Using **icenReg** for interval censored data in **R**

Clifford Anderson-Bergman

December 6, 2015

Contents

Chapter 1

Introduction

1.1 Interval Censoring

Interval censoring occurs when a response is known only up to an interval. A classic example is testing for diseases at a doctor's clinic; if a subject tests negative at t_1 and positive at t_2, all that is known is that the subject acquired the disease in (t_1, t_2), rather than an exact time. Other classic examples include examining test mice for tumors after sacrifice (results in *current status* or *case I* interval censored data, in which all observations are either left or right censored, as opposed to the more general *case II*), customer choice models in economics (customers are presented a price for a product and chose to purchase or not, researcher wants to know distribution of maximum spending amount; this results in current status data again), data reduction methods for sensor analyses (to reduce load on sensor system, message is intentionally surpressed if outcome is in an expected region) and data binning (responses reported only up to an interval, in some cases to keep the subjects anonymous, in some cases to reduce size of data).

Often interval censoring is ignored in analysis. For example, age is usually reported only up to the year, rather than as a continuous variable. In the case that these intervals are relatively short, the bias introduced by ignoring the interval censoring may be small enough to be safely ignored. However, in the case that the width of intervals is non trivial, statistical methods that account for this should be used for reliable analysis.

1.2 Classic Estimators

The topic of interval censoring began emerging in the field of survival analysis. Although it is now considered in other fields of study (such as *tobit regression*), at this time **icenReg** focusses on survival models.

One of the earliest models is the *Non-Parametric Maximum Likelihood Estimator* (NPMLE), also referred to as *Turnbull's Estimator*. This is a generalization of the Kaplan Meier curves (which is a generalization of the empirical distribution function) that allows for interval censoring. Unlike the Kaplan Meier curves, the solution is not in closed form and several algorithms have been proposed for efficient computation. A special topic regarding the NPMLE is the bivariate NPMLE; this is for the special case of two interval censored outcomes, in which the researcher wants a non-parametric estimator of the joint distribution. This is especially computationally intense as the number of parameters can be up to n^2.

Semi-parametric models exist in the literature as well; two classic regression models are the Cox-PH model and the proportional odds model. Unlike the special example of right-censored data, the baseline parameters *must* be estimate concurrently with the regression parameters. The model can be kept semi-parametric (i.e. no need to decide on a parametric baseline distribution) by using the Turnbull estimator, modified to account for the given regression model, as the baseline distribution. The semi-parametric model can be computationally very difficult, as the number of baseline parameters can be quite high (up to n), which must follow shape constraints (i.e. either a set of probability masses or a cumulative hazard function, which must be strictly increasing) and there is no closed form solution to either regression or baseline parameters.

Fully parametric models exist as well and can be calculated using fairly standard algorithms. There are slight complications in that the interval censoring causes the log likelihood function to be non-concave. However, only slight modifications are required to adress this issue. In practice, fully-parametric models should be used with caution; the lack of observed values means that model inspection can be quite difficult; there are no histograms, etc., to be made. As such, even if fully parametric models are to be used for the final analysis, it is strongly encouraged to use semi-parametric models at least for model inspection.

1.3 Models fit with icenReg

The author's motivation for building **icenReg** was to build a set of fast, reliable tools for analysts' to use on applied data. At this time, the following set of models can be fit (name in paratheses is function call in **icenReg**):

- NPMLE (`ic_sp` can be used to fit univariate NPMLE, alternatively `ICNPMLE` can be used to fit univariate or bivariate NPMLE)
- Semi-parametric model (`ic_sp`, with options `model = "ph"` for porportional hazards, `"po"` for proportional odds)
- Fully parametric model (`ic_sp` in additon to `model` option, also have a choice of `dist`, with options `"exponential", "gamma", "weibull", "lnorm", "loglogistic" and "generalgamma"`)

In addition, **icenReg** includes various diagnostic tools. These include

- Plots for diagnosising baseline distribution (`diag_baseline`)
- Plots for diagnosising covariate effects (`diag_covar`)
- Cross validation via multiple imputations (`icenReg_cv`)

1.4 Data Examples in icenReg

The package includes 4 sources of example data: two functions that simulate data and two sample data sets. The simulation functions are `simIC_weib`, which simulates interval censored regression data with a Weibull baseline distribution and `simBVCen`, which simulates bivariate interval censored data. The sample data sets are `miceData`, which contains current status data regarding lung tumors from two groups of mice and `essIncData`, which includes data from the European Social Survey. In this case, wages were only recorded up to an interval to protect the identity of the subjects. The dataset `essIncData_small` is a smaller subset of `essIncData` (n = 500 instead of 6,712), which is used in many of the examples only so that CRAN's testing of the package runs quicker. In practice, using all these models on n = 6,712 is trival to do, rarely taking more than a few seconds even on a slower laptop.

Chapter 2

Fitting Models in icenReg

An important note about **icenReg** is that in all models, it is assumed that the response interval is **closed**, i.e. the event is known to have occurred within $[t_1, t_2]$, compared with $[t_1, t_2)$, (t_1, t_2), etc. This is of no consequence for fully parametric models, but does mean the solutions may differ somewhat in comparison with semi- and non-parametric models that allow differnt configurations of open and closed response intervals.

2.1 Non-parametric models

As noted earlier, for univariate interval censored data, the model may be fit with either `ic_sp` or `ICNPMLE`. For large datasets (i.e. n > 50,000), `ic_sp` will become faster than `ICNPMLE`. In addition, `ic_sp` can readily be provided to the `plot` method. For bivariate data, `ICNPMLE` is currently the only choice.

If the data set is relatively small and the user is interested in non-parametric tests, such as the log-rank statistic, we actually advise using the **interval** package, as this provides several testing functions. However, **icenReg** is several fold faster than **interval**, so if large datasets are used (i.e. n > 1,000), the user may have no choice but to use **icenReg**.

To fit an NPMLE model for interval censored data, we will consider the `miceData` provided in **icenReg**. This dataset contains three variables: `l, u` and `grp`. `l` and `u` represent the left and right side of the interval containing the event time (note: data is current status) and `grp` is a group indicator with two categories.