A flexible modelling framework using linear splines for censored regression models to estimate serological profiles.

Lower limit of detection: censored regression

The lower limit of detection limit means that observations below the DL are left-censored. To account for this, we use censored regression, specifically we assume a Gaussian linear model for y, the natural logarithm of the measured IgG concentration, with a single explanatory variable, x, age:

$$y = \beta_0 + \beta_1 x + \epsilon = f(x, \beta) + \epsilon \qquad (1)$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

We do not directly observe y, but only y_M, where values below the detection limit are censored:

$$y_M = \begin{cases} y, & y > \tau \\ \tau, & y \leq \tau \end{cases}$$

where τ is the lower limit of detection ($\tau = 0.15$ ug/mL in our data).

Since we assume a Gaussian model, this allows us to write down, for an observed data set $\{x_i, y_{M,i}\}_{i=1}^n$, the log-likelihood below[1], thereby making the usual maximum likelihood inference theory available:

$$l(\beta, \sigma^2) = \sum_{yM,i>\tau} \log\left(\phi((y_{M,i} - f(x_i, \beta))/\sigma)\right) + \sum_{y_{M,i} \leq \tau} \log\left(\Phi(f(x_i, \beta) - \tau)\right) \qquad (2)$$

Identify and estimate change points: linear regression splines

To address the objectives, we need fit a model that allows the linear relationship between age and IgG to change over time. A natural candidate for this are linear splines. Splines fit piece-wise linear regression models, allowing the slope to change at specific values of the predictor variable. These locations, commonly referred to as knots, define intervals over which the relationship between predictor and response is linear.

Mathematically, for a response y and a predictor x, a linear spline with 3 knots at locations $x = k_1, k_2, k_3$ can be written down as:

$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot (x - k_1)_+ + \beta_3 \cdot (x - k_2)_+ + \beta_4 \cdot (x - k_3)_+ + \epsilon \qquad (3)$$

where $(x)_+ = x$ if $x \geq 0$ and 0 otherwise.

This means that the fitted curve has different slopes in different intervals of values of x:

interval	slope
(-Inf, k1)	β_1
[k_1,k_2)	$\beta_1 + \beta_2$
[k_2,k_3)	$\beta_1 + \beta_2 + \beta_3$
[k_3,Inf)	$\beta_1 + \beta_2 + \beta_3 + \beta_4$

Statistical inference for number and locations of knots: bootstrap and AIC

Usually splines are used to produce smooth functional fits (typically one would use cubic rather than linear functions) and knot locations are either user-provided or set to regular intervals in the predictor variable.

For the purpose of the pneumococcus serology data however, the number and the locations of the knots, i.e. the change points, are of interest themselves and we would like to do statistical inference on the knot locations. To this end, we treat the knot locations $k_1, k_2, k_3, \ldots$ as model parameters and estimate them using numerical optimisation. We then use the bootstrap to derive confidence intervals for the locations of the knots.

To identify which number of change points best fits the data, we fit models with several numbers of knots, then use Akaike Information Criterion (AIC) to select the optimal number of knots.

Putting it all together

Given the nature of the data (IgG titre with a lower detection limit) and the fact that we consider several plausible biological models for vaccine response profiles (with changing slopes) for which it is important to know when the changes in IgG levels occur, we need our modelling framework to satisfy several requirements:

1. Properly account for data values that are left censored (below the detection limit).
2. Treat the knot locations as parameter values to be estimated during model fitting.
3. Return a fit for the geometric mean IgG concentration as a function of age.
4. Select the optimal number of knots / changepoints.

This means that our model is defined by the likelihood (2), with $f(x, \beta)$ in (1) replaced by $f(x, \beta, k)$ using the functional form from (3).

In summary, the key characteristics of our modelling framework are:

1. We fit censored regression models with linear splines for the age predictor variable.
2. Knot locations are treated as model parameters and fitted by numerical maximisation of the likelihood of the censored regression model. We use bootstrap techniques to derive confidence intervals for slope parameters (as well as predicted average concentration at any given age) and knot locations. Specifically, the bootstrap percentile method[2] is used for CIs.
3. Before model fitting, the IgG data are log-transformed, so that the fitted arithmetic mean corresponds to the log of the geometric mean on the original data scale (i.e. we can exponentiate the fitted curve to have a fit for the geometric mean).
4. For every serotype, models are fitted with $k = 0, 1, 2, 3$ knots / changepoints and the optimal value for k is selected using AIC.

Our implementation was done in the `R` environment for statistical programming, making use of packages `rms`[3] (for the linear spline terms), `censReg`[4] (for the censored regression likelihood), `boot`[2,5] (for the bootstrapped CIs) and custom code to estimate model parameters and selecting the number of knots / changepoint.

Conclusion

We present a flexible modelling framework to estimate both the number and location of changepoints in serological profiles. Key benefits of our approach are the principled statistical analysis of serological profile data which are increasingly common and the grouping of profiles, giving insights into the underlying biology.

References

1. Kleiber, C. & Zeileis, A. (2008), *Applied Econometrics with R*, Springer, New York, ISBN 978-0-387-77316-2
2. Davison, A.C. & Hinkley, D.V. (1997), *Bootstrap Methods and Their Applications*, Cambridge University Press, Cambridge, ISBN 0-521-57391-2

3. Harrell, F.E. Jr (2021). *rms: Regression Modeling Strategies*, R package version 6.2-0, https://CRAN.R-project.org/package=rms

4. Henningsen, A. (2020), *censReg: Censored Regression (Tobit) Models*, R package version 0.5-32, https://CRAN.R-project.org/package=censReg

5. Canty, A. & Ripley B. (2021), *boot: Bootstrap R (S-Plus) Functions*, R package version 1.3-28, https://CRAN.R-project.org/package=boot