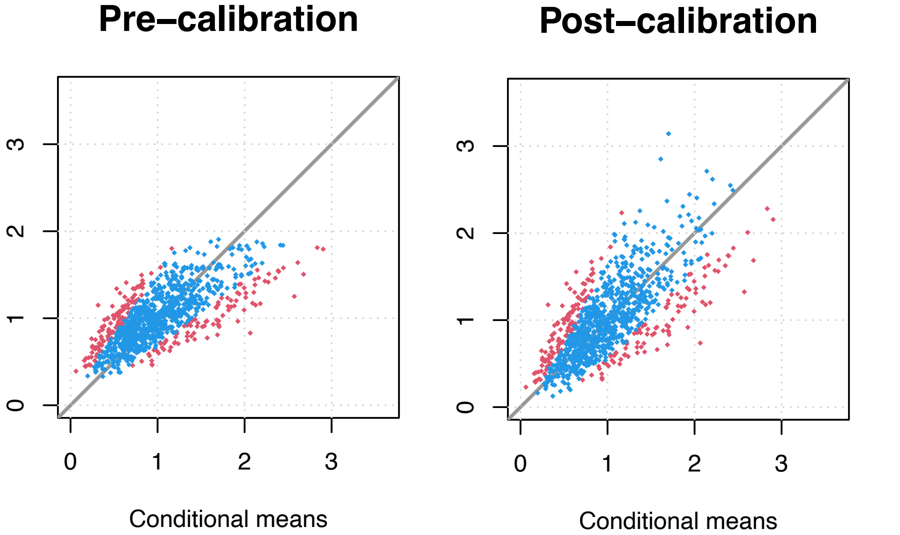
P(Y=1) before frailties

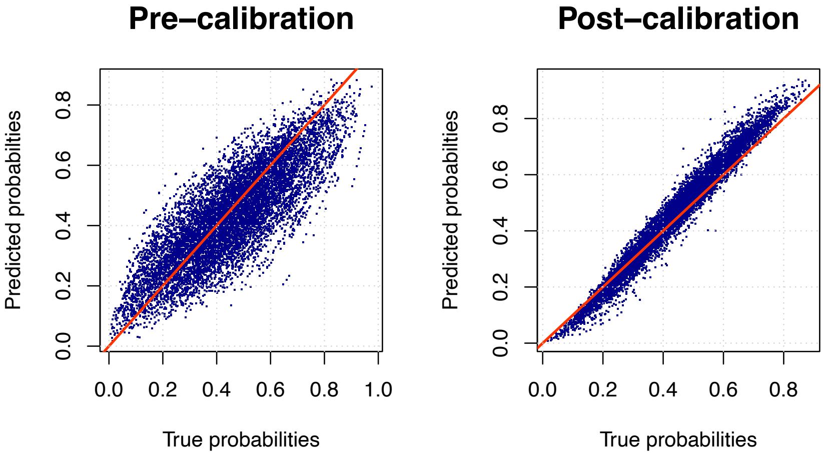


P(Y=1) after frailties

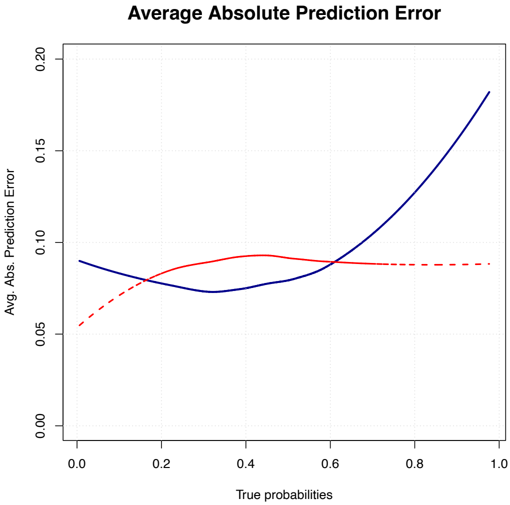

Pre–calibration
Post–calibration
Prediction
0.0
1.0
2.0
3.0
True Frailty


Pre–calibration
Post–calibration
0
1
2
3
0
1
2
3
Confidence intervals
0
1
2
3
0
1
2
3
Confidence intervals

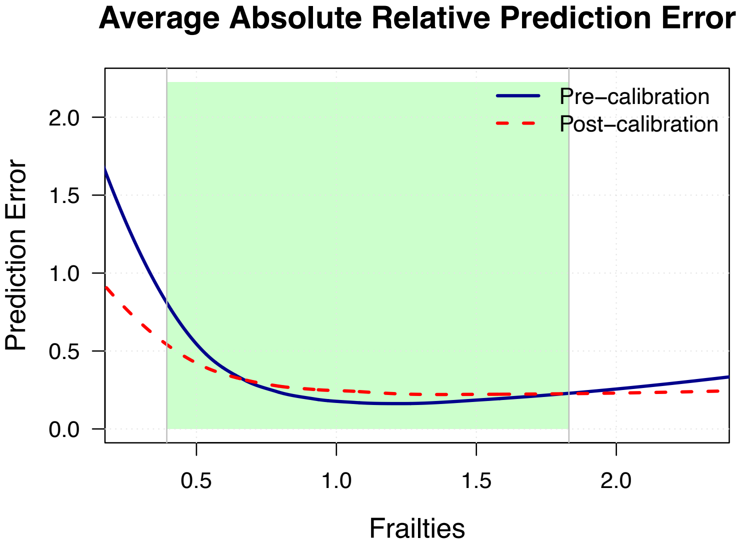

Pre–calibration
Post–calibration
0
1
2
3
0
1
2
3
Conditional means
0
1
2
3
0
1
2
3
Conditional means

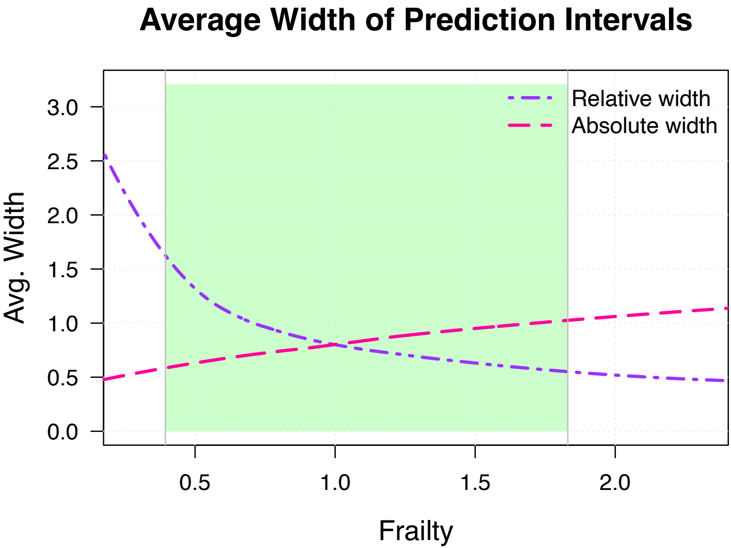


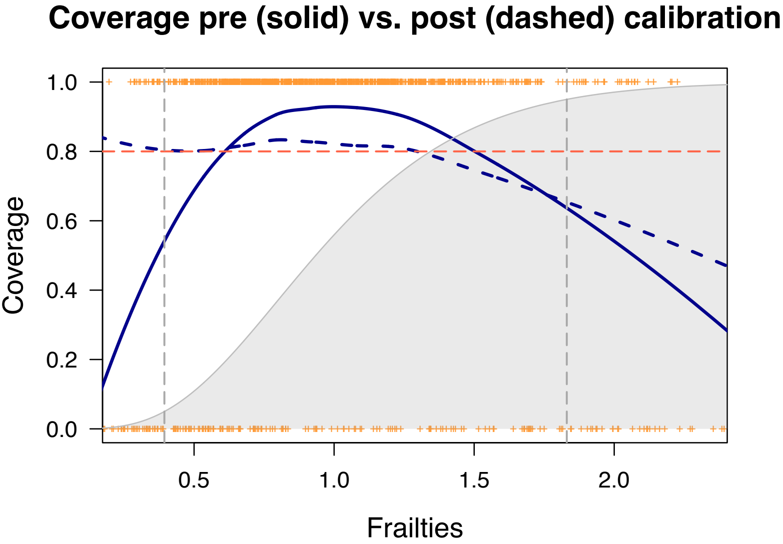
(Prediction of the probabilities $p_{ij}^{\gamma_i}$, all bunched together)



(Prediction of the probabilities $p_{ij}^{\gamma_i}$, all bunched together)


Average Absolute Relative Prediction Error
Pre−calibration
Post−calibration
Prediction Error
0.0
0.5
1.0
1.5
2.0
0.5
1.0
1.5
2.0
Frailties


Average Width of Prediction Intervals


Coverage pre (solid) vs. post (dashed) calibration
Coverage
0.0
0.2
0.4
0.6
0.8
1.0
Frailties
0.5
1.0
1.5
2.0

θ	k	n	CovPreCal	CovPostCal	AbsRelPredErrF_Pre	AbsRelPredErrF_Post
1	5	1000	0.8110	0.8050	1.8724	1.0627
		10000	0.8004	0.8072	1.6266	1.0598
	10	1000	0.8120	0.8150	0.9501	0.8073
		10000	0.7896	0.8086	2.8735	1.9153
	20	1000	0.7870	0.7940	0.4861	0.4063
		10000	0.7956	0.8015	0.8122	0.5467
5	5	1000	0.7760	0.7120	0.3785	0.3954
		10000	0.8011	0.7490	0.3578	0.3791
	10	1000	0.7960	0.7560	0.2984	0.3049
		10000	0.7933	0.7676	0.3066	0.3119
	20	1000	0.8060	0.7760	0.2286	0.2401
		10000	0.8003	0.7835	0.2427	0.2438
10	5	1000	0.8060	0.7160	0.2546	0.2914
		10000	0.7946	0.7214	0.2551	0.2891
	10	1000	0.8170	0.7570	0.2269	0.2461
		10000	0.7984	0.7401	0.2311	0.2529
	20	1000	0.8120	0.7680	0.1918	0.2065
		10000	0.7894	0.7593	0.2029	0.2133

θ	k	n	RelW	Width	PredErrP_Pre	PredErrP_Post
1	5	1000	4.7824	1.4559	0.1063	0.1105
		10000	4.2345	1.3790	0.1072	0.1077
	10	1000	2.6147	1.0812	0.0789	0.0787
		10000	6.7939	1.1151	0.0834	0.0831
	20	1000	1.4346	0.8513	0.0597	0.0598
		10000	2.1354	0.8562	0.0608	0.0610
5	5	1000	1.1108	0.9384	0.0794	0.0892
		10000	1.1003	0.9306	0.0762	0.0862
	10	1000	0.9370	0.8220	0.0662	0.0716
		10000	0.9375	0.8251	0.0674	0.0727
	20	1000	0.7427	0.6805	0.0528	0.0572
		10000	0.7576	0.6838	0.0550	0.0573
10	5	1000	0.7888	0.7240	0.0588	0.0703
		10000	0.7924	0.7225	0.0591	0.0697
	10	1000	0.7224	0.6679	0.0528	0.0596
		10000	0.7242	0.6670	0.0537	0.0608
	20	1000	0.6275	0.5849	0.0453	0.0497
		10000	0.6268	0.5855	0.0479	0.0515

Insights

- Maximization of the conditional likelihood was fast, and the mles had the expected increasing precision for growing sample size.
- The standard errors of the mles of the regression parameters were found to be quite stable.
- Large values of θ were estimated with a large variance.
- Conditional means $\simeq$ conditional medians.
- The probabilities of correct classification

$$P(true\ frailty < 20th\ percentile \mid cond.\ mean < 20th\ percentile)$$
$$P(true\ frailty > 80th\ percentile \mid cond.\ mean > 80th\ percentile)$$

  were both well above 20%, and in particular they are around 60% and 50%, respectively.
- Marginal coverage very accurate for conditional means, sometimes sligthly lower after calibration.
- Calibration shows some improvement in the stability of the conditional coverage across the (useful) range of the true frailties.

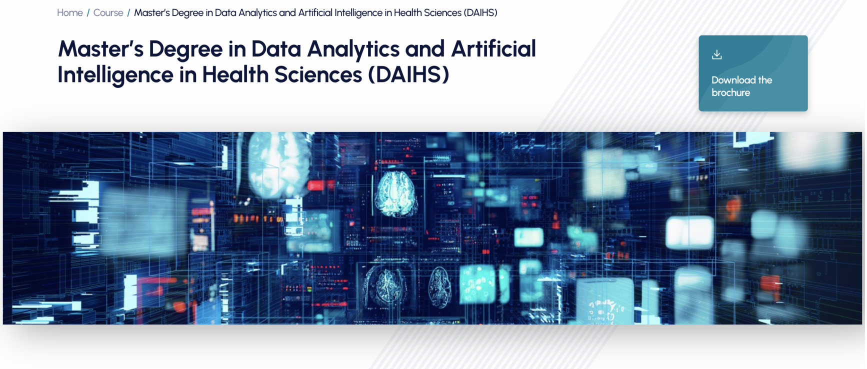
Conclusions

- Shared frailty models for CR survival data and for longitudinal binary data can be useful addition to the toolbox for disease onset and response (e.g. compliance) problems.
- Individual-level risk prediction is difficult but very relevant.
- The binary model is an alternative (with just one random effect) to Bernoulli GLLMs which assume

$$P\left(Y_{ij}=1|\gamma_{ij},x_{ij}\right)=\frac{e^{\beta' x_{ij}+\gamma_i' z_{ij}}}{1+e^{\beta' x_{ij}+\gamma_i' z_{ij}}} \tag{2}$$

  with γ_i is the vector of random effect for the i individual, with $\left(\gamma_{i1},\ldots,\gamma_{ik}\right)^T \sim N_k(0,\Sigma)$.
- We experienced convergence problems only rarely, possibly thanks to the closed form nature of the likelihood function.
- Some more "To Do" items: (i) Extensive simulation studies; (ii) Experimenting on data sets ▸ Projects; (iii) Quantification and interpretation of covariate effects in longitudinal model; (iv) Multiple (correlated) frailty terms γ_{ij}, $i=1,\ldots,n$ and $j=1,\ldots,k$.

Advertisement: New MSc program at Bocconi University joint with Humanitas University https://www.hunimed.eu/course/master-daihs/



Home / Course / Master's Degree in Data Analytics and Artificial Intelligence in Health Sciences (DAIHS)

Master's Degree in Data Analytics and Artificial Intelligence in Health Sciences (DAIHS)

marco.bonetti@unibocconi.it

Some references

[1] L. Bondi, M. Bonetti, D. Grigorova, and A. Russo. Approximate bayesian computation for the natural history of breast cancer, with application to data from a Milan cohort study. *Stat in Med*, 42(18):3093–3113, 2023.

[2] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[3] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

[4] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[5] E. L. Lehmann. The power of rank tests. *Annals of Mathematical Statistics*, 24(1):23–43, 1953.

[6] S. H. Moolgavkar and D. J. Venzon. Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. *Mathematical biosciences*, 47(1-2):55–77, 1979.

[7] M. S. Pepe et al. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, 2003.

[8] J. Tyrer, S. W. Duffy, and J. Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine*, 23(7):1111–1130, 2004.

Applications (ongoing)

- MV CR model:

  [1.] Analysis of the Swedish Multi-Generational Breast Cancer registry. The dataset concerns a cohort of $n = 1,603,920$ Swedish families, consisting of a total of $4,267,803$ women.

- Longitudinal binary model:

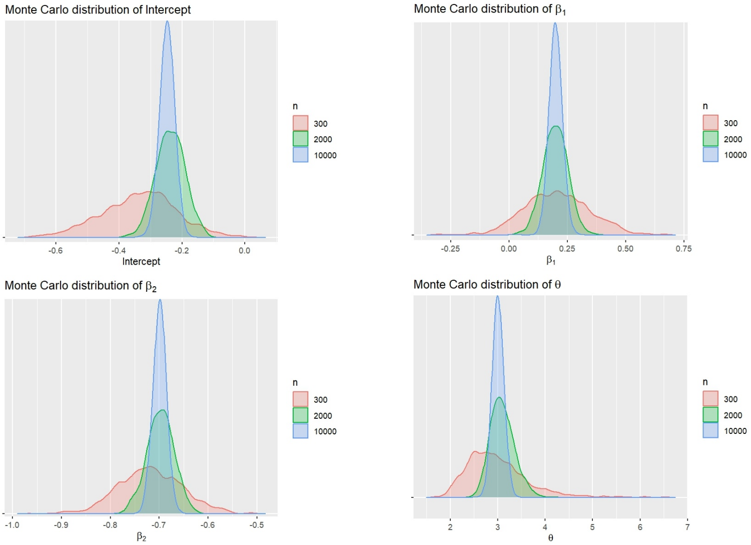
  [2.] Analysis of adherence to screening invitations in the HIP breast cancer screening trial (NCI).
  Prediction of the individual risk of non-adherence above and beyond measured risk factors, so that the more non-adherence-prone women may be identified as targets for compliance-inducing interventions, as well as for the further study of the determinants of their behavior.

  [3.] Tooth decay ("presence of at least one cavity") in pre-school children in the Italian region of Monza and Brianza. Use of mixed models to identify school-specific heterogeneity beyond the effect of observed child-specific covariates, in particular on nutrition.

  (For both, comparison with GLMMs results.)

▸ Back

MLEs



Monte Carlo distribution of Intercept

Monte Carlo distribution of β_1

Monte Carlo distribution of β_2

Monte Carlo distribution of θ

Monte Carlo distributions of the estimators (1000 simulations).

Monte Carlo results, 1000 simulations. Standard errors were close to the empirical SEs.

Measure / Parameter	**n = 300**	**n = 2 000**	**n = 10 000**
β_0 **(true =** -0.25**)**			
Estimate	−0.3350	−0.2369	−0.2477
Bias	−0.0850 (0.0038)	0.0131 (0.0014)	0.0023 (0.0007)
Relative Bias	0.3399 (0.0151)	−0.0526 (0.0057)	−0.0091 (0.0026)
Empirical SE	0.1194 (0.0027)	0.0453 (0.0010)	0.0208 (0.0005)
Coverage (95%)	0.9260 (0.0083)	0.9470 (0.0071)	0.9620 (0.0060)
β_1 **(true =** 0.20**)**			
Estimate	0.2055	0.1984	0.1998
Bias	0.0055 (0.0044)	−0.0016 (0.0016)	−0.0002 (0.0008)
Relative Bias	0.0275 (0.0221)	−0.0081 (0.0082)	−0.0010 (0.0039)
Empirical SE	0.1397 (0.0031)	0.0521 (0.0012)	0.0246 (0.0006)
Coverage (95%)	0.9510 (0.0068)	0.9470 (0.0071)	0.9400 (0.0075)
β_2 **(true =** -0.70**)**			
Estimate	−0.7184	−0.6975	−0.6993
Bias	−0.0184 (0.0021)	0.0025 (0.0009)	0.0007 (0.0004)
Relative Bias	0.0263 (0.0030)	−0.0036 (0.0012)	−0.0010 (0.0006)
Empirical SE	0.0661 (0.0015)	0.0273 (0.0006)	0.0122 (0.0003)
Coverage (95%)	0.9640 (0.0059)	0.9390 (0.0076)	0.9530 (0.0067)
θ **(true =** 3**)**			
Estimate	2.9749	3.0987	3.0064
Bias	−0.0251 (0.0201)	0.0987 (0.0079)	0.0064 (0.0033)
Relative Bias	−0.0084 (0.0067)	0.0329 (0.0026)	0.0021 (0.0011)
Empirical SE	0.6363 (0.0142)	0.2503 (0.0056)	0.1044 (0.0023)
Coverage (95%)	0.9210 (0.0085)	0.9680 (0.0056)	0.9590 (0.0063)

▸ Back