Laplace approximations in double additive cure survival models with exogenous time-varying covariates

Philippe Lambert[1,2]
Joint work with Michaela Kreyenfeld (Hertie School, Berlin).

1. UR Mathematics
Université de Liège, Belgique

2. Institut de Statistique, Biostatistique et
sciences Actuarielles (ISBA)
Université catholique de Louvain.

http : //www.statsoc.ulg.ac.be/

Cure survival models

- In classical survival analysis, it is often implicitly assumed that all units will experience the event of interest if one waits for sufficiently long.
- Mathematically, this translates into a population survival function equal to zero at infinity:

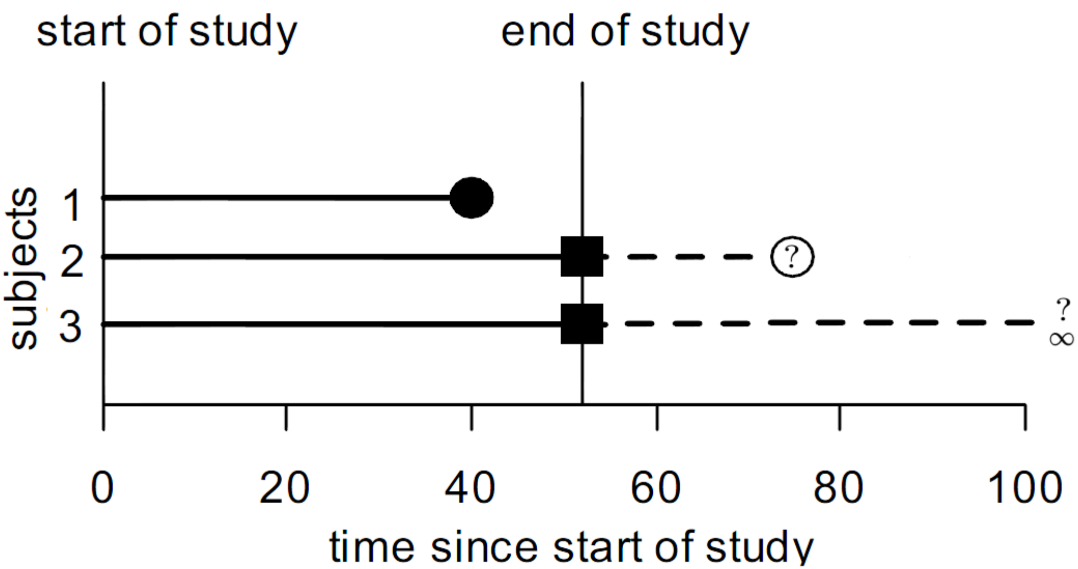
$$S_p(+\infty) = \lim_{t \to +\infty} \Pr(T > t) = 0$$

- However, in many practical situations, an unidentified cure fraction of the population will never experience the event of interest:
  - ▷ Immunotherapy: patients under treatment might be totally cured when the immune system is reactivated.
  - ▷ Fertility study: women under follow-up might have decided not to have a child.
- In these situations, $S_p(\cdot)$ is an improper (population) survival function:

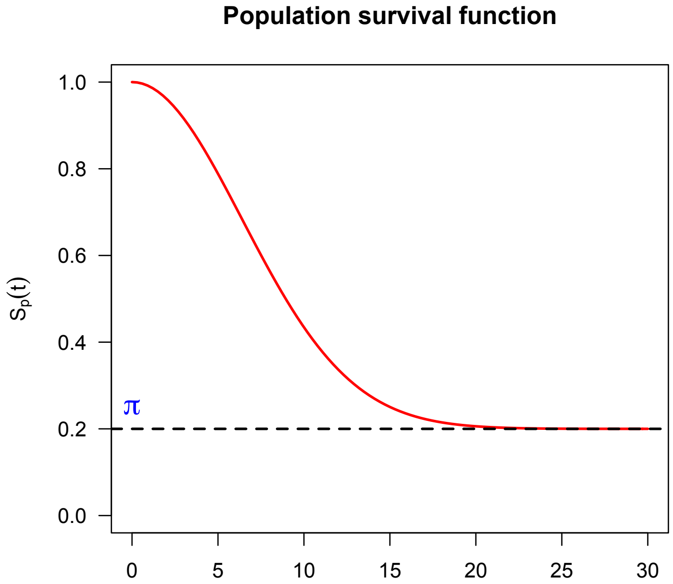
$$S_p(+\infty) = \pi > 0 \quad \text{(Long-term survival)}$$

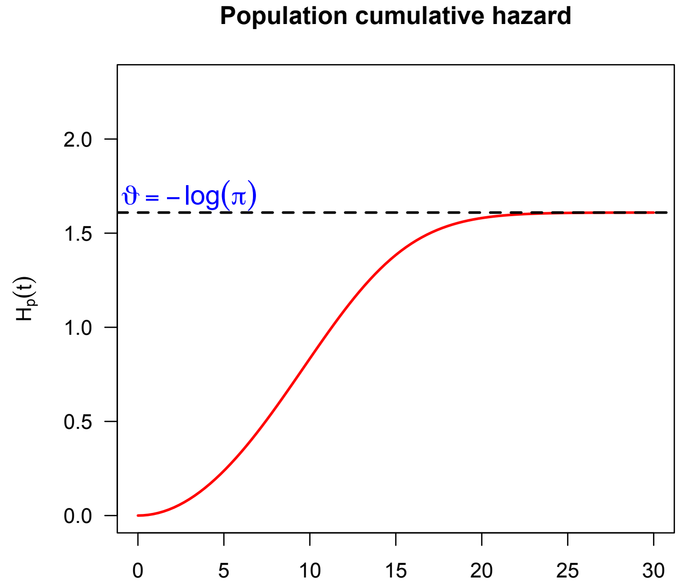
Right censoring and cure models

- The major difficulty when a cure fraction is present is the status of subjects with a right-censored response.
- Will a subject with a right-censored response finally experience the event ?



The promotion time cure model



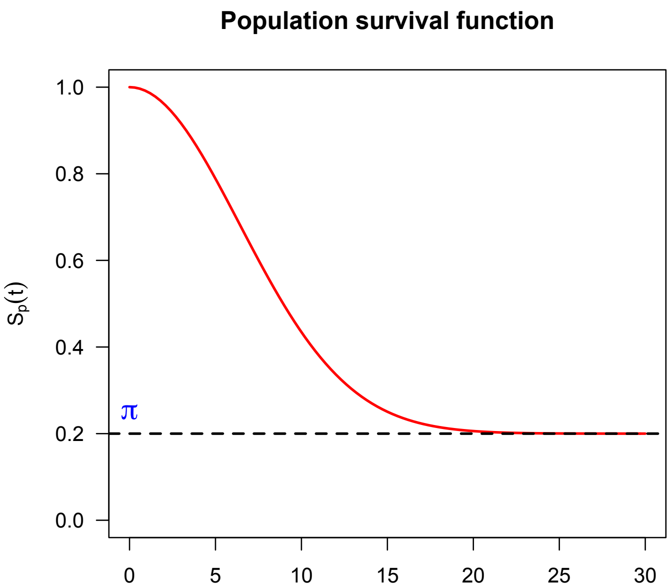


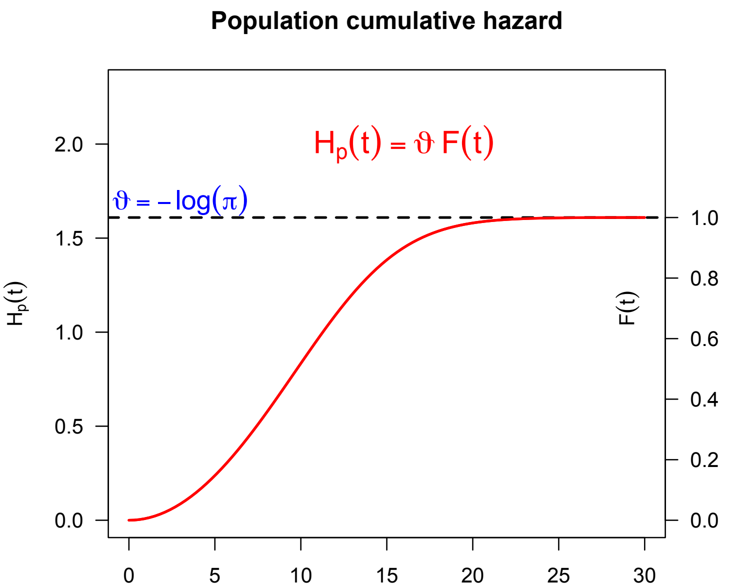
$$\text{where} \quad S_p(t) = \exp\left(-H_p(t)\right)$$

- The existence of a cured fraction π implies that the *cumulative hazard* $H_p(t)$ at the population level is bounded: it remains constant when $t \geq T$:

$$t \geq T \Longrightarrow H_p(t) = -\log S_p(+\infty) = -\log \pi = \vartheta$$

$H_p(t) = \vartheta \times F(t)$



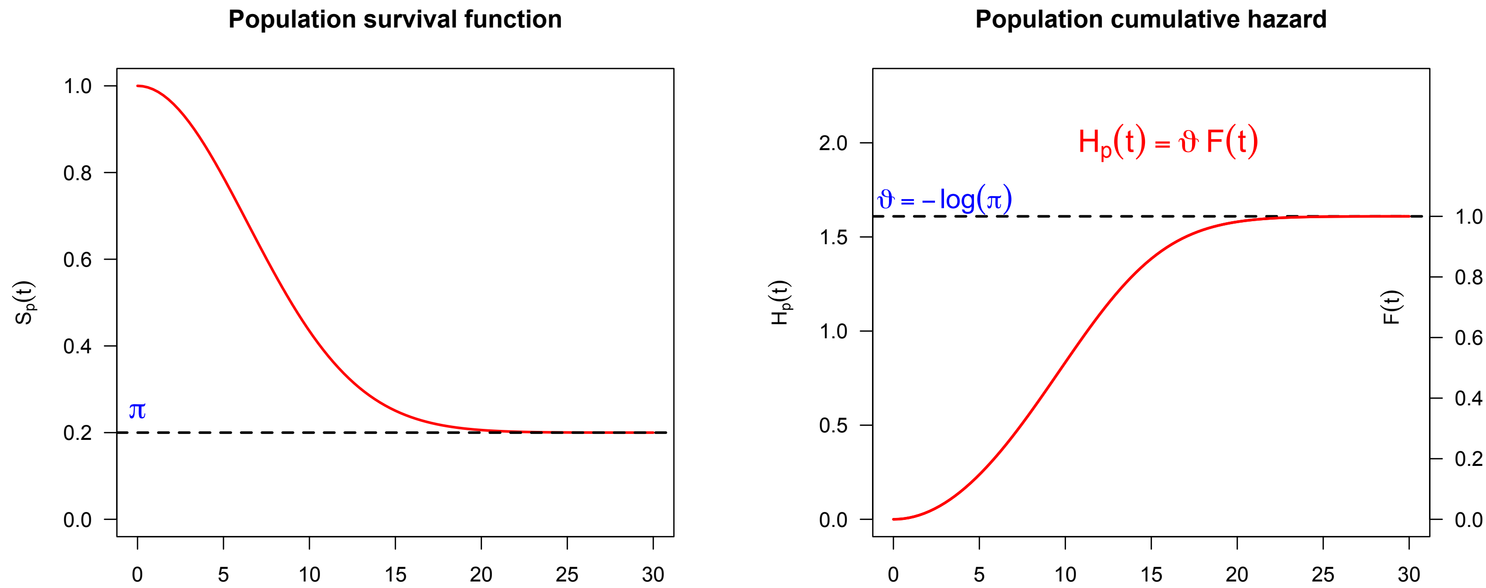


- Given that $H_p(t)$ is a non decreasing function with $H_p(0) = 0$, one has

$$H_p(t) = \vartheta \times F(t)$$

where $F(\cdot)$ is a c.d.f. such that $F(0) = 0$ and $F(T) = 1$.

$$S_p(t) = \exp\{-\vartheta F(t)\}$$



- Cure fraction $\Longrightarrow$ $\boxed{S_p(t) = \exp\{-\vartheta F(t)\}}$ with P(cured) $= \pi = \exp(-\vartheta)$
- Note: $F(t)$ is the *normalized cumulative hazard*.
  It governs the dynamics of $H_p(t)$ (i.e. short-term survival) towards ϑ.
- The *bounded cumulative hazard model* or *promotion time cure model*, was initially motivated by a stochastic model of tumor latency (Yakovlev & Tsodikov 1996).

Covariate inclusion in the promotion time model

- $H_p(t) = \vartheta(\mathbf{v}) \times F(t|\tilde{\mathbf{v}})$
  - ▷ $\mathbf{v} = (\mathbf{z}, \mathbf{x})$: covariates influencing the cure probability (long-term survival);
  - ▷ $\tilde{\mathbf{v}} = (\tilde{\mathbf{z}}, \tilde{\mathbf{x}})$: covariates influencing the hazard dynamics (short-term survival).
- One or more covariates could be shared by $\mathbf{v}$ and $\tilde{\mathbf{v}}$.
- Where is $H_p(t)$ going ?

$$\vartheta(\mathbf{v}) = -\log \underbrace{\pi(\mathbf{v})}_{\text{P(cure)}} = \exp\{\eta_\vartheta(\mathbf{v})\} \text{ where } \eta_\vartheta(\mathbf{v}) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{z} + \sum_j \underbrace{f_j(x_j)}_{\text{P-splines}}$$

- How fast does $H_p(t)$ go to ϑ ?

$$F(t|\tilde{\mathbf{v}}) = 1 - \underbrace{S_0(t)}_{\text{P-splines}}{}^{\exp(\eta_F(\tilde{\mathbf{v}}))} \text{ where } \eta_F(\tilde{\mathbf{v}}) = \boldsymbol{\gamma}^\top \tilde{\mathbf{z}} + \sum_j \underbrace{\tilde{f}_j(\tilde{x}_j)}_{\text{P-splines}}$$

- Then, the population survival function becomes

$$S_p(t|\mathbf{v}, \tilde{\mathbf{v}}) = \exp\{-H_p(t)\} = \exp\{-\vartheta(\mathbf{v}) F(t|\tilde{\mathbf{v}})\}$$

Data from German pension registers

- Event of interest: age at 1st pregnancy (1st birth minus 9 months).
- Random sample of 15 248 women from 'West' Germany.
- 5 cohorts: 1950-54, 1955-59, 1960-64, 1965-69 and 1970-74.
- Monthly information from Age 20 till 45 (maximum) on:
  - ▷ Occupation status ('Educ' ; 'Unemployed' ; 'Employed' ; 'Other') ;
  - ▷ Relative earnings (with 1.00 = Average gross earnings in a given year).

  $\Longrightarrow$ time-varying covariates !!
- Possibility of right censoring (i.e. no baby after a follow-up ending before age 45):

		Person	Mother by Age 45		
Cohort	n	months	Yes	No	Right-cens.
1950-54	2423	277668	1900 (78.4%)	519 (21.4%)	4 (0.2%)
1955-59	2388	313692	1762 (73.8%)	613 (25.7%)	13 (0.5%)
1960-64	3029	431144	2186 (72.2%)	797 (26.3%)	46 (1.5%)
1965-69	3385	511119	2410 (71.2%)	887 (26.2%)	88 (2.6%)
1970-74	4023	655489	2761 (68.6%)	1130 (28.1%)	132 (3.3%)

Extension to time-varying covariates

- Remember that

$$S_p(t|\mathbf{v},\tilde{\mathbf{v}}) = \exp\{-H_p(t)\} = \exp\{-\vartheta(\mathbf{v})F(t|\tilde{\mathbf{v}})\}$$

$$\Longrightarrow \qquad h_p(t|\mathbf{v},\tilde{\mathbf{v}}) = \mathrm{e}^{\eta_\vartheta(\mathbf{v})+\eta_F(\tilde{\mathbf{v}})} h_0(t) S_0(t)^{\exp(\eta_F(\tilde{\mathbf{v}}))}$$

- The key idea is to start from the population hazard specification:

Substitute: $\mathbf{v} \longrightarrow \mathbf{v}(t) \;\; ; \;\; \tilde{\mathbf{v}} \longrightarrow \tilde{\mathbf{v}}(t)$

$$\Longrightarrow \quad h_p(t|\mathbf{v}(t),\tilde{\mathbf{v}}(t)) = \mathrm{e}^{\eta_\vartheta(\mathbf{v}(t))+\eta_F(\tilde{\mathbf{v}}(t))} h_0(t) S_0(t)^{\exp(\eta_F(\tilde{\mathbf{v}}(t)))}$$

- Likelihood contribution for woman i during month m (with $t = m\Delta$):
  - ▷ No event: $\exp\{-h_p(t|\mathbf{v}_i(t),\tilde{\mathbf{v}}_i(t))\Delta\}$
  - ▷ Event: $h_p(t|\mathbf{v}_i(t),\tilde{\mathbf{v}}_i(t)) \, \exp\{-h_p(t|\mathbf{v}_i(t),\tilde{\mathbf{v}}_i(t))\Delta\}$

- Smoothness priors on the splines parameters in $f_0(t)$, $f_j(t)$ and $\tilde{f}_j(t)$:

$$p(\boldsymbol{\phi}|\omega) \propto \exp\big(-.5\,\boldsymbol{\phi}'(\omega\mathbf{P}_0)\,\boldsymbol{\phi}\big)$$

$$p(\boldsymbol{\theta}_j|\lambda_j) \propto \exp\big(-.5\,\boldsymbol{\theta}_j'(\lambda_j\mathbf{P})\,\boldsymbol{\theta}_j\big) \;\; ; \;\; p(\tilde{\boldsymbol{\theta}}_j|\tilde{\lambda}_j) \propto \exp\big(-.5\,\tilde{\boldsymbol{\theta}}_j'(\tilde{\lambda}_j\tilde{\mathbf{P}})\,\tilde{\boldsymbol{\theta}}_j\big).$$

Inference

Iterate the following steps till convergence: $\boldsymbol{\tau}=(\omega,\boldsymbol{\lambda},\tilde{\boldsymbol{\lambda}})$; $\boldsymbol{\zeta}=(\boldsymbol{\phi},\beta_0,\boldsymbol{\beta},\boldsymbol{\psi},\boldsymbol{\gamma},\tilde{\boldsymbol{\psi}})$

1. Estimation of the regression and spline parameters $\boldsymbol{\zeta}$ for given penalty parameters:

   ▷ Fast Maximization of the conditional posterior using Newton-Raphson:

$$\log p(\boldsymbol{\zeta}|\boldsymbol{\tau},\mathcal{D})=\ell-\frac{\omega}{2}\underbrace{\boldsymbol{\phi}^\top\mathbf{P}_0\boldsymbol{\phi}}_{h_0(t)}-\frac{1}{2}\underbrace{(\boldsymbol{\psi}-\mathbf{b})^\top\mathbf{K}_\lambda(\boldsymbol{\psi}-\mathbf{b})}_{\beta_0,\ \boldsymbol{\beta},\ f_j(t)\ (j=1,\ldots J)}-\frac{1}{2}\underbrace{(\tilde{\boldsymbol{\psi}}-\mathbf{g})^\top\tilde{\mathbf{K}}_{\tilde{\lambda}}(\tilde{\boldsymbol{\psi}}-\mathbf{g})}_{\boldsymbol{\gamma},\ \tilde{f}_j(t)\ (j=1,\ldots\tilde{J})}$$

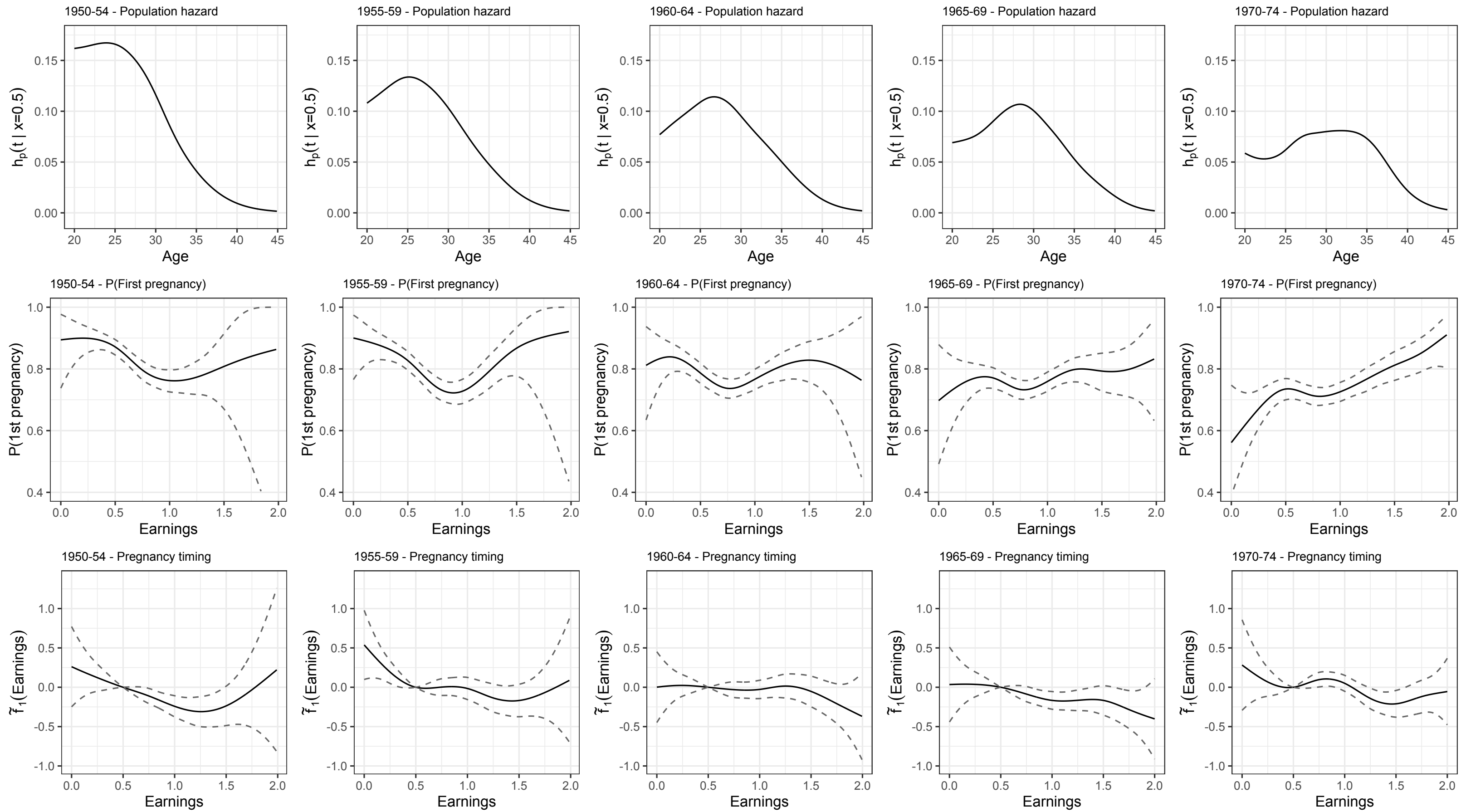
   ▷ Explicit forms exist for gradients and Hessian matrices.

2. Selection of the penalty parameters $\boldsymbol{\tau}=(\omega,\boldsymbol{\lambda},\tilde{\boldsymbol{\lambda}})$:

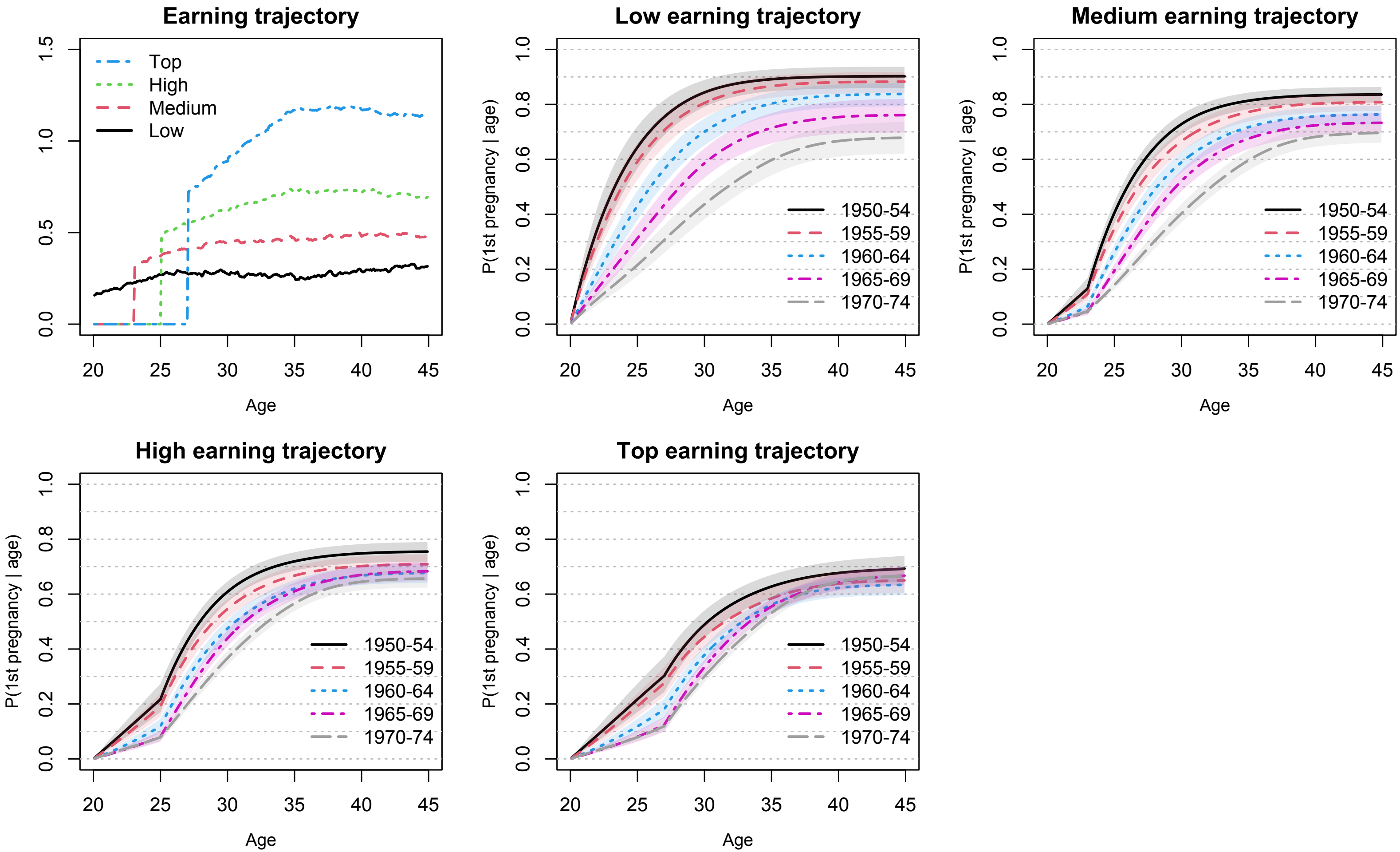
Maximize
$$\begin{aligned}p(\boldsymbol{\tau}|\mathcal{D})&=\frac{p(\boldsymbol{\zeta},\boldsymbol{\tau}|\mathcal{D})}{p(\boldsymbol{\zeta}|\boldsymbol{\tau},\mathcal{D})} &&\longleftarrow \text{Identity}\\ &\approx\frac{p(\boldsymbol{\zeta},\boldsymbol{\tau}|\mathcal{D})}{p_G(\boldsymbol{\zeta}|\boldsymbol{\tau},\mathcal{D})} &&\longleftarrow \text{Laplace approximation}\\ &\propto p(\hat{\boldsymbol{\zeta}}_\tau,\boldsymbol{\tau}|\mathcal{D})\left|\Sigma_\tau^{-1}\right|^{-1/2} &&\longleftarrow \text{\ldots evaluated at the MAP}\end{aligned}$$

⇝ Maximization can be made using the fixed point method.

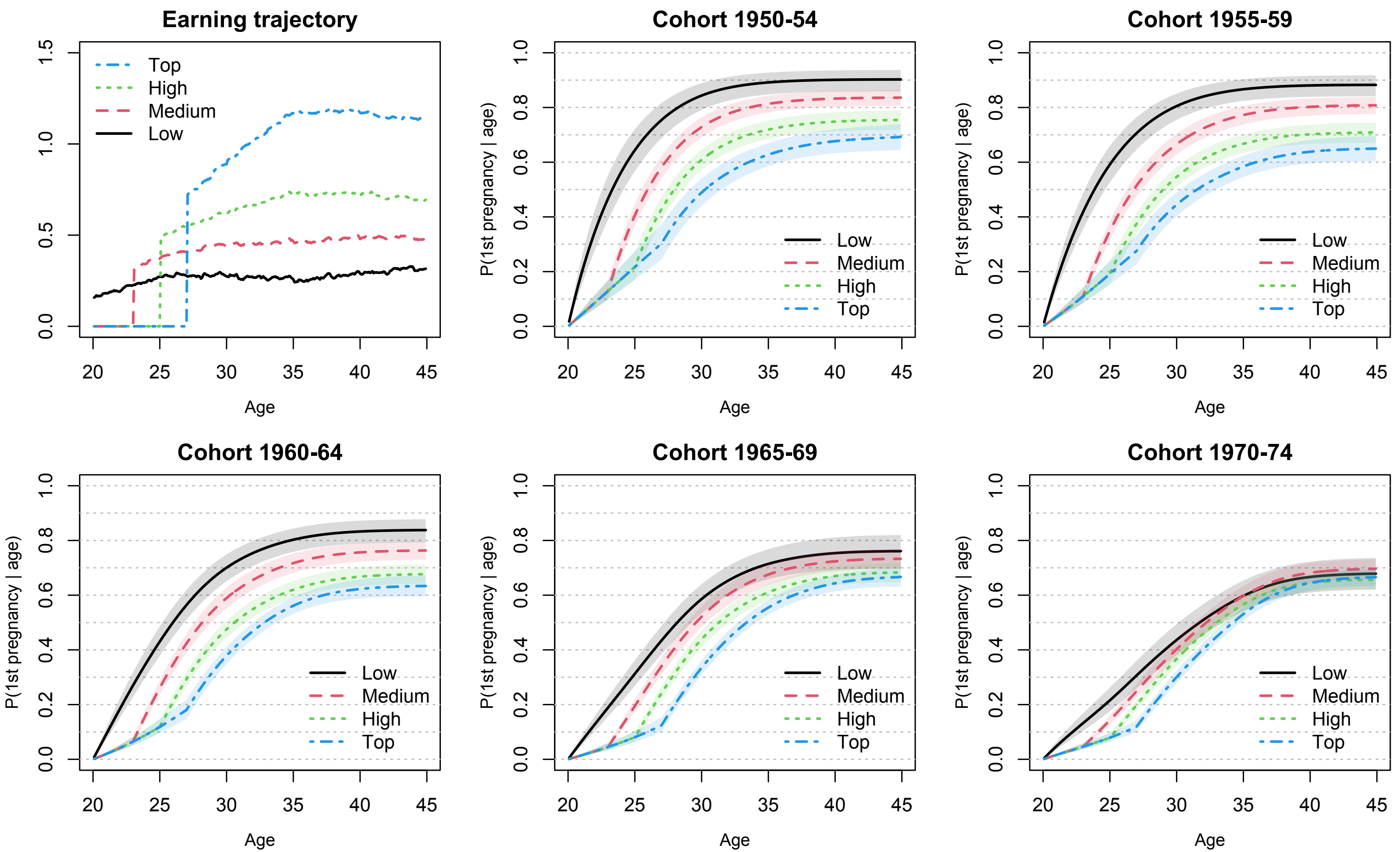
Estimated model components per cohort



Probability of pregnancy for prototypical earnings profiles



. . . from a cohort viewpoint



Discussion

- We described a very flexible cure survival model with:
  - ▷ Additive model for the cure probability ;
  - ▷ Flexible specification for the event dynamics.
- Time-varying covariates can be included in the two sub-models.
- Penalty parameters are selected to maximize their marginal posterior
  $\hookrightarrow$ no over-fitting (despite the rich information).
- Parameter interpretation in the presence of TV covariates is better understood by visualizing the estimated survival functions for given proptotypical trajectories.

  Here, the employment status and earnings of a woman can vary over time in a non trivial way.
- R-package `tvcure` will soon be available on CRAN and on Github

  https://github.com/plambertULiege/

Some references: http : //www.statsoc.ulg.ac.be/ p.lambert@uliege.be

- Lambert, P. and Kreyendeld, M. (2023). Exogenous time-varying covariates in double additive cure survival model with application to fertility. *arXiv:2302.00331*.
- Lambert, P. (2023) R-package **tvcure** - Soon on CRAN & GitHub: plambertULiege

- Lambert P. and Gressani, O. (2023). Penalty parameter selection and asymmetry corrections to Laplace approximations in Bayesian P-splines models. *Statistical Modelling*, **23(5-6)**: 409-423. R package **ordgam**.
- Lambert P. (2021). Fast Bayesian inference using Laplace approximations in nonparametric double additive location-scale models with right- and interval-censored data. *Computational Statistics and Data Analysis*, **161**: 107250. R package **DALSM**.
- Gressani, O. and Lambert P. (2021). Laplace approximation for fast Bayesian inference in generalized additive models based on P-splines. *Computational Statistics and Data Analysis*, **154**: 107088. R package **blapsr**.
- Gressani, O. and Lambert P. (2018). Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. *Computational Statistics and Data Analysis*, **124**: 151-167.

- Lambert, P. and Bremhorst, V. (2020). Inclusion of time-varying covariates in cure survival models with an application in fertility studies. *JRSS-A*, 183(1): 333-354.
- Lambert, P. and Bremhorst, V. (2019). Estimation and identification issues in the promotion time cure model when the same covariates influence long- and short-term survival. *Biometrical Journal*, 61: 275-279.
- Bremhorst, V., Kreyenfeld M. and Lambert P. (2019). Nonparametric double additive cure survival models: an application to the estimation of the nonlinear effect of age at first parenthood on fertility progression. *Statistical Modelling*, 19(3): 248-275.
- Bremhorst, V., Kreyenfeld M. and Lambert P. (2016). Fertility progression in Germany: an analysis using flexible nonparametric cure survival models. *Demographic Research*, 35: 505-534.