Using **icenReg** for interval censored data in **R** v1.3.5

Clifford Anderson-Bergman

May 22, 2016

Contents

1 Introduction

This manual is meant to provide an introduction to using **icenReg** to analyze interval censored data. It is written with expectation that the reader is familiar with basic survival analysis methods. Familiarity with the Kaplan Meier curves and Cox proportional hazards model should be sufficient.

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*, which allows for any interval), 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; when a subject reports that their age is 33, the information we have is really that their age is in the interval [33,34). In the case that these intervals are very short relative to the question of interest, such as with reported age when the scientific quesiton is about age of onset of type II diabetes, the bias introduced by ignoring the interval censoring may besmall 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.

Standard notation for interval censoring is that each observation contains a response interval $[l_i, r_i]$ such that the true event time is known to have occurred within. Note that this allows for uncensored observations ($l_i = r_i$), right censored($r_i = \infty$), left censored ($l_i = 0$) or none of the above ($0 < l_i < r_i < \infty$).

In **icenReg**, the response value is allowed to be interval censored. If our data contains the values `L` and `R`, representing the left and right sides of the response interval, we can pass our response to a regression model using either

```
cbind(L, R)
Surv(L, R, type = "interval2")
```

It is worth nothing that other R packages, specifically for non-parametric estimation, allow you to declare whether the response intervals are open, closed or a combination of partially opened, for example $[l_i, r_i)$. In **icenReg**, it is always assumedthat the intervals are closed.

1.2 Classic Estimators

The topic of interval censoring began 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 fit by **icenReg** are the Cox-PH model and the proportional odds model. The well known Cox-PH, or proportional hazards regression model, has the property that

$$h(t|X,\beta) = h_o(t)e^{X^T\beta}$$

where $h(t|X,\beta)$ is the hazard rate conditional on covariates X and regression parameters β, with h_o as the baseline hazard function. This relation is equivalent to

$$S(t|X,\beta) = S_o(t)^{e^{X^T\beta}}$$

where $S(t|X,\beta)$ is the conditional survival and $S_o(t)$ is the baseline survival function.

The less known proportional odds model can be expressed as

$$\text{Odds}(S(t|X,\beta)) = e^{X^T\beta}\text{Odds}(S_o(t))$$

or

$$\frac{S(t|X,\beta)}{1-S(t|X,\beta)} = e^{X^T\beta}\frac{S_o(t)}{1-S_o(t)}$$

Unlike the special example of the Cox PH model with right-censored data, the baseline parameters *must* be estimated 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. One of the contribution the algorithms in **icenReg** make to the field of statistical computing is efficient computation of the non-parametric and semi-parametric estimators, allowing for relatively efficient estimation on standard computers (*i.e.* less than one second) of relatively large samples (n = 10,000 for the semi-paramtric model, n = 100,000 for the non-parametric model), although the semi-parametric models are still significantly slower than fully-parametric models.

Fully parametric models exist as well and can be calculated using fairly standard algorithms. In addition to the proportional hazards and odds models, the accelerated failure time model can be used for parameteric modeling. These models have the following relationship:

$$S(t|X,\beta) = S_o(te^{X^T\beta})$$

For technical reasons not discussed here, this model is very simple to implement for a fully parameteric model, but very difficult for a semi-parametric model. As such, **icenReg** contains tools for a fully-parametric accelerated failure time model, but not a semi-parametric one.

There are slight complications in that the interval censoring can cause the log likelihood function to be non-concave. However, for reasonable sized data, the log likelihood function is usually locally concave near the mode and only slight modifications are required to address 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. **icenReg** fits fully parametric accelerated failure time, proportional odds and proporitonal hazard models for interval censored data.

1.3 Models fit with icenReg

At this time, the following set of models can be fit (name in paratheses is function call in **icenReg**):

- NPMLE (`ic_np`)
- Semi-parametric model (`ic_sp`)
  - `model` for model type
    - `"po"` for proportional odds
    - `"ph"` for proportional hazards
- Fully parametric model (`ic_par`)
  - `model` for model type
    - `"po"` for proportional odds
    - `"ph"` for proportional hazards
    - `"aft"` for accelerated failure time model
  - `dist` for baseline distribution
    - `"exponential"`
    - `"gamma"`
    - `"weibull"`
    - `"lnorm"`
    - `"loglogistic"`
    - `"generalgamma"`

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

- Plots for diagnosising baseline distribution (`diag_baseline`)
- Plots for diagnosising covariate effects (`diag_covar`)

1.4 Data Examples in icenReg

The package includes 3 sources of example data: one function that simulates data and two sample data sets. The simulation function is `simIC_weib`, which simulates interval censored regression data with a Weibull baseline distribution The sample data sets are `miceData`, which contains current status data regarding lung tumors from two groups of mice and `IR_diabetes`, which includes data on time from diabetes to diabetic nephronpathy in which 136 of 731 observations are interval censored due to missed follow up.

2 Fitting Models using 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 different configurations of open and closed response intervals.

2.1 Non-parametric models

The non-parametric maximum likelihood estimator can be using `ic_np`. 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**. In discussions with the author of **interval**, it was indicated that a user could build a wrapper around `ic_np` to be used in **interval**'s logrank tests, but this has not been done yet.

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.

```
> data(miceData)
> head(miceData, 3)

  l   u grp
1 0 381  ce
2 0 477  ce
3 0 485  ce
```

We can fit a non-parametric estimator for each group by

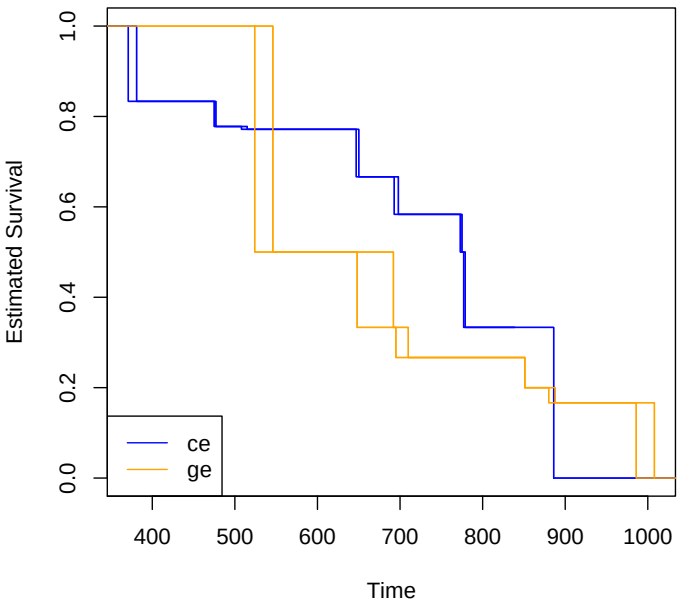
```
> np_fit = ic_np(cbind(l, u) ~ grp, data = miceData)
```

If we wanted only a single fit for both groups, this can be done in two ways. The two fits are equivalent, but are just used to demonstrate differing possible syntax.

```
> groupedFit1 <- ic_np(cbind(l,u) ~ 0, data = miceData)
> groupedFit2 <- ic_np(miceData[,c('l', 'u')])
```

The fits can be plotted as follows:

```
> plot(np_fit, col = c('blue', 'orange'),
+      xlab = 'Time', ylab = 'Estimated Survival')
```



Looking at the plots, we can see a unique feature about the NPMLE for interval censored data. That is, there are *two* lines used to represent the survival curve. This is because with interval censored data, the NPMLE is not always unique (in fact, it usually is not); any curve that lies between the two lines has the same likelihood. For example, any curve that lies between the two blues lines maximizes the likelihoodassociated with `"ge"` group of mice.

Formal statistical tests using the NPMLE are not currently supported by **icenReg**. We recommend using the **interval** package for this.

2.2 Semi-parametric models

Semi-parametric models can be fit with `ic_sp` function. This function follows standard regression syntax. As an example, we will fit the `IR_diabetes` dataset, which contains data on time from diabetes to diabetic nephropathy. In this dataset, we have the left and right sides of the observation interval containing the true response time and the gender of the patient.

```
> data("IR_diabetes")
> head(IR_diabetes, 3)
```

```
  left right gender
1   24    27   male
2   22    22 female
3   37    39   male
```

We fit the model below. Note that this may be time consuming, as the semi-parametric model is somewhat computationally intense and we are taking `bs_samples` bootstrap samples of the estimator.

```
>   fit_ph <- ic_sp(cbind(left, right) ~ gender, model = 'ph',
+                   bs_samples = 100, data = IR_diabetes)
>   fit_po <- ic_sp(cbind(left, right) ~ gender, model = 'po',
+                   bs_samples = 100, data = IR_diabetes)
```

The first model by default fits a Cox-PH model, while the second fits a proportional odds model. We can look at the results using either the `summary` function, or just directly looking at the results (what is displayed is the same).

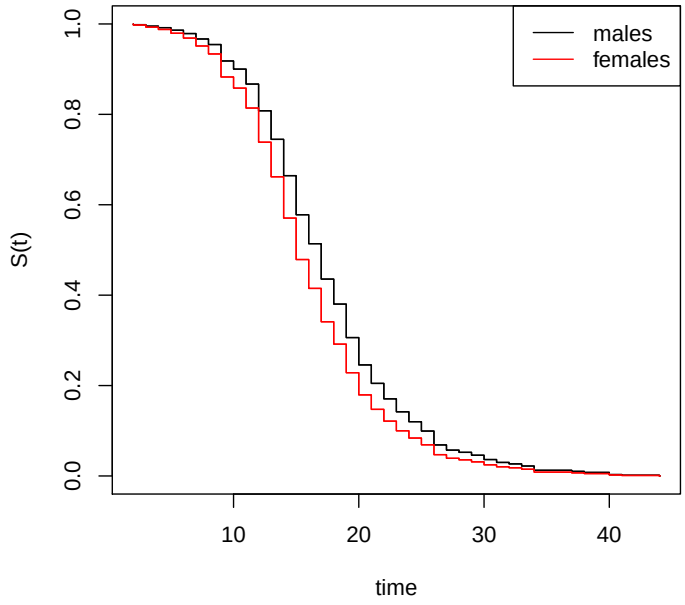
```
>   fit_po

Model:  Proportional Odds
Baseline:  semi-parametric
Call: ic_sp(formula = cbind(left, right) ~ gender, data = IR_diabetes,
    model = "po", bs_samples = 100)

           Estimate Exp(Est) Std.Error z-value        p
gendermale    0.399     1.49    0.1302   3.065 0.002178

final llk =  -1956.969
Iterations =  21
Bootstrap Samples =  100

>   fit_ph

Model:  Cox PH
Baseline:  semi-parametric
Call: ic_sp(formula = cbind(left, right) ~ gender, data = IR_diabetes,
    model = "ph", bs_samples = 100)

           Estimate Exp(Est) Std.Error z-value     p
gendermale  -0.1392     0.87    0.0886  -1.572 0.116

final llk =  -1959.489
Iterations =  38
Bootstrap Samples =  100
```

For the semi-parametric models, bootstrap samples are used for inference on the regression parameters. The reason for this is that as far as we know, the limiting distribution of the baseline distribution is currently not characterized. In fact, to our knowledge, even using the bootstrap error estimates for the baseline distribution is not valid. Because the regression parameters cannot be seperated in the likelihood function, using the negative inverse of the Hessian for the regression standard errors is not generally valid. However, it has been shown that using the bootstrap for inference *on the regression parameters* leads to valid inference.

We can use these fits to create plots as well. The `plot` function will plot the estimated survival curves or CDF for subjects with the set of covariates provided in the `newdata` argument. If `newdata` is left equal to `NULL`, the baseline survival function will be plotted.

Below is a demonstration of how to plot the semi-parametric fit for males and females.

```
>   newdata <- data.frame(gender = c('male', 'female') )
>   rownames(newdata) <- c('males', 'females')
>   plot(fit_po, newdata)
```



2.3 Parametric Models

We can fit parametric models in **icenReg** using the `ic_par` function. The syntax is essentially the same as above, except that the user needs to specify `dist`, the parametric family that the baseline distribution belongs to. The current choices are `"exponential"`, `"weibull"` (default), `"gamma"`, `"lnorm"`, `"loglogistic"` and `"generalgamma"` (generalized gamma distribution). The user must also select `model = "ph"`, `"po"`, or `"aft"` as the model type.

It is not necessary to specify `bs_samples` for parametric models, as inference is done using the asymptotic normality of the estimators. Fitting a parametric model is typically faster than the semi-parametric model, even if no bootstrap samples are taken for the semi-parametric model. This is because the fully-parametric model is of lower dimensional space without constraints.

Suppose we wanted to fit a proportional odds model to the `IR_diabets` data with a generalized gamma distribution. This could be fit by

```
> fit_po_gamma <- ic_par(cbind(left, right) ~ gender,
+     data = IR_diabetes, model = "po", dist = "gamma")
```

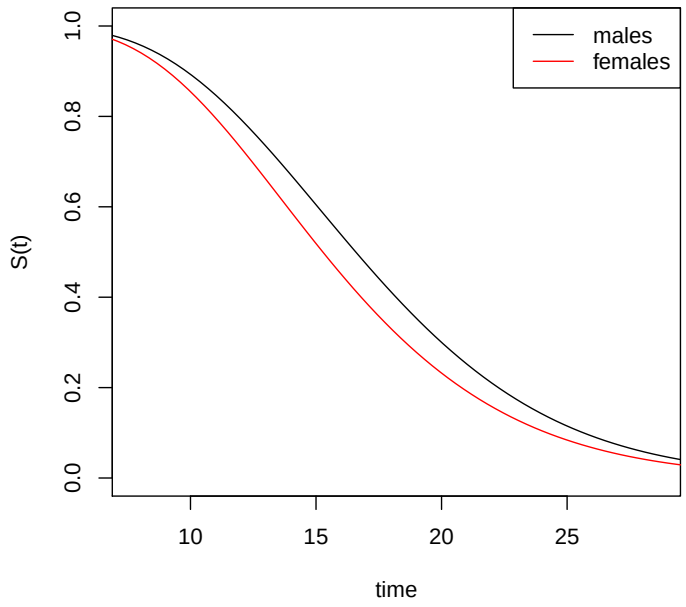
We can examine the regression coefficients in the same way as with the semi-parametric model.

```
> fit_po_gamma
```

```
Model:  Proportional Odds
Baseline:  gamma
Call: ic_par(formula = cbind(left, right) ~ gender, data = IR_diabetes,
    model = "po", dist = "gamma")

           Estimate Exp(Est) Std.Error z-value        p
log_shape    1.9980    7.377   0.05447   36.69 0.000000
log_scale    0.8248    2.281   0.05560   14.83 0.000000
gendermale   0.3496    1.419   0.13550    2.58 0.009876

final llk =  -2006.619
Iterations =  4
```

We can also examine the survival/cdf plots in the same way.

```
> plot(fit_po_gamma, newdata, lgdLocation = "topright")
```



3 Inspecting model fit

3.1 Examining Baseline Distribution

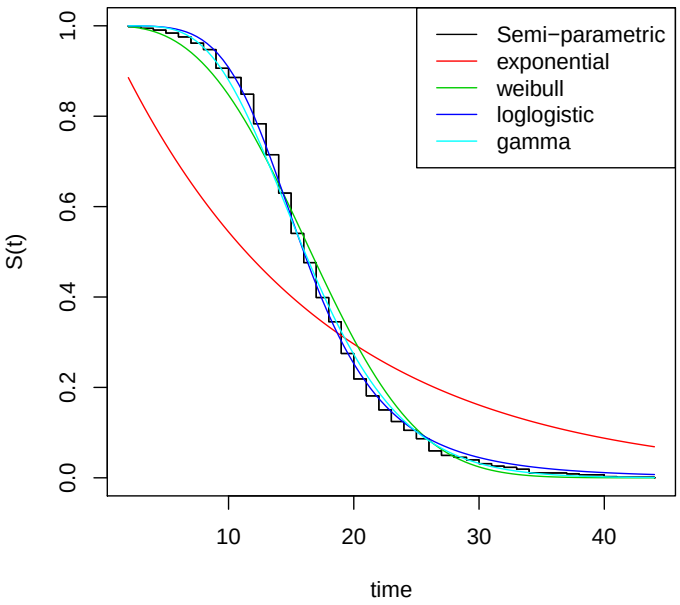
Althought the semi-parametric model is more flexible, and thus more robust to unusual baseline distributions, there are many reasons one may decide to use a parametric model instead. One reasons is that, as stated earlier, we are not aware of any general distributional theory regarding the baseline distribution, outside of the univariate case with case I interval censored data. Even in this

case, the estimator is highly inefficient, observing convergence rates of $n^{1/3}$ instead of the more standard $n^{1/2}$. Because of this, making inference about values that directly require the baseline distribution, such as creating a confindence interval for the median for subjects with a given set of covariates, cannot be done with the semi-parametric model.

However, even if a parametric model is used for final inference, the semi-parametric model is still useful for assessing model fit. This is especially important for interval censored data, as we do not have the option of examining typical residuals or histograms as we would if the outcome was uncensored. **icenReg** has the function `diag_baseline` that plots several choices of parametric baseline distributions against the semi-parametric estimate. If the parametric distribution shows no systematic deviations from the semi-parametric fit, this implies the choice of parametric family may do a reason job of describing the underlying distribution. If there are clear deviations, this model should not be trusted.

To use `diag_baseline`, you must provide either a fitted model, or a formula, data and model. You then select the parametric families that you would like plotted against the non-parametric estimate (default is to fit all available). As an example, suppose we wanted to examine the different parametric fits for the `IR_diabetes` dataset. This could be done with

```
> diag_baseline(cbind(left, right) ~ gender,
+     model = "po",
+     data = IR_diabetes,
+     dists = c("exponential", "weibull",
+               "loglogistic", "gamma"),
+     lgdLocation = "topright")
```



Alternatively, using the fits from earlier, we can just call

```
> diag_baseline(fit_po, lgdLocation = "topright",
+               dists = c("exponential", "weibull",
+                         "loglogistic", "gamma")
+               )
```

Visual diagnostics are always subjective, but in this case we definitively know that the exponential fit is not appropriate and we believe the gamma baseline is most appropriate for the proporitonal odds model (although there is not overwhelming evidence that it is best).

3.2 Examining Covariate Effect

Although semi-parametric models do not make assumptions about the parametric family of the baseline distribution, both fully-parametric and semi-parametric models make assumptions about the form of the covariate effect, akin to the link function in generalized linear models.

A rule of thumb for identifying gross violations of proportional hazards is to check if the Kaplan Meier curves cross; if they do, and this cross appears not purely by chance, the proportional hazards assumption seems inappropriate.

This can naturally extend to the case of interval censored data by replacing the Kaplan Meier curves with the NPMLE. Also, this informal test can be generalized to the proportional odds model; the proportional odds assumption also implies that survival curves that differ only by a constant factor of the odds of survival should not cross.

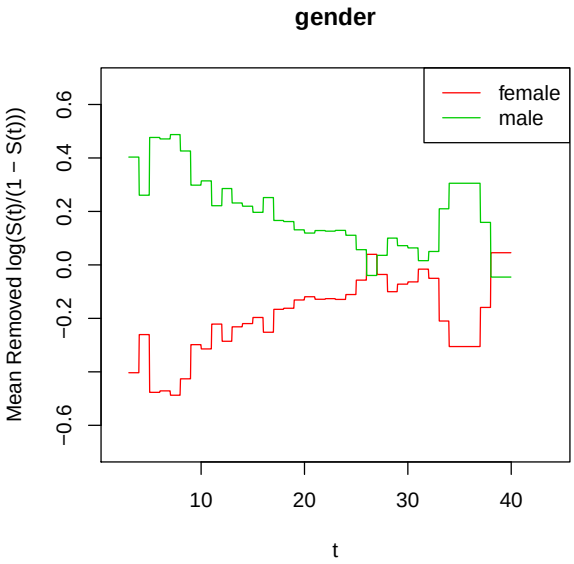
Another method of assessing involves transforming yoursurvival estimates such that if the assumptions are met, the difference in transformed survival will be constant. For the proportional hazards model, this is the complementary log-log tranformation (i.e. $\log(-\log(s))$). For the proportional odds model, this is the logit transformation (i.e. $\log(s/(1-s))$).

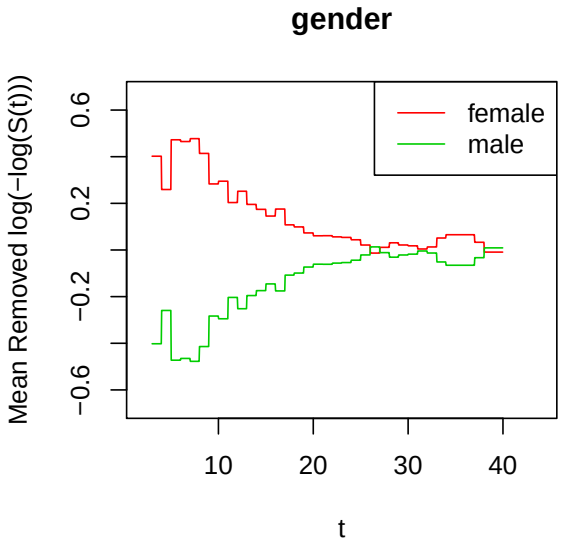
Plotting these functions can be done automatically in **icenReg** using the `diag_covar` function. The basic flow is that function takes in the fit, divides the data up on a covariate of interest. If it is categorical, it simply breaks up by category, if it is numeric, it attempts to find break point to evenly split up the data. Then, for each subset of the data, it fits the corresponding semi-parametric model and plots the transformation of the baseline distribution.

To demonstrate, suppose we wanted to assess whether the Cox-PH or proportional odds model was more appropriate for the `IR_diabetes`. This could be done by

```
> diag_covar(fit_po, lgdLocation = "topright",
+            main = "Checking Proportional Odds")
> diag_covar(fit_ph, lgdLocation = "topright",
+            main = "Checking Proportional Hazards")
```

We see that especially for gender, the porportional odds seems somewhat more appropriate (the difference between transformed values seems more constant). This agrees with the fact that the likelihood is approximately 2.5 greater for the proportional odds model than Cox-PH.





4 Appendix

4.1 Parallel Bootstrapping

Bootstrapping can be very computationally intensive. Fortunately, it is also embarrassingly parallel. As such, icenReg is written to work seamlessly with `doParallel`

```
> library(doParallel)
> myCluster <- makeCluster(4)  #uses 4 cores
```

```
> registerDoParallel(myCluster)
> fit <- ic_sp(cbind(left, right) ~ gender,
+            data = IR_diabetes, model = "po",
+            bs_samples = 50, useMCores = TRUE)
>stopCluster(myCluster)
```