Model fitting and goodness-of-fit for generalized linear models when covariates are interval-censored

Andrea Toloba Klaus Langohr Guadalupe Gómez Melis

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya•BarcelonaTech

July 11, 2024

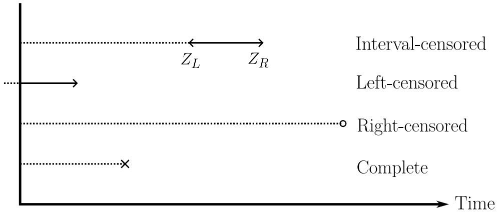
Overview

1. Interval-censored covariates
   - What's interval censoring?
   - Construction of the likelihood function
2. Parameter estimation

   Gómez G, Espinal A and Lagakos SW (2003) Inference for a linear regression model with an interval-censored covariate. *Stat in Med*, 22(3), 409–425
   - Alternative approach that doesn't rely on discretization
   - In the context of GLMs
3. Goodness-of-fit
   - Typical residuals for GLMs are not well-defined
   - Extending definitions / exploring new residuals (work in progress)
4. Chromatography illustration

Interval censoring: Survival illustration

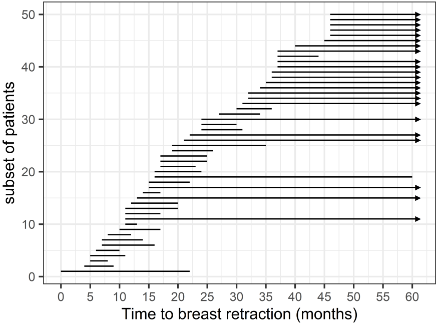
The **time-to-event variable** Z is interval-censored in $\lfloor Z_L, Z_R \rfloor$ if the exact value of Z is not observed, but it is known to lie within the **time interval** $\lfloor Z_L, Z_R \rfloor$.



Interval censoring: Survival illustration

The **time-to-event variable** Z is interval-censored in $\lfloor Z_L, Z_R \rfloor$ if the exact value of Z is not observed, but it is known to lie within the **time interval** $\lfloor Z_L, Z_R \rfloor$.

- **Response variable:** Time to breast retraction in early breast cancer patients
  - Radiotherapy and adjuvant chemotherapy v.s. Radiotherapy alone
  - **Main goal:** Effect of treatment in cosmetic appearance
  - Cosmetic deterioration = manifestation of breast retraction
  - Scheduled visits every 4 to 6 months



Beadle et al. (1984) Cosmetic results following primary radiation therapy for early breast cancer. *Cancer*, 54(12), 2911–2918

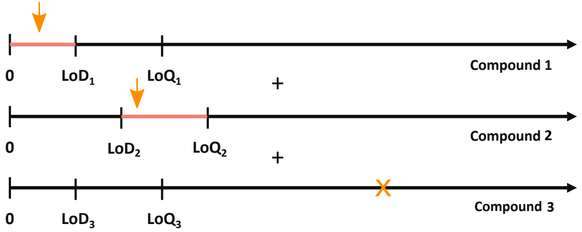
Interval censoring: Chromatography illustration

The **measurement variable** Z is interval-censored in $\lfloor Z_L, Z_R \rfloor$ if the exact value of Z is not observed, but it is known to lie within the interval $\lfloor Z_L, Z_R \rfloor$.

Interval censoring: Chromatography illustration

The **measurement variable** Z is interval-censored in $\lfloor Z_L, Z_R \rfloor$ if the exact value of Z is not observed, but it is known to lie within the interval $\lfloor Z_L, Z_R \rfloor$.

- **Explanatory variable:** Total plasma carotenoid concentration (Z)
  - Carotenoids are a family of antioxidant compounds that we obtain from fruits and vegetables.
  - **Main goal:** Predictive value of blood carotenoid concentration in cardiometabolic health.
  - Carotenoid components are measured in a laboratory using techniques with specific **limits of detection and quantification.**



$$Z \in \left[\ \text{LoD}_2 + C_3,\quad \text{LoD}_1 + \text{LoQ}_2 + C_3\ \right]$$

Interval censoring: Chromatography illustration

The **measurement variable** Z is interval-censored in $\lfloor Z_L, Z_R \rfloor$ if the exact value of Z is not observed, but it is known to lie within the interval $\lfloor Z_L, Z_R \rfloor$.

- **Explanatory variable:** Total plasma carotenoid concentration (Z)
  - Carotenoids are a family of antioxidant compounds that we obtain from fruits and vegetables.
  - **Main goal:** Predictive value of blood carotenoid concentration in cardiometabolic health.
  - Carotenoid components are measured in a laboratory using techniques with specific **limits of detection and quantification.**

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*, 9, 967967

INSA-UB: Research Institute of Nutrition and Food Safety at the University of Barcelona
Predimed-Plus: Spanish multicenter randomized trial of primary cardiovascular prevention

Gómez Melis G, Marhuenda-Muñoz M and Langohr K (2022) Regression Analysis with Interval-Censored Covariates. Application to Liquid Chromatography. In: Sun J and Chen DG (eds) *Emerging Topics in Modeling Interval-Censored Survival Data* (pp. 271–294)

Generalized linear model

$$\mu = \mathsf{E}(Y|\boldsymbol{X}, Z) = g^{-1}(\alpha + \boldsymbol{\beta'}\boldsymbol{X} + \gamma Z)$$

where

- $g(\cdot)$ monotonic differentiable link function
- $\boldsymbol{X} = (X_1, \ldots, X_p)'$ covariates
- Z with distribution function $W(\cdot)$ and $Z \in [Z_L, Z_R]$
- Y discrete or continuous, belonging to ψ-exponential family of distributions

$$f(y \mid \psi = \psi(\mu), \phi) = h(y, \phi) \exp[\{y\psi - a(\psi)\}/\phi]$$

- First two moments of Y: $\mu = \dot{a}(\psi)$ and $\mathrm{Var}(Y \mid \boldsymbol{X}, Z) = \phi\, \ddot{a}(\psi)$

Goal: Estimate $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta'}, \gamma, \phi)'$ where ϕ represents the dispersion of the model.

Likelihood functions: full and simplified

$$L_{\text{full}} = \prod_{i=1}^{n} P\big(Y \in dy_i, \boldsymbol{X} \in d\boldsymbol{x_i}, Z_i \in [z_{l_i}, z_{r_i}], Z_L \in dz_{l_i}, Z_R \in dz_{r_i}\big)$$

Likelihood functions: full and simplified

$$L_{\text{full}} = \prod_{i=1}^{n} P\big(Y \in dy_i, \boldsymbol{X} \in d\boldsymbol{x_i}, Z_i \in [z_{l_i}, z_{r_i}], Z_L \in dz_{l_i}, Z_R \in dz_{r_i}\big)$$

$$\begin{aligned}
L_{\text{simp}}(\boldsymbol{\theta}, W(\cdot)) &= \prod_{i=1}^{n} P\big(Y \in dy_i, \boldsymbol{X} \in d\boldsymbol{x_i}, Z_i \in [z_{l_i}, z_{r_i}]\big) \\
&= \prod_{i=1}^{n} \int_{z_{l_i}}^{z_{r_i}} f_{Y|\boldsymbol{X},Z}(y_i \mid \boldsymbol{x_i}, s;\, \boldsymbol{\theta})\, dW(s \mid \boldsymbol{x_i})\, P(\boldsymbol{X} \in d\boldsymbol{x_i})\, ds \\
&\propto \prod_{i=1}^{n} \int_{z_{l_i}}^{z_{r_i}} f_{Y|\boldsymbol{X},Z}(y_i \mid \boldsymbol{x_i}, s;\, \boldsymbol{\theta})\, dW(s)
\end{aligned}$$

Likelihood functions: full and simplified

$$L_{\text{full}} = \prod_{i=1}^{n} P\big(Y \in dy_i, \boldsymbol{X} \in d\boldsymbol{x_i}, Z_i \in [z_{l_i}, z_{r_i}], Z_L \in dz_{l_i}, Z_R \in dz_{r_i}\big)$$

$$\begin{aligned}
L_{\text{simp}}(\boldsymbol{\theta}, W(\cdot)) &= \prod_{i=1}^{n} P\big(Y \in dy_i, \boldsymbol{X} \in d\boldsymbol{x_i}, Z_i \in [z_{l_i}, z_{r_i}]\big) \\
&= \prod_{i=1}^{n} \int_{z_{l_i}}^{z_{r_i}} f_{Y|\boldsymbol{X},Z}(y_i \mid \boldsymbol{x_i}, s; \boldsymbol{\theta})\, dW(s \mid \boldsymbol{x_i})\, P(\boldsymbol{X} \in d\boldsymbol{x_i})\, ds \\
&\propto \prod_{i=1}^{n} \int_{z_{l_i}}^{z_{r_i}} f_{Y|\boldsymbol{X},Z}(y_i \mid \boldsymbol{x_i}, s; \boldsymbol{\theta})\, dW(s)
\end{aligned}$$

Assumptions for $L_{\text{simp}} \propto L_{\text{full}}$:

- Non-informative censoring [1]

$$dW(z \mid Z_L = z_l, Z_R = z_r) = \frac{dW(z)}{P(z_l \leq Z \leq z_r)}$$

- Y and (Z_L, Z_R) conditional independent given Z

[1] Oller R, Gómez Melis G and Calle ML (2004) Interval censoring: model characterizations for the validity of the simplified likelihood. *Can J Stat*, 32(3), 315–326

Observations y_i provide crucial information about $\widehat{W}$

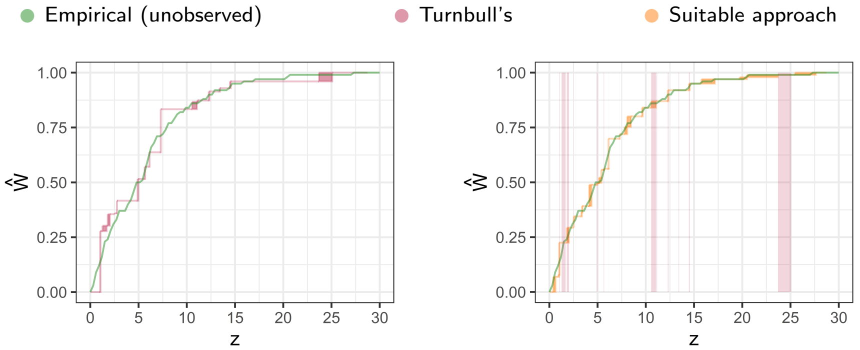


Fig.1: NPMLE of W using only $\{z_{l_i}, z_{r_i}\}_{i=1}^n$
i.e. **Turnbull's estimator**

Fig.2: NPMLE of W using $\{z_{l_i}, z_{r_i}, y_i\}_{i=1}^n$
i.e. the $\widehat{W}$ that maximizes

$$L(\boldsymbol{\theta},\, W(\cdot)) = \prod_{i=1}^{n} \int_{z_{l_i}}^{z_{r_i}} f(y_i \mid \boldsymbol{x_i}, s;\, \boldsymbol{\theta})\, dW(s)$$

Parameter estimation: an EM-type algorithm

Maximization of

$$l(\boldsymbol{\theta},\, W(\cdot)) = \sum_{i=1}^{n} \log\left\{ \int_{z_{l_i}}^{z_{r_i}} f(y_i \mid \boldsymbol{x_i}, s;\, \boldsymbol{\theta})\, dW(s) \right\}$$

over $\boldsymbol{\theta} \in \mathbb{R}^{p+2} \times \mathbb{R}^{+}$ and $W : \Omega \subseteq \mathbb{R} \to [0,1]$ distribution function.

- Set up initial conditions and iterate between the maximization of l with respect to W and $\boldsymbol{\theta}$.
- Differentiating the functional $l(W \mid \boldsymbol{\theta})$ and equating to zero yields the self-consistent equations in **A)**.
- The EM-type algorithm is defined by

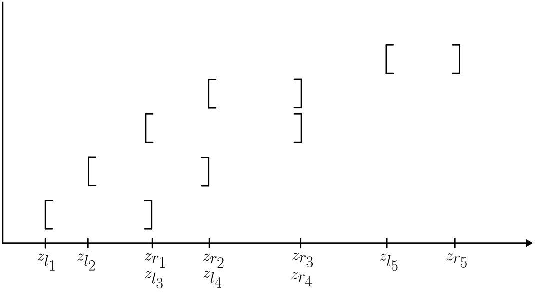
A) $$\widehat{W}(z_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\displaystyle\int_{z_{l_i}}^{z_{r_i}} f(y_i \mid \boldsymbol{x_i}, s;\, \hat{\boldsymbol{\theta}})\, dW(s \wedge z_0)}{\displaystyle\int_{z_{l_i}}^{z_{r_i}} f(y_i \mid \boldsymbol{x_i}, s;\, \hat{\boldsymbol{\theta}})\, dW(s)}$$

B) $$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{i=1}^{n} \log\left\{ \int_{z_{l_i}}^{z_{r_i}} f(y_i \mid \boldsymbol{x_i}, s;\, \boldsymbol{\theta})\, d\widehat{W}(s) \right\}$$

where $s \wedge z_0 = \min\{s, z_0\}$.

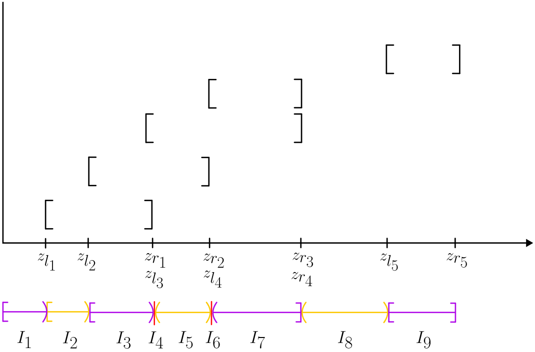
Parameter estimation: construction of partition intervals

$$l(\boldsymbol{\theta},\, W(\cdot)) = \sum_{i=1}^{n} \log \left\{ \int_{z_{l_i}}^{z_{r_i}} f(y_i \mid \boldsymbol{x_i}, s;\, \boldsymbol{\theta})\, dW(s) \right\}$$



Parameter estimation: construction of partition intervals

$$l(\boldsymbol{\theta},\, W(\cdot)) = \sum_{i=1}^{n} \log\left\{ \int_{z_{l_i}}^{z_{r_i}} f(y_i \mid \boldsymbol{x_i}, s;\, \boldsymbol{\theta})\, dW(s) \right\}$$



$\{I_j\}_{j=1}^{m_n}$ is a partition of the support $\Omega = [0, z_{r_5}]$ such that

$$l(\boldsymbol{\theta},\, W(\cdot)) = \sum_{i=1}^{n} \log\left\{ \sum_{j=1}^{m_n} \kappa_j^i \int_{I_j} f(y_i \mid \boldsymbol{x_i}, s;\, \boldsymbol{\theta})\, dW(s) \right\}$$

where $\kappa_j^i = \mathbb{1}\{I_j \subseteq \lfloor z_{l_i}, z_{r_i} \rfloor\}$.

Parameter estimation: redefinition of the maximization problem

- **Assume W is uniform in I_j for all $j = 1, \ldots, m_n$.**

- Then the maximization problem rewrites to

$$l(\boldsymbol{\theta},\, \boldsymbol{w}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m_n} \kappa_j^i \frac{\hat{w}_j}{|I_j|} \int_{I_j} f(y_i \mid \boldsymbol{x_i}, s;\, \boldsymbol{\theta}) ds \right\}$$

where $|I_j|$ denotes the length of I_j,

over $\boldsymbol{\theta} \in \mathbb{R}^{p+2} \times \mathbb{R}^{+}$ and $\boldsymbol{w}$ s.t. $\sum^{m_n} w_j = 1$ and $w_j \geq 0$.

- And the EM-type algorithm ($j = 1, \ldots, m_n$):

$$\text{A)} \quad w_j^{(l+1)} = \frac{1}{n} \sum_{i=1}^{n} \kappa_j^i \, \frac{\dfrac{w_j^{(l)}}{|I_j|} \displaystyle\int_{I_j} f(y_i \mid s; \hat{\boldsymbol{\theta}}) ds}{\displaystyle\sum_{k=1}^{m_n} \kappa_k^i \, \frac{w_k^{(l)}}{|I_k|} \int_{I_k} f(y_i \mid s; \hat{\boldsymbol{\theta}}) ds}$$

$$\text{B)} \quad \hat{\boldsymbol{\theta}} = \ \text{argmax } l(\boldsymbol{\theta} \mid \hat{\boldsymbol{w}})$$

Solved by Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm, a quasi-Newton method for the numerical search of local maxima.

Diagnostics for GLM assumptions: Pearson and deviance residuals

- **Pearson residuals** are defined as $r_i^{(P)} = (y_i - \hat{\mu}_i)/\sqrt{V(\hat{\mu}_i)}$, where $V(\cdot)$ is the variance function in $\mathrm{Var}(Y_i) = \phi V(\mu_i)$. Asymptotic normality of $r_i^{(P)}$ follows from the Central Limit Theorem (CLT) applied to Y_i.

Asymptotics for Pearson residuals in case of $Y_i \sim$ Gamma with shape ν and scale $\lambda_i = \mu_i/\nu$.

$$Y_i = \sum_{k=1}^{\nu} U_k \text{ with } U_k \sim_{i.i.d.} \mathrm{Exp}(1/\lambda_i)$$

By the CLT, the Pearson residual $\sqrt{\nu}\,\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i^2}} \to_d N(0,1)$ as $\nu \to \infty$.
For the ith Pearson residual to be asymp. normal, the data dispersion $\phi = 1/\nu$ should be low.

- **Deviance residuals** are defined as $r_i^{(D)} = \mathrm{sgn}(y_i - \hat{\mu}_i)\sqrt{d(y_i, \hat{\mu}_i)}$, where $d(y_i, \hat{\mu}_i) = 2\left\{y_i\big(\psi(y_i) - \psi(\hat{\mu}_i)\big) - b(\psi(y_i)) + b(\psi(\hat{\mu}_i))\right\}$ is the unit deviance. Asymptotic normality derives from the saddle-point approximation of Y_i's distribution to the normal.

- **Discard Pearson and deviance residuals** because their asymptotics are approximations that, in most cases, do not hold.[2]

[2] Smyth GK and Dunn PK (2018) Generalized Linear Models With Examples in R. Section 8.6.

Diagnostics for GLM assumptions: Quantile residuals

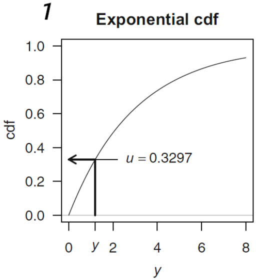
In the context of a GLM defined by $E[Y \mid \boldsymbol{X_i}, Z_i] = \mu_i = g^{-1}(\alpha + \boldsymbol{\beta}'\boldsymbol{X_i} + \gamma Z_i)$, with predicted mean $\hat{\mu}_i = g^{-1}(\hat{\alpha} + \hat{\boldsymbol{\beta}}'\boldsymbol{x_i} + \hat{\gamma} z_i)$,

Quantile residuals are defined by

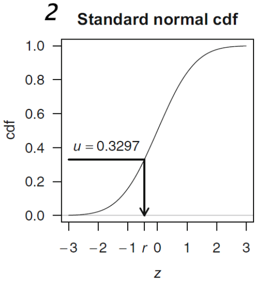
$$r_i = \Phi^{-1}(F(y_i; \hat{\mu}_i, \hat{\phi})),$$

where Φ is the cdf of the standard Normal distribution.

- Consider a gamma GLM with $\phi = 1$ fitted to data
- Observation with $y = 1.2$ and $\hat{\mu} = 3$




```
> y <- 1.2; mu <- 3
> cum.prob <- pexp(y, rate=1/mu)
[1] 0.32968
```




```
> rq <- qnorm(cum.prob)
[1] -0.4407971
```

Smyth GK and Dunn PK (2018) Generalized Linear Models With Examples in R. Section 8.3.4

Diagnostics for the distributional assumption

- Denote by F^* the true distribution of Y_i. Then $U_i = F^*(y_i) \sim U(0,1)$.
- Quantile residuals are normally distributed if $F(\cdot; \hat{\mu}_i, \hat{\phi})$ is good enough for each i.
- $F(y_i; \hat{\mu}_i, \hat{\phi}) = F(y_i \mid X = \boldsymbol{x_i}, Z = z_i; \hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\phi})$, so **we define**

$$\begin{aligned} r_i &= \Phi^{-1}(F(y_i \mid \boldsymbol{x_i}, Z_i \in \lfloor z_{l_i}, z_{r_i} \rfloor; \hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\phi})) \\ &= \Phi^{-1}(E_{Z_i}[F(y_i; \hat{\mu}_i, \hat{\phi}) \mid z_{l_i}, z_{r_i}]) \end{aligned}$$

- It is defined under the true distribution of Z_i. We choose to estimate r_i assuming that Z_i is uniformly distributed within $\lfloor z_{l_i}, z_{r_i} \rfloor$.

Simulation analysis to assess the power of these residuals in validating the distribution assumption.

Next steps

- Possible improvements of the estimation algorithm
  - ☐ Alternatives to the assumption of W being uniform in I_j
  - ☐ Elaborate a B step analogous to IRLS to improve computational efficiency
  - ☐ Kuhn–Tucker conditions to check that $\widehat{W}$ is a global maximum
- Check consistency of the estimator $\hat{\boldsymbol{\theta}}$
- Derive standard error and confidence intervals for $\hat{\boldsymbol{\theta}}$
- Adapt diagnostic tools for GLM assumptions
  - ☑ Quantile residuals to check the distributional assumption
  - ☐ Working residuals[3] to check the linearity of covariates and link function assumptions
  - ☐ Outliers / influential observations (Cook's distance[3])

If everything goes as planned, we'll be publishing by the end of September!

[3] McCullagh P and Nelder JA (1989) Generalized linear models, 2nd ed. Sections 2.5.1 and 12.6

Chromatography illustration

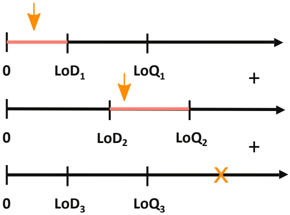
Marhuenda-Muñoz M et al. (2022) Circulating carotenoids are associated with favorable lipid and fatty acid profiles in an older population at high cardiovascular risk. *Front Nutr*, 9, 967967

INSA-UB: Research Institute of Nutrition and Food Safety at the University of Barcelona

Predimed-Plus: Spanish multicenter randomized trial of primary cardiovascular prevention

Total plasma carotenoid concentration (Z)

- Carotenoids are a family of antioxidant compounds that we obtain from fruits and vegetables.
- **Main goal:** Predictive value of blood carotenoid concentration in cardiometabolic health.
- Carotenoid components are measured in a laboratory using techniques with specific **limits of detection and quantification.**



$$Z \in \left[\ \text{LoD}_2 + C_3, \quad \text{LoD}_1 + \text{LoQ}_2 + C_3\ \right]$$

Chromatography illustration

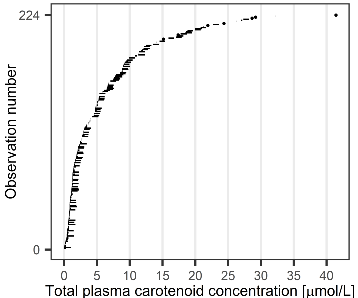
Marhuenda-Muñoz M et al. (2022) Circulating carotenoids are associated with favorable lipid and fatty acid profiles in an older population at high cardiovascular risk. *Front Nutr*, 9, 967967

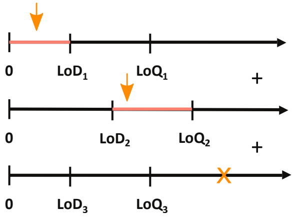
INSA-UB: Research Institute of Nutrition and Food Safety at the University of Barcelona

Predimed-Plus: Spanish multicenter randomized trial of primary cardiovascular prevention

Total plasma carotenoid concentration (Z)

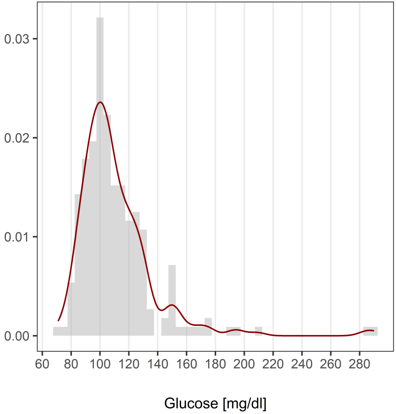
- Carotenoids are a family of antioxidant compounds that we obtain from fruits and vegetables.
- **Main goal:** Predictive value of blood carotenoid concentration in cardiometabolic health.
- Carotenoid components are measured in a laboratory using techniques with specific **limits of detection and quantification.**





$$Z \in \left[\ \text{LoD}_2 + C_3, \quad \text{LoD}_1 + \text{LoQ}_2 + C_3\ \right]$$

$E[\text{glucose}] = g^{-1}(\alpha + \gamma \cdot \text{Total plasma carotenoid concentration})$



- Y has **right-skewed distribution** $\rightarrow g = \log$

i.e. assume Z_i is related to Y_i in log scale

$$E[Y_i \mid Z_i] = \exp\{\alpha + \gamma Z_i\}$$

- $Y_i \mid Z_i$ Gamma or Gaussian distributed

Gaussian $\rightarrow \mathrm{Var}(Y_i \mid Z_i) = \phi$

Gamma $\rightarrow \mathrm{Var}(Y_i \mid Z_i) = \phi\mu_i^2$

Estimation results

$$E[\text{glucose}] = g^{-1}(\alpha + \gamma \cdot \text{Carotenoid concentration})$$

Regression parameters:

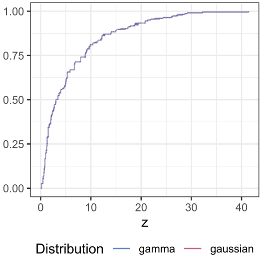
	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\phi}$
Gaussian	4.76	-0.009	720
Gamma	4.76	-0.008	0.043

The distinction is on the variance:

Gaussian $\rightarrow \mathrm{Var}(Y \mid Z_i) = 720$

Gamma $\rightarrow \mathrm{Var}(Y \mid Z_i) \in [312, 584]$

The resulting $\widehat{W}$ under both models:



For $z \in I_k = \lfloor q_j, p_j \rfloor$,

$$\widehat{W}(z) = \sum_{I_j \prec I_k} \hat{w}_j + \hat{w}_k \frac{z - q_k}{p_k - q_k}$$

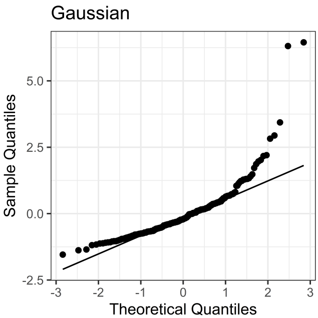
- Mean glucose levels decrease 0.9% for each unit increase in total plasma carotenoid concentration $\left(E[Y \mid z+1] = e^{\hat{\gamma}} \times E[Y \mid z]\right)$.
- From interval-censored measurements, the model is able to identify the non-parametric estimator distribution of W.

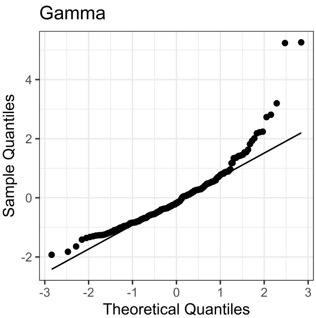
Quantile residuals

For each model and individual i,

$$r_i = \Phi^{-1}(F(y_i \mid \boldsymbol{x_i}, Z_i \in \lfloor z_{l_i}, z_{r_i} \rfloor; \hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\phi}))$$
$$= \Phi^{-1}(E_{Z_i}[F(y_i; \hat{\mu}_i, \hat{\phi}) \mid z_{l_i}, z_{r_i}])$$

and $r_i \sim N(0, 1)$ if the distribution resembles the true one.





Summary

- We have developed an algorithm for modeling responses with interval-censored covariates that does not require prior knowledge of the covariate support.

- Essential: derive the standard error and asymptotic distribution to provide confidence intervals for $\hat{\boldsymbol{\theta}}$.
- Desirable: proof for the consistency of $\hat{\boldsymbol{\theta}}$.

- Provide an R package to facilitate its use in applied research.