Generalized linear models with interval-censored covariates

Andrea Toloba[1,2], *Klaus Langohr*[2], *Guadalupe Gómez Melis*[2]

[1]andrea.toloba@upc.edu

[2]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya

Abstract

Interval-censored covariates have been appearing in clinical studies lately, but the lack of methodological content has not allowed researchers to analyse such data properly. We introduce the GEL technique for generalized linear models, and develop novel work on model checking. The proposed methodology is illustrated with data from metabolomics research area.

Keywords: interval censoring; generalized linear models; residual analysis

1. Introduction

Interval-censoring is typically encountered in the analysis of time-to-event data, and so statistical methods to analyse such data have been extensively studied; see, for example, Gómez et al (2009) for a thorough review [1]. On the contrary, scientific literature on regression models with an interval-censored covariate is rather scarce.

The GEL technique was first introduced in Gómez et al (2003) to estimate the regression parameters of a simple linear model where the independent variable is an interval-censored time [2]. Langohr et al (2014) revisited it and provided an implementation in `R` of the algorithm [3]. More recently, Morrison et al (2022) adapted the GEL technique to accommodate left-truncated interval-censored data, and Gómez et al (2022) rewrote the algorithm for generalized linear models [4, 5].

The definition of residuals has been even less explored than model estimation itself. Concerning linear regression, Langohr et al (2014) offer an exhaustive discussion of residuals in presence of interval-censoring [3]. The aim of this work is hence to propose a residual analysis procedure that accounts for interval-censoring, and address its suitability to check the model's goodness of fit.

2. Motivating data

The development of the methodology presented is motivated by metabolic data from 104 female participants of the PREDIMED-Plus trial. Clinical interest was on association of plasma carotenoid concentrations and cardiovascular risk factors [5]. Briefly, carotenoids are a family of eight phytochemical compounds produced by plants that are thought to be responsible for the health benefits fruits and vegetables provide, and are present in human blood acquired through diet. In order to study the global action of carotenoids as a whole, the sum of plasma concentrations was a relevant biomarker.

The plasma concentration of a carotenoid is measured in the laboratory by a technique called high-performance liquid chromatography. This technique is able to identify, extract and quantify the molecules of the compound, but it fails when concentrations are too low, so a mass spectrometry detection method is usually employed to obtain narrower intervals for the measure. Consequently, the data of a single plasma carotenoid determination depends on a limit of detection (LoD) and a limit of quantification (LoQ), in such a way that the value is exactly known over the LoQ, interval-censored between LoQ and LoD, and left-censored below LoD. These limits are compound-specific, so the sum of plasma concentrations gives rise to an interval-censored variable where the limits appear blurred, hence the resulting observed intervals are overlapping. This feature is an advantage over single determinations, for which very little information can be drawn under the limit of quantification.

3. Model formulation

For each carotenoid compound C_j, let C_{Lj}, C_{Rj} be two random variables denoting the observable intervals $[C_{Lj}, C_{Rj}]$, which correspond to the potential observations $[0, \text{LoD}_j)$, $[\text{LoD}_j, \text{LoQ}_j)$, and $[C_j, C_j]$. Note the observable intervals can be assumed closed because any quantification method has a limited decimal precision. The sum of carotenoids is denoted by Z, and the corresponding observable random variables are defined by $Z_L = \sum_{j=1}^{8} C_{Lj}$ and $Z_R = \sum_{j=1}^{8} C_{Rj}$. Hence, Z is an interval-censored variable, and the censoring interval $[Z_L, Z_R]$ verifies the non-informative conditions particular of this type of censoring. Figure 1 illustrates the observed data for one of the carotenoids, and for the sum of all of them. Additionally, let Y be a random variable from the exponential family denoting the response variable, and X a p-dimensional vector of fully-observed covariates.

Generalized linear models are an extension of classical linear models that aim to fit regression models with response variables whose distribution cannot be approximated by a normal random variable. These models relate the expected mean response $\mu = E[Y \mid X, Z]$ with the linear predictor $\eta = \alpha + \beta' X + \gamma Z$ through a monotonic differentiable link function g, that is $g(\mu) = \eta$. Many of the properties they possess arise from assuming that the probability density function of Y, in terms of η, pertains to the θ-parameter exponential family with shape

$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y; \phi)\right\},$$

where ϕ is known as dispersion parameter, and θ is a function of the mean $\mu = g^{-1}(\eta)$. For instance, an expression can be found for the variance of Y, $\text{Var}(Y) = a(\phi)\ddot{b}(\theta)$, which implies that it is not constant with respect to the mean μ; and the variance function $V(\mu)$ is defined so that $\text{Var}(Y) = a(\phi)V(\mu)$.

4. Parameter estimation

The GEL (Gómez - Espinal - Lagakos) technique is a parameter estimation method based on maximizing the log-likelihood function through an EM-type algorithm that contemplates an interval-censored covariate. It assumes the interval-censored covariate Z is discrete with support $\mathcal{S} = \{s_1, s_2, \ldots, s_m\}$ and probability law $w_j := P(Z = s_j)$ for $j = 1, \ldots, m$, yet no additional distributional restrictions are made.

Non-informative conditions are assumed in the sense that Z_L, Z_R provide no additional information about Z than being inside the interval, and that Y depends on Z_L, Z_R only through Z. These are standard requirements for interval-censored data, and can be found for instance in Gómez et al (2009) [1].

Given n independent realizations of (Y, X, Z_L, Z_R), denote by $(y_i, x_i, z_{Li}, z_{Ri})$ the observed data of the ith individual. Non-informative conditions make it possible to ignore the distribution function of Z_L, Z_R from the log-likelihood maximization, since the likelihood function becomes proportional to

$$L(\theta, \phi, w) \propto \prod_{i=1}^{n} \sum_{j=1}^{m} \kappa_{ij}\ f(y_i \mid Z = s_j, X = x_i;\ \theta, \phi)\ w_j, \tag{1}$$

where $\kappa_{ij} = 1\{s_j \in [z_{Li}, z_{Ri}]\}$ indicates whether the support point s_j is included in the observed interval of the ith individual, and $w = (w_1, \ldots, w_m)$ is a parameter vector enclosing the probabilities of Z in $\mathcal{S}$. In order to estimate α, β, γ in presence of the nuisance parameter w, an iterative two-step algorithm is proposed:

1. Consider $\theta = \theta(\mu_i)$ fixed, and solve the following self-consistent equations for w.

$$w_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\kappa_{ij} f(y_i \mid Z = s_j, X = x_i;\ \theta) w_j}{\sum_{k=1}^{m} \kappa_{ik} f(y_i \mid Z = s_k, X = x_i;\ \theta) w_k}, \quad j = 1, \ldots, m$$

2. Given an estimate for w, the log-likelihood is maximized with respect to regression parameters α, β, γ. Notice that (1) does not correspond to the particular likelihood of generalized linear models, so standard tools such as iterative reweighted least squares or Fisher scoring algorithm are not applicable, and generic numerical algorithms for maximization shall be used instead.

5. Residual analysis

After model fitting it is indispensable to check the adequacy of the selected model to the data, since misspecification may invalidate findings and predictions. Common sources of model misspecification include a bad choice of link function g, a wrong assumption of the mean-variance relationship induced by the variance function $V(\mu)$, a wrong pre-specified dispersion ϕ, the possible omission of non-linear effects from covariates, and the presence of unusual response observations that might influence parameter estimation.

Diagnostic plots of residuals can reveal and help to identify these problems. For instance, plotting residuals against fitted values to seek patterns provides a first glimpse of systematic inconsistencies in the model, whereas non-linear effects can be checked with component-plus-residual plots, which show numeric covariates against its partial residuals. In addition, goodness-of-fit tests computed from residuals offer a valuable overview of model suitability.

It is well known that Pearson and deviance residuals, together with their standardized form, are the most employed in generalized linear models. Briefly, the former are defined as $r_{P,i} = (y_i - \hat{\mu}_i)/\sqrt{V(\hat{\mu}_i)}$, and arise from the Pearson's χ^2 statistic. Conversely, deviance residuals $r_{d,i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$ are defined so that $D = \sum_i d_i$, where $D = 2\phi\{l(y) - l(\hat{\mu})\}$ is the deviance of the model. Unfortunately, both residuals rely on exact values of covariates, so are not totally suited for interval-censored data.

In this work we propose a residual analysis procedure based on the *pseudo-observations* estimating method originally conceived for incomplete data [6]. It will be compared with the alternative approach of computing the expected residuals $\hat{r}_{P,i} = E_Z[r_{P,i} \mid z_{Li}, z_{Ri}]$, $\hat{r}_{d,i} = E_Z[r_{d,i} \mid z_{Li}, z_{Ri}]$ under the Turnbull's non-parametric estimator of $P(Z = s_j)$.

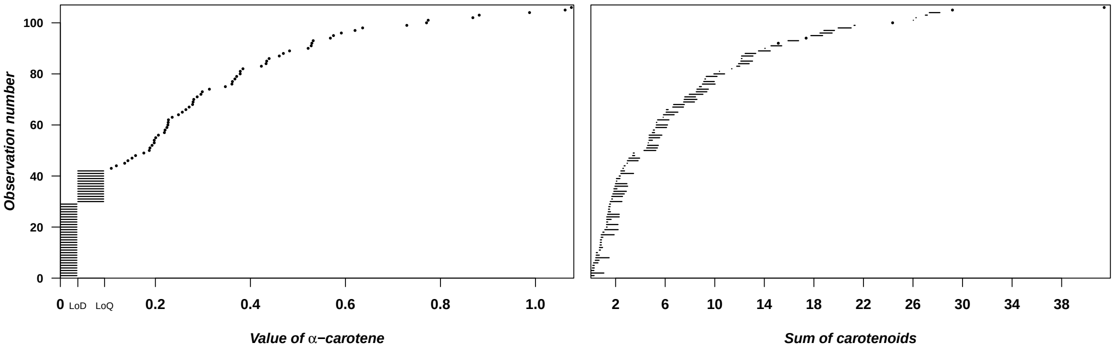

Figure 1: Ordered observed data from 104 female participants of a study on the association of circulating carotenoids and cardiovascular risk factors. Lines represent interval-censored observations, points exactly quantified determinations.

6. Aknowledgements

This work was funded by the Ministerio de Ciencia e Innovación (Spain) [Grant: PID2019-104830RB-I00/ DOI (AEI): 10.13039/501100011033] and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya (Spain) [Grant: 2021 SGR 01421].

7. Bibliography

[1] Gómez G., Calle M.L., Oller R., and Langohr K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling*, 9 (4), 259–297.

[2] Gómez G., Espinal A., and Lagakos S.W. (2003). Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine*, 22 (3), 409–425.

[3] Langohr K. and Gómez G. (2014). Estimation and residual analysis with R for a linear regression model with an interval-censored covariate. *Biometrical Journal*, 56 (5), 867–885.

[4] Morrison D., Laeyendecker O., and Brookmeyer R. (2022). Regression with interval-censored covariates: Application to cross-sectional incidence estimation. *Biometrics*, 78 (3), 908–921.

[5] Gómez G., Marhuenda-Muñoz M., and Langohr K. (2022). Regression analysis with interval-censored covariates. Application to liquid chromatography. In Sun J., Chen DG. (eds) *Emerging topics in modeling interval-censored survival data*. ICSA Book Series in Statistics. Springer, Cham.

[6] Andersen P.K., Klein J.P., and Rosthøj S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90 (1), 15–27.