What's additional and/or new since the first printing?

Tableman

November 7, 2004

A. *New functions*

- `g.emphaz`: This function replaces the function `emphazplot` intoduced on page 45. It prints the empirical hazards values similar to the output at bottom of page 45. It draws the empirical hazards plots, only nicer than those in Figure 2.5 on page 46. The required arguments are:

  `data`: a `Surv` object or a list of `Surv` objects
  `type`: what should be drawn? "ht" for hitilde or "hhat" for hihat

  Example: The AML data

```
Surv0<-Surv(aml$weeks[aml$group==0],aml$status[aml$group==0])
Surv1<-Surv(aml$weeks[aml$group==1],aml$status[aml$group==1])
data<-list(Surv1,Surv0)
g.emphaz(data=data,type="ht",main="hitilde",
                    legend=c("maintained","nonmaintained"))
g.emphaz(data=data,type="hhat",main="hihat",
                    legend=c("maintained","nonmaintained"))
```

- `extcox.twochange`: Extends the `extcox.1Et`, page 193, to incorporate two change points. That is, it determines three intervals over which we hope the PH assumption is satisfied.

- `optimal.change.point`: See the description in B. *Additional material* below.

- `qq.reg.resid`: For parametric regression models, this constructs a Q-Q plot of ordered residuals $e_i = (y_i - \hat{y}_i)/\hat{\sigma}$ against the log-parametric standard quantiles z_i of either the Weibull, log-normal, or log-logistic distribution. See Errata Sheet, **item p. 147**, for a detailed description and example.

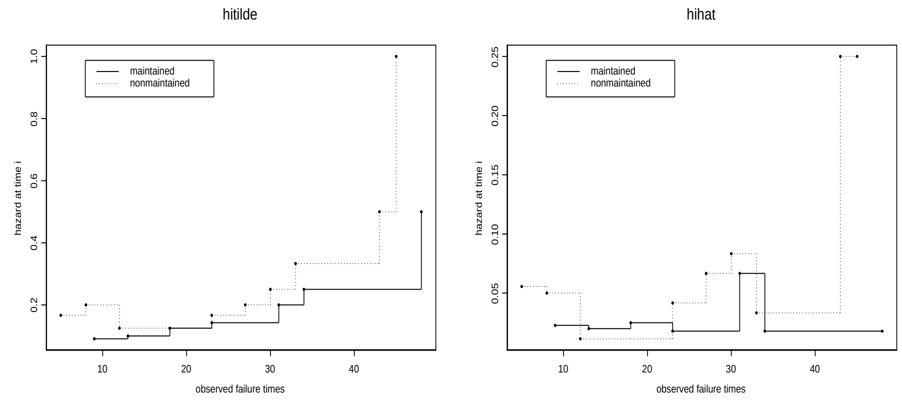

Figure 2.5: *A comparison of empirical hazards. Left plot displays* $\tilde{h}(t_i)$. *Right plot displays* $\hat{h}(t)$.

B. *Additional material*

7.1 Extended Cox Model

This is a continuation of **Part IV: An extended Cox model analysis**, which begins on page 192. Kleinbaum visually chooses one year (365 days) to be the change point as this is where the two survivor curves appear to begin to diverge. One can also employ the *profile log-likelihood approach* to determine the optimal change point. This approach was introduced in Chapter 6.3.8, where we used the criterion of maximizing the profile log-likelihood to determine the cut point. The function `optimal.change.point` computes the profile log-likelihoods for values of t_0 ranging over the default quantiles, `seq(.1,.9,.01)`, of the uncensored survival times. Figure 1 displays their graph.

In order to use the function `optimal.change.point` pick any time point within the scope of your data to start. We pick 100 days.

Caution: Be sure the exposure variable is in column 2, the status variable is in column 3, and the time variable is in column 4 of your data frame.

```
> attach(ADDICTS)
> out <- extcox.1Et(ADDICTS,100) # Puts in Andersen-Gill counting
                                 # process form.
> temp.ext <- coxph(Surv(Start,Stop,Status)~Prison+Dose+ET1+ET2,
               data=out)   # temp.ext is the coxph object that
  # gives the necessary formula within the function
  # optimal.change.point.
> best <- optimal.change.point(data=ADDICTS,time=Days.survival,
             status=Status,object=temp.ext)
> cbind(best$t0+.00001,best$loglik) # Prints out the values.
```

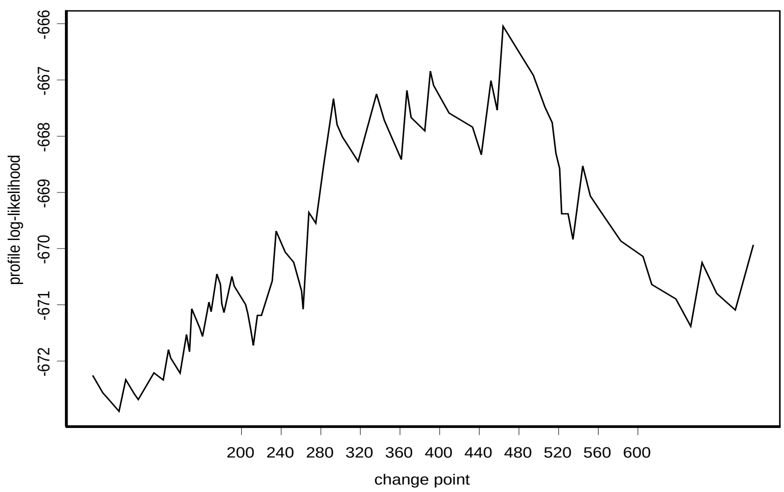

Figure 1: *Profile log-likelihoods for the change point t_0.*

```
> plot(best$t0+.0001,best$loglik,type="l",xlab="change point",
   ylab="profile log-likelihood",lwd=2)   # Figure 1
> out <- extcox.1Et(ADDICTS,464) # Optimal change point is 464
                                 # days.
> fit4 <- coxph(Surv(Start,Stop,Status) ~Prison+Dose+ET1+ET2,
                 data=out)
```

Some selected output follows:

```
> best <- optimal.change.point(data=ADDICTS,time=Days.survival,
               status=Status, object = temp.ext)
    69%
 464.05  # The optimal change point in days

> fit4
Call: coxph(formula=Surv(Start,Stop,Status)~Prison+Dose+ET1+ET2,
              data=out)

          coef exp(coef) se(coef)     z        p
Prison  0.3890     1.476  0.16859  2.31 2.1e-002
  Dose -0.0354     0.965  0.00645 -5.48 4.3e-008
   ET1  0.4887     1.630  0.23396  2.09 3.7e-002
   ET2  2.3970    10.990  0.52996  4.52 6.1e-006

Likelihood ratio test=79  on 4 df, p=3.33e-016  n= 337
```

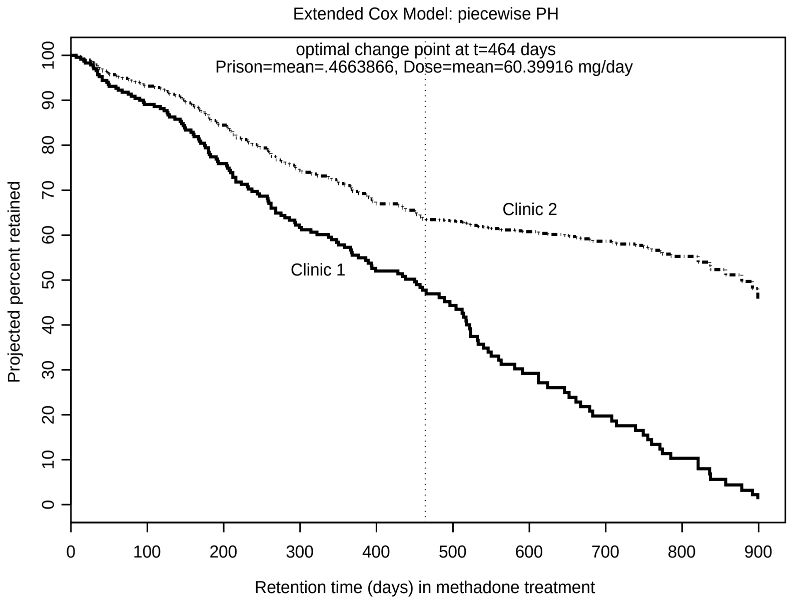

Figure 2: *K-M curves adjusted for Prison and Dose effects in the extended Cox model.*

The following S code provides a plot of the projected survival probabilities, which here is the projected percent retention in each clinic. The output has been modified. Figure 2 displays the plot.

```
> fit.1 <- survfit(fit4, data.frame(Start = c(0,464),
       Stop = c(464,1100), Status = c(0,1), ET1 = c(1,0),
       ET2 = c(0,1),Prison =c(0.4663866,0.4663866),Dose =
         c(60.39916, 60.39916)), individual=T)
> fit.2 <- survfit(fit4, data.frame(Start = c(0,464),
       Stop = c(464,1100), Status = c(0,1), ET1 = c(0,0),
       ET2 = c(0,0), Prison = c(0.4663866,0.4663866), Dose =
       c(60.39916,60.39916)), individual=T)

> fit.1
   n events mean se(mean) median 0.95LCL 0.95UCL
 236    150  434       16    450     358     518
> fit.2
   n events mean se(mean) median 0.95LCL 0.95UCL
 236    150  632     31.9    878     612      NA
```

```
> plot(fit.1,type="l",lty=1,lwd=3,lab=c(10,10,7),
    xlab="Retention time (days) in methadone treatment",ylab=
    "Projected percent retained", yscale=100, conf.int=F)
> lines(fit.2,lty=3,lwd=3)
> abline(v=464,lty=2,lwd=2)
> mtext("Extended Cox Model: piecewise PH",3,line=1)
> mtext("optimal change point at t=464 days",3,line=-1)
> mtext("Prison mean=.4663866,Dose mean=60.39916 mg/day",3,
     line=-2)                # Figure 2
```

Results:

- The difference here is that now the clinic effect is significant over both intervals of time. The $\widehat{\mathrm{HR}} = 1.63$ with p-value $= 0.037$ for the effect of clinic when time $t < 464$ days. For $t \geq 464$, $\widehat{\mathrm{HR}} = 10.99$ with p-value $= 6.1 \times 10^{-6}$. Clinic 2 is always doing significantly better in retention of patients than Clinic 1.

- Within the first 464 days, Clinic 2 is 1.63 times more likely to retain patients longer than Clinic 1. After 464 days, Clinic 2 is nearly 11 times more likely to retain patients longer than Clinic 1. Equivalently, Clinic 2 has $\frac{1}{11} \approx 9\%$ the risk of Clinic 1 of patients leaving its methadone treatment program.

- The risks, the rates at which patients leave the two clinics' treatment programs, are visually represented in Figure 2 by the slopes of the survivor curves at any time point. The slope of the Clinic 1 curve appears constant, whereas the slope of the Clinic 2 curve significantly slows after $t_0 = 464$.

7.2 Competing risks: cumulative incidence estimator

The following example was cleverly formulated by Peter Sparks, a former student in our master's program. To the best of our knowledge, Peter's competing risks analysis of the Case K employment data is novel.

Case K employment data example:

The dataset `CaseK` chosen to illustrate a competing risk analysis is in the datasets archive `statLib` located at `http://lib.stat.cmu.edu/datasets` under "`employment`". It was originally used by Kadane and Woodworth (2004) in their paper "Hierarchical Models for Employment Decisions" to investigate a claim of age biased firing (terminated involuntarily) by a company we shall refer to as company K. Individuals 40 years or older are federally protected against age discrimination in employment decisions concerning hiring, firing, and promotion. The methods Kadane and Woodworth used are not discussed in this example. Their conclusion, however, was that the data supported the claim.

For a sample of 416 company K employees followed over time, birth dates, hire dates, end of employment dates, and termination indicators were recorded. The dates were of the form MM/DD/YYYY. The table below is a partial list of the original data. The variables are defined as follows:

mob = month of birth
dob = day of birth
yob = year of birth
moh = month of hire
doh = day of hire
yoh = year of hire
mox = end of employment month
(= 99 if still employed at the end of the study)
dox = end of employment day
(= 99 if still employed at the end of the study)
yox = end of employment year
(= 1999 if still employed at the end of the study)
t = 1 if involuntary termination; 0 if not

obs	**mob**	**dob**	**yob**	**moh**	**doh**	**yoh**	**mox**	**dox**	**yox**	**t**
1	11	24	1972	2	11	1991	99	99	1999	0
2	3	22	1955	3	4	1985	99	99	1999	0
3	11	13	1941	2	4	1991	10	2	1992	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
15	4	16	1930	12	28	1990	1	24	1992	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Competing risks with right censored data formulation

The failure time of interest is "*time from hired to fired*". We use "*terminated*" as a euphemism for "*fired*". Failures from a competing risk, such as quit or retired, are referred to as "*other*". Censored individuals (those still with the company at the end of the study period) had for their end of employment dates 99/99/1999. For example, the employees corresponding to observations 1 and 2 are censored. Four employees' birth dates are missing. The following newly created variables are stored in the data frame `CaseK`:

CaseK data frame variables
`ftime`(in days) = hire time minus end of employment time (or end of study if censored)
`fstatus` = 0 if censored, 1 if fired, 2 if other (quit, retired, died, etc.)
`age` = age (in years) at end of employment or end of study if censored
`age40` = 0 if age less than 40, 1 if age greater or equal to 40
`f1status` = 0 if not fired, 1 if fired

As the actual date of end of study was not available at the time of this writing, the last uncensored end of employment date, 01/27/1995, was used in its place. Thus, the original data are transformed into a set of variables which fit into the framework of competing risks with right censored data.

`cmprsk` **Library**

The `cmprsk` library, downloadable from `biowww.dfci.harvard.edu/~gray/`, contains a number of S functions for use in analysis of competing risks data. Below is a brief description of functions in the library. Recall the **cumulative incidence (CI) function** defined in expression (7.6) is a **subdistribution function** since it increases to $P(T_1 < T_2)$, a quantity less than 1.

- `cuminc()` computes the CI estimator (7.7) and its variance estimates, and performs a nonparametric test for equality of subdistributions across groups.
- `crr()` fits the **proportional subdistribution hazards regression model** described in Fine and Gray (1999). The residuals returned are analogous to the *scaled Schoenfeld residuals* (page 164) in ordinary survival models.
- The functions `print.cuminc()`, `plot.cuminc()`, and `timepoints()` are titled descriptively and illustrated with examples.

S code and analysis

```
> library(cmprsk)
> xx <- cuminc(CaseK$ftime,CaseK$fstatus)
> xx # Estimates and Variances:
 $est:
         2000    4000    6000    8000    10000     12000    14000
 1 1   0.1944  0.2444  0.2747  0.3121   0.3578    0.4008   0.4331
 1 2   0.2572  0.2946  0.3102  0.3479   0.3731    0.3731   0.3731
```

```
$var:
        2000     4000      6000      8000     10000     12000
1 1  0.00045  0.00066   0.00082   0.00105   0.00155   0.00216
1 2  0.00053  0.00067   0.00076   0.00101   0.00123   0.00123

       14000
1 1  0.00282
1 2  0.00123
> plot.cuminc(xx,main="Cumulative Incidence for Termination and
   Other",curvlab=c("Termination","Other"),xlab="Days employed",
     lty=1:2)           # Figure 1
```

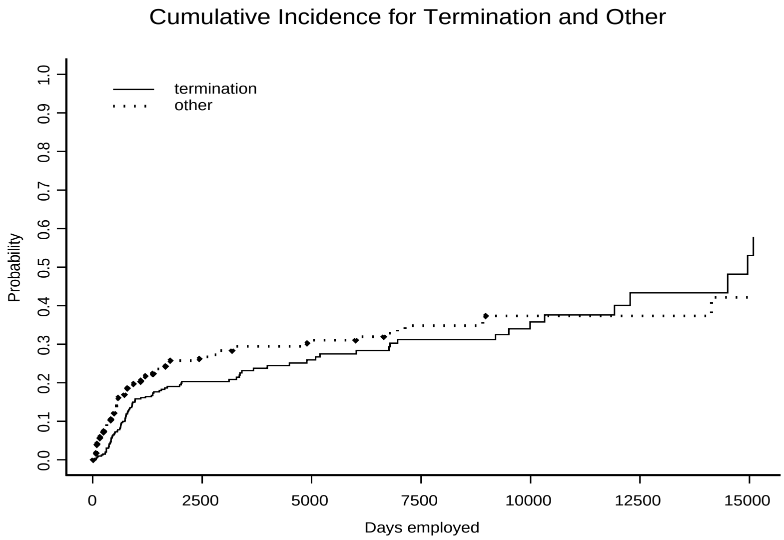

Figure 1: *Estimated cumulative incidence curves for the two competing risks "termination" and "other".*

In Figure 1 we observe the curve for "*other*" lies above the one for "*termination*" until about 11000 days (about 30 years). Then the curves cross. This suggests the presence of an age based discrimination in firing practices of company K.

We can obtain estimates of CI along with variance estimates at survival times of our choice using the `timepoints` function. For example,

```
> timepoints(xx,c(1826,3625,7304,10950,14600))
   # CI evaluated at 5, 10, 20, 30,and 40 years
$est:
        1826     3625      7304     10950     14600
1 1   0.1902   0.2314    0.3121    0.3757    0.4816
1 2   0.2572   0.2946    0.3479    0.3731    0.4215
```

```
$var:
        1826      3625        7304       10950       14600
1 1  0.00043   0.00059     0.00105     0.00175     0.00455
1 2  0.00053   0.00067     0.00101     0.00123     0.00349
```

We now illustrate the error introduced when we treat failures from a competing risk as censored observations. The function `plot.cuminc.f1` is a modification of `plot.cuminc` that only plots the curve for failure of type 1.

```
> ww <- cuminc(CaseK$ftime, CaseK$fstatus)
> xx <- cuminc(CaseK$ftime, CaseK$f1status)
> plot.cuminc.f1(xx,main="CI for Termination: Other as a Com.
  Risk and Other as Censored",curvlab=c(""),xlab="Days employed",
     lty=2)
> lines(ww$"1 1"$time,ww$"1 1"$est, type="s",lty=1)
> legend(0,.9,c("other treated as censored","other as com.risk"),
     lty=2:1)           # Figure 2
```

As expected, when we treat a competing risk failure as censored, we overestimate cumulative incidence of the failure type of interest. This is clearly observed in Figure 2.

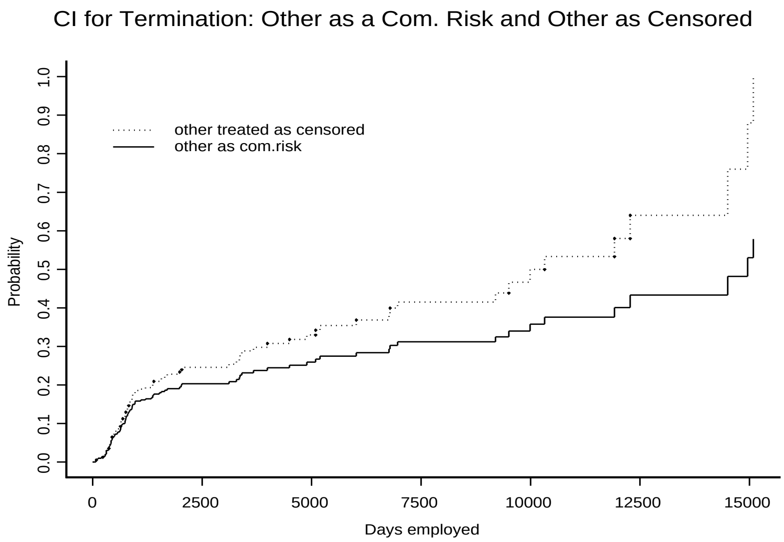

Figure 2: *Estimated cumulative incidence curve for "termination" is the solid line. The dotted line represents the 1-KM curve for "termination" since the competing risk failures are treated as censored.*

Stratifying on age40 to test for age biased firing

We now run `cuminc` while stratifying on the variable age40. When stratifying on levels of a group, `cuminc` conducts tests comparing the subdistrbution functions across groups for each failure type. For our example this means that we test the alternative hypotheses: 1) the "*termination*" subdistributions for the younger and older groups are not equal and 2) the "*other*" subdistributions for the younger and older groups are not equal. The test statistics are described in Gray (1988). `cuminc` also gives estimates of CI at certain times in the range of failure times and corresponding variance estimates for each combination of failure type and group. As `print.cuminc` (see below) also reports these estimates, we omit them under `cuminc` and include them under `print.cuminc`.

```
> CaseK.bday <- na.exclude(CaseK)
       # omits subjects with missing birthdates
> ci.to <- cuminc(CaseK.bday$ftime,CaseK.bday$fstatus,
              group=CaseK.bday$age40,na.action=na.exclude)
> ci.to
 Tests:
       stat       pv  df
 1 12.10549  0.0005   1
 2 12.12225  0.0005   1
```

The first p-value indicates there is a significant difference between the "*termination*" sudistribution for those 40 or older and the subdistribution for those younger than 40. The function `print.cuminc` yields much of the same information as the output of `cuminc`. The number of estimates is a function of `ntp` (number of time points).

```
> print.cuminc(ci.to,ntp=3)
 Tests:
       stat       pv  df
 1 12.10549  0.0005   1
 2 12.12225  0.0005   1

 Estimates and Variances: $est:
        5000     10000       15000
 0 1  0.1604        NA          NA
 1 1  0.3245    0.4268      0.6030
 0 2  0.4118        NA          NA
 1 2  0.2249    0.2979      0.3474

 $var:
        5000     10000       15000
 0 1  0.0023        NA          NA
 1 1  0.0012    0.0019      0.0058
 0 2  0.0029        NA          NA
 1 2  0.0009    0.0015      0.0038
```

The following command plots the CI for each combination of `age40` and failure type. These curves are displayed in Figure 3.

```
> plot.cuminc(ci.to,main="CI for the Four Combinations of Group
   and Failure",curvlab=c("age40=0, terminated", "age40=1,
    terminated","age40=0,other","age40=1,other"),
     xlab="Days employed")        # Figure 3
```

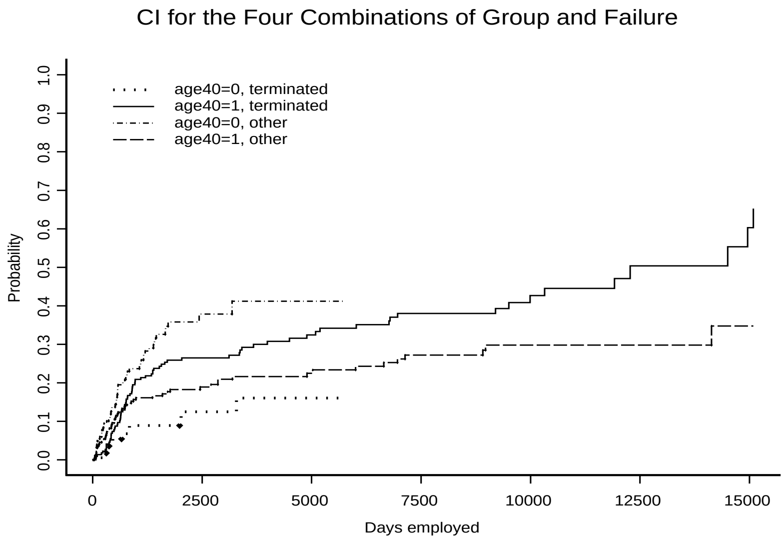

Figure 3: *Estimated cumulative incidence curves for "termination". "Other" is the competing risk.*

The curve for cumulative incidence of "*termination*" for the `age40` = 1 group lies entirely above the one for `age40` = 0. Thus, older individuals at hire experience more chances to be fired. This supports the claim of age based discrimination in termination practices of company K. Also, within older individuals at hire, there is a greater incidence of "*fires*" than the "*others*". On the other hand, within younger individuals at hire, there is a greater incidence of "*others*" than "*fires*" perhaps because younger individuals move more often due to job opportunities, kids' education, etc.

Regression analysis

Gray's `cmprsk` library also includes functions to fit a **proportional subdistribution hazards regression model**, compute and store *scaled Schoenfeld type residuals* for such a model, compute CI estimates and estimator variance estimates, plot the CI estimates, and conduct statistical tests for this model. The function `crr` fits the data to a proportional subdistribution hazards regression model. `crr` returns estimated coefficients along with their standard errors (se) so one can compute point and confidence interval estimates of the sudistribution hazards ratio (SDHR). The default computes the subdistribution hazard function for the type 1 failure "*termination*".

```
> CaseK.reg <- crr(CaseK$ftime,CaseK$fstatus,CaseK$age40)
   # The default computes results for the type 1 failure.
4 cases omitted due to missing values
> CaseK.reg
 convergence:  TRUE
```

```
coefficients:
[1] 1.097
standard errors:
[1] 0.2677
two-sided p-values:
[1] 0.000042
```

For this model the coefficient of age40 is significantly different from zero with p-value = 0.000042. It is significantly greater than zero as the null reference distribution is approximately normal so that the p-value for the one-sided test is 0.000021. The estimated SDHR is exp(coef) = exp(1.097) = 3.00. This value means that employees 40 or older have an estimated 3.00 times the risk or hazard of being terminated as those younger than 40 at any time during their period of employment. The general form of a 95% confidence interval for the SDHR is $\exp(\text{coef} \pm 1.96 \times \text{se(coef)})$. Then from the S output we have $\exp(1.097 \pm .2677)$ which yields a 95% confidence interval estimate of $[1.77, 5.06]$.

The functions `predict.crr` and `plot.predict.crr` are now illustrated in the following code.

```
> z <- predict.crr(CaseK.reg,c(0,1))
    # Computes predictions of CI at levels of age40
> plot.predict.crr(z,main="Regression Curves for Termination:
   age40=0,age40=1",xlab="Days employed",ylab="Probability")
> legend(0,.6,legend=c("age40=0","age40=1"),lty=2:1,bty="n")
                        # Figure 4
```

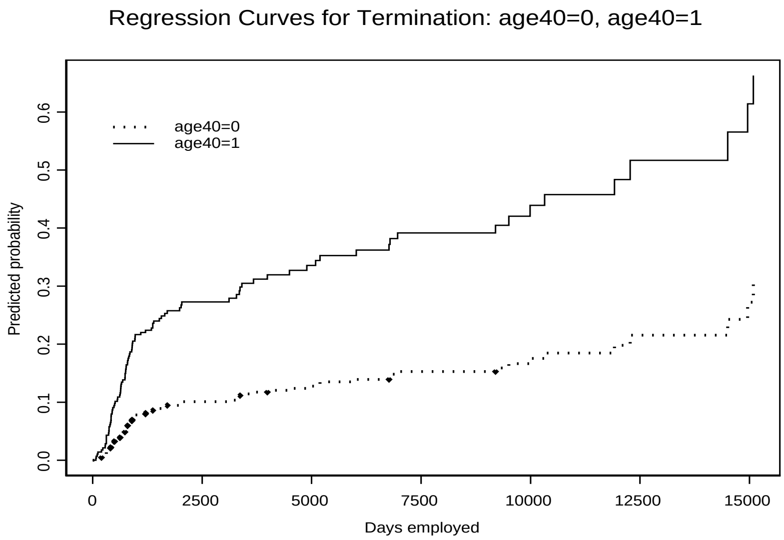

Figure 4: *Predicted cumulative incidence curves for "termination". "Other" is the competing risk.*

In Figure 4, the CI curve for those 40 or older lies entirely above that of the younger than 40 group, which again supports the claim of age based discrimination in firing practices. The curves exhibit the same pattern as that observed between the two "terminated" curves in Figure 3.

We now plot the *scaled Schoenfeld type residuals* versus the unique failure times.

```
> scatter.smooth(CaseK.reg$uftime,CaseK.reg$res,type="p",
   main="Residuals for age40 vs. Unique Failure Times to
    Assess PH Fit", xlab="Days employed",ylab="Residual")
                        # Figure 5
```

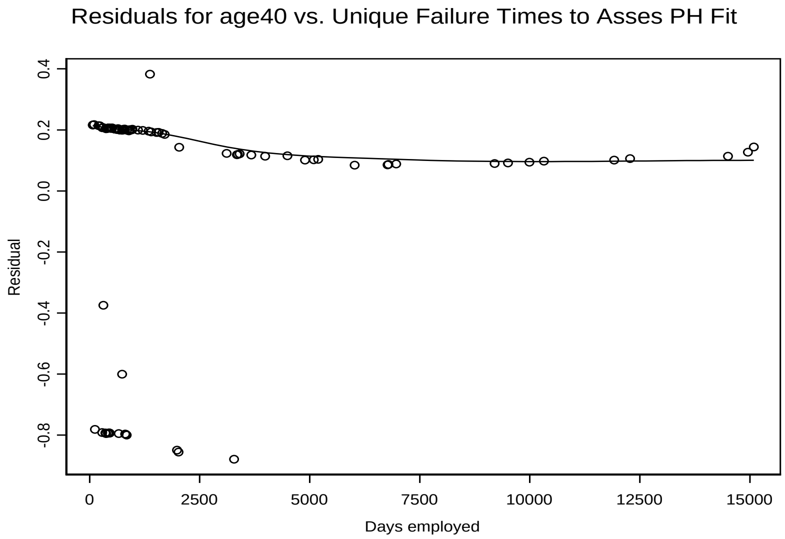

Figure 5: *Scaled Schoenfeld type residuals to assess the fitted subdistribution hazards regression model for "termination" with respect to the PH assumption. A spline smoother is used.*

Fine and Gray (1999) write "The residuals should locally have mean 0 across time, and patterns other than a constant local average indicate lack of fit." The plot in Figure 5 indicates the proportional subdistribution hazards model adequately fits the data.

Gray's `crr` function also allows for time dependent hazard ratios. Another model for this data could then, for example, be one which is piecewise PH. We let the reader investigate this and other models.

References

Fine, J. P. and Gray, R.J. (1999). A proportional hazards model for the subdistribution of a competing risk. *J. Amer. Statist. Assoc.*, **94**, 496−509.

Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Statist.*, **16**, 1141−1154.

Gray, R.J. (2002). `cmprsk` library, competing risks library for S-PLUS. `http://biowww.dfci.harvard.edu/ gray/`.

Kadane, J. and Woodworth, G. (2001). `employment` in `http://lib.stat.cmu.edu/datasets.`

Kadane, J. and Woodworth, G. (2004). Hierarchical models for employment decisions. *J. Business and Economic Statist.*, **22**.

D. *Coming attractions*

1. An example of crossing survival curves: data from a colon cancer study