# Generalized linear model with two correlated interval-censored covariates linked by a Gumbel copula

Andrea Toloba López-Egea

Joint work with **Ingrid Van Keilegom**[*], **Klaus Langohr** and **Guadalupe Gómez Melis**

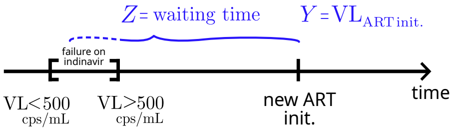
[*]KU Leuven, Belgium

Statistics and Operations Research Department and Institute for Research and Innovation in Health (IRIS)
Universitat Politècnica de Catalunya - BarcelonaTech

July 1st, 2026

# AIDS Clinical Trials Group Study 359: a real-data case from 2000

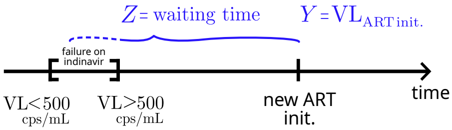
- Viral load at treatment start is a clinically relevant prognostic marker in HIV studies.
- This study enrolled individuals with **virologic failure on indinavir** who were about to start a new antiretroviral regimen.




Could the **waiting time** until the new regimen ($Z$)
be associated with **viral load at regimen initiation** ($Y$)?

# AIDS Clinical Trials Group Study 359: a real-data case from 2000

- Viral load at treatment start is a clinically relevant prognostic marker in HIV studies.
- This study enrolled individuals with **virologic failure on indinavir** who were about to start a new antiretroviral regimen.




Could the **waiting time** until the new regimen ($Z$)
be associated with **viral load at regimen initiation** ($Y$)?

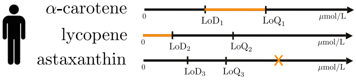
- Gómez, Espinal and Lagakos (2003)

Linear model with interval-censored time-to-event $Z$

$$\log_{10} Y = \alpha + \gamma Z + \beta \text{age} + \varepsilon, \quad \varepsilon \sim N(0,1)$$

# Study of carotenoids

- Carotenoids are a family of antioxidant plasma compounds acquired from vegetables, for example:



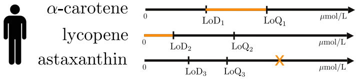
# Study of carotenoids

Research Institute of Nutrition and Food Safety
at the University of Barcelona

Spanish multicenter randomized trial
of primary cardiovascular prevention

Marhuenda-Muñoz M et al. (2022) Circulating carotenoids are associated with favorable lipid and fatty acid profiles in an older population at high cardiovascular risk. *Front Nutr*

- Carotenoids are a family of antioxidant plasma compounds acquired from vegetables, for example:



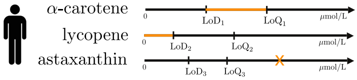
# Study of carotenoids

Research Institute of Nutrition and Food Safety
at the University of Barcelona

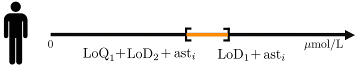
Spanish multicenter randomized trial
of primary cardiovascular prevention

Marhuenda-Muñoz M et al. (2022) Circulating carotenoids are associated with favorable lipid and fatty acid profiles in an older population at high cardiovascular risk. *Front Nutr*

- Carotenoids are a family of antioxidant plasma compounds acquired from vegetables, for example:




- $Z =$ Total circulating carotenoids



# Study of carotenoids

INSA UB

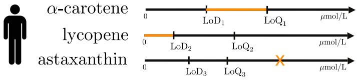
Research Institute of Nutrition and Food Safety
at the University of Barcelona

predimed plus

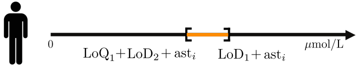
Spanish multicenter randomized trial
of primary cardiovascular prevention

Marhuenda-Muñoz M et al. (2022) Circulating carotenoids are associated with favorable lipid and fatty acid profiles in an older population at high cardiovascular risk. *Front Nutr*

- Carotenoids are a family of antioxidant plasma compounds acquired from vegetables, for example:




- $Z$ = Total circulating carotenoids




- $Z$ is interval-censored between $Z_L$ and $Z_R$

$$\mathbb{E}[Y|\boldsymbol{X}, Z] = g^{-1}(\alpha + \boldsymbol{\beta X} + \gamma Z)$$

Continuous interval-censored covariate

# Regression model with an interval-censored covariate

- Likelihood-based approach

$$\ell(\boldsymbol{\theta}, W(\cdot)) = \sum_{i=1}^{n} \log\left( \int_{z_{l_i}}^{z_{r_i}} f(y_i|\boldsymbol{x}_i, z; \boldsymbol{\theta})\, dW(z) \right) \qquad \text{where } W(z) = \Pr(Z \leq z).$$

(valid under non-informative censoring and conditional independence)

- GLM for $Y|\boldsymbol{X}, Z$:

$$f(y; \mu_i, \phi) = \exp\left\{ \frac{y\psi(\mu_i) - b(\psi(\mu_i))}{\phi} + c(y, \phi) \right\}, \qquad \begin{aligned} \mu_i &= \mathbb{E}[Y|\boldsymbol{x}_i, z] = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_i + \gamma z, \\ \boldsymbol{\theta} &= (\alpha, \boldsymbol{\beta}', \gamma, \phi)'. \end{aligned}$$

› If $W(z)$ is **known** (or estimated with negligible error) **plug it in** and maximize with respect to $\boldsymbol{\theta}$.

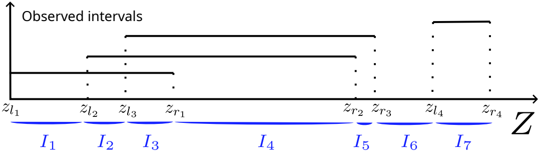
› When the estimation error in $W$ must be accounted for:

- Parametric $\widehat{W}(z; \lambda)$: jointly estimate $\lambda$ and $\boldsymbol{\theta}$
- Semi-/nonparametric $\widehat{W}$: iterative procedure alternating between estimation of $W$ and maximization of $\ell(\boldsymbol{\theta})$. Gómez et al. (2003), Langohr et al. (2004), Sizelove et al. (2024), Li et al. (2026)

# The *Gomez-Espinal-Lagakos continuous* (GELc) approach

- Observed intervals determine a partition for the support of $Z$, or domain of $W(\cdot)$

$$I_j = \lfloor a_j, b_j \rfloor,\ j = 1, \ldots, m_n \quad \text{such that} \quad \lfloor z_{l_i}, z_{r_i} \rfloor = \bigcup_{j: I_j \cap \lfloor z_{l_i}, z_{r_i} \rfloor \neq \varnothing} I_j$$

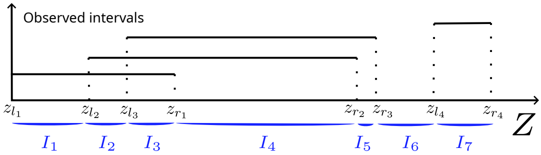



**Augmented Turnbull intervals**

# The *Gomez-Espinal-Lagakos continuous* (GELc) approach

- Observed intervals determine a partition for the support of $Z$, or domain of $W(\cdot)$

$$I_j = \lfloor a_j, b_j \rfloor,\ j = 1, \ldots, m_n \quad \text{such that} \quad \lfloor z_{l_i}, z_{r_i} \rfloor = \bigcup_{j: I_j \cap \lfloor z_{l_i}, z_{r_i} \rfloor \neq \varnothing} I_j$$




**Augmented Turnbull intervals**

- If $W(z|\boldsymbol{x}) = W(z)$, the **increase** of $W(\cdot)$ in these intervals can be estimated as the **fixed point of**

$$\begin{aligned} w_j^{(l+1)} &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[ Z \in I_j \mid y_i, \boldsymbol{x}_i, z_{l_i}, z_{r_i};\ \boldsymbol{w}^{(l)} \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \kappa_j^i \frac{w_j^{(l)} \cdot C_{ij}(\boldsymbol{\theta})}{\sum_{k=1}^{m} \kappa_k^i w_k^{(l)} \cdot C_{ik}(\boldsymbol{\theta})}, \qquad j = 1, \ldots, m \end{aligned}$$

$$\kappa_j^i = \mathbb{1}\{I_j \cap \lfloor z_{l_i}, z_{r_i} \rfloor \neq \varnothing\},$$

Indicator that subject $i$ contributes to estimating the increase of $W$ on $I_j$

$$C_{ij}(\boldsymbol{\theta}) = \left[ \int_{I_j} f(y_i | \boldsymbol{x}_i, z; \boldsymbol{\theta})\ dz \right] / |I_j|,$$

Average conditional density of $y_i$ given $\boldsymbol{x}_i$ over $z \in I_j$

# The GELc estimator $\hat{\theta}_{\text{GELc}}$

Consider the **class of distribution functions that are piecewise uniform** on the augmented Turnbull intervals:

$$\mathcal{W}_n = \left\{ W(z) = \sum_{k=1}^{j-1} w_k + (z - a_j)\frac{w_j}{|I_j|} \quad \text{for } z \in I_j = \lfloor a_j, b_j \rfloor,\ w_j \geq 0,\ j = 1, \ldots, m_n,\ \sum_{j=1}^{m_n} w_j = 1 \right\}$$

**Theorem** *[Toloba, Langohr, and Gómez-Melis]*. For fixed $\boldsymbol{\theta}$, the **augmented Turnbull estimator**

$$\widehat{W}_n(z; \boldsymbol{\theta}) = \sum_{k=1}^{j-1} \widehat{w}_k + (z - a_j)\frac{\widehat{w}_j}{|I_j|} \quad \text{for } z \in I_j = [a_j, b_j]$$

maximizes $\ell\big(\boldsymbol{\theta}, W(\cdot)\big)$ over $\mathcal{W}_n$.

Uniqueness and uniform convergence to the true $W$ hold under additional identifiability and regularity conditions.

The GELc estimator $\hat{\boldsymbol{\theta}}_{\text{GELc}}$ is obtained by alternating until convergence between

$$\widehat{W}^{(r+1)} = \operatorname*{argmax}_{W \in \mathcal{W}_n} \ell\big(\boldsymbol{\theta}^{(r)}, W\big) \quad \text{and} \quad \boldsymbol{\theta}^{(r+1)} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \ell\big(\boldsymbol{\theta}, \widehat{W}^{(r+1)}(\cdot; \boldsymbol{\theta}^{(r)})\big).$$

**Theorem**. If $\widehat{W}_n$ is uniformly consistent at a rate faster than $n^{-1/4}$, the effect of estimating $W$ on $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{GELc}} - \boldsymbol{\theta}_0)$ is asymptotically negligible and, under regularity conditions,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{GELc}} - \boldsymbol{\theta}_0) \xrightarrow[n \to \infty]{d} \mathcal{N}\left(\mathbf{0}, \mathcal{I}^{-1}\right).$$

Standard errors can be obtained from the Hessian with respect to $\boldsymbol{\theta}$ at the final iteration.

# The `ICenCov` R package

| | |
|---|---|
| `icglm` | Fit a GLM with one interval-censored covariate using GELc. |
| `plotint` | Visualize the observed censoring intervals. |
| `censoring` | Simulate non-informative interval-censored data. |
| **Model diagnostics** | `plot_ic_mapping`<br>`plot_icres_cdf`<br>`plot_icres_qq` |
| **Example datasets** | `carotenoids`<br>`actg359` |

github.com/atoloba/ICenCov

# Two interval-censored covariates linked by a Gumbel copula

- GLM $\qquad \mathbb{E}[Y|\boldsymbol{X},U,V] = g^{-1}(\alpha + \boldsymbol{\beta}\boldsymbol{X} + \gamma U + \zeta V), \qquad U$ and $V$ interval-censored

$$F_{U,V}(u,v) = \mathcal{C}_\xi\big(G(u),H(v)\big), \qquad G \text{ and } H \text{ marginal distributions}$$

$$= \exp\Big\{-\Big[\big(-\ln G(u)\big)^\xi + \big(-\ln H(v)\big)^\xi\Big]^{1/\xi}\Big\}, \qquad \xi \geq 1.$$

- Log-likelihood function

$$\ell(\boldsymbol{\theta},G,H,\xi) = \sum_{i=1}^{n} \log \int_{u_{l_i}}^{u_{r_i}} \int_{v_{l_i}}^{v_{r_i}} f(y_i|\boldsymbol{x}_i,u,v;\boldsymbol{\theta})\, c_\xi\big(G(u),\ H(v)\big)\, dG(u)\, dH(v).$$

- GELc estimation approach: Alternate between

$$\widehat{G}^{(r+1)} = \operatorname*{argmax}_{G\in\mathcal{G}_n} \ell\big(\boldsymbol{\theta}^{(r)},G,\widehat{H}^{(r)},\xi^{(r)}\big), \qquad \widehat{H}^{(r+1)} = \operatorname*{argmax}_{H\in\mathcal{H}_n} \ell\big(\boldsymbol{\theta}^{(r)},\widehat{G}^{(r+1)},H,\xi^{(r)}\big)$$

and $\quad (\boldsymbol{\theta},\xi)^{(r+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta,\xi\geq 1} \ell\big(\boldsymbol{\theta},\widehat{G}^{(r+1)},\widehat{H}^{(r+1)},\xi\big).$

# Extending the augmented Turnbull estimator to copula-linked covariates

The observed intervals for $U$ determine the augmented Turnbull intervals $I_j^{(U)} = \lfloor a_j, b_j \rfloor$ , $j = 1, \dots, m_U$.

Fixing $\boldsymbol{\theta}$, $\xi$, and $H$, and **assuming that $G$ is piecewise uniform**, the increments $w_j = G(b_j) - G(a_j)$ are updated by the self-consistency equation $(j = 1, \dots, m_U)$

$$w_j^{(l+1)} = \frac{\displaystyle\sum_{i=1}^{n} \frac{\kappa_{ij}^{(U)} \dfrac{w_j^{(l)}}{|I_j^{(U)}|} \displaystyle\int_{I_j^{(U)}} \int_{v_{l_i}}^{v_{r_i}} f(y_i|\boldsymbol{x}_i, u, v; \boldsymbol{\theta})\, c_\xi\big(G^{(l)}(u), H(v)\big)\, dH(v)\, du \;+\; \displaystyle\int_{u_{l_i}}^{u_{r_i}} \int_{v_{l_i}}^{v_{r_i}} f(y_i|\boldsymbol{x}_i, u, v; \boldsymbol{\theta})\, \dfrac{\partial}{\partial r} c_\xi(r,s)\Big|_{\substack{r = G^{(l)}(u)\\ s = H(v)}} \Delta_j(u)\, dH(v)\, dG^{(l)}(u)}{L_i(\boldsymbol{\theta}, G^{(l)}, H, \xi)}}{n + \displaystyle\sum_{i=1}^{n} \frac{\displaystyle\int_{u_{l_i}}^{u_{r_i}} \int_{v_{l_i}}^{v_{r_i}} f(y_i|\boldsymbol{x}_i, u, v; \boldsymbol{\theta})\, \dfrac{\partial}{\partial r} c_\xi(r,s)\Big|_{\substack{r = G^{(l)}(u)\\ s = H(v)}} G^{(l)}(u)\, dH(v)\, dG^{(l)}(u)}{L_i(\boldsymbol{\theta}, G^{(l)}, H, \xi)}},$$

$$\Delta_j^{(l)}(u) = \begin{cases} 0, & u \le a_j, \\ \dfrac{w_j^{(l)}}{|I_j^{(U)}|}(u - a_j), & a_j < u < b_j, \\ w_j^{(l)}, & u \ge b_j. \end{cases}$$

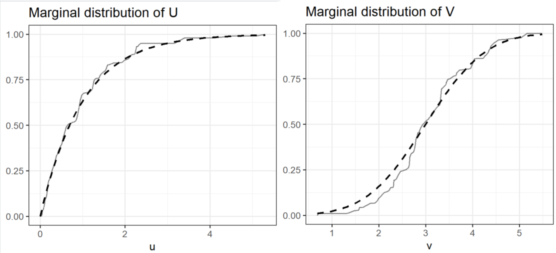
The equations for $H$ are analogous.

# Toy example: GELc with two interval-censored covariates

- Accelerated failure time model (AFT)

$$\log(T) = \alpha + \gamma U + \zeta V + \sigma E, \qquad E \sim \text{standard Gumbel.}$$

The covariates were simulated with marginals $U \sim \mathrm{Exp}(1)$ and $V \sim N(3,1)$ (under Kendall's $\tau = 0$), and then interval-censored. The response was generated with approximately $30\%$ right-censoring.




Estimated (solid lines) and true (dashed lines) marginal distributions.

| Parameters | $\alpha$ | $\gamma$ | $\zeta$ | $\sigma$ |
|---|---|---|---|---|
| True | $0$ | $0.50$ | $2.00$ | $0.70$ |
| Estimated | $1.27$ | $0.40$ | $1.89$ | $0.88$ |

# Summary and final remarks

- Interval-censored covariates arise naturally in biomedical studies, for example when biomarkers or event times are only known to lie within ranges of values.
- For one interval-censored covariate, the GELc approach combines:
  - a likelihood-based regression model for $Y \mid \boldsymbol{X}, Z$,
  - a nonparametric estimator of $W$ based on augmented Turnbull intervals,
  - an iterative maximization procedure alternating between $W$ and $\boldsymbol{\theta}$.
- The augmented Turnbull estimator is the NPMLE over the class of piecewise-uniform distribution functions, with consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}_{\mathrm{GELc}}$ under additional conditions.
- Current work is to extend the idea to two interval-censored covariates linked through a copula.
- Preliminary simulations suggest that the approach is promising, but further work is needed to study identifiability, asymptotic properties, and performance under heavier censoring.