eAppendix 1: Implementation guide of the additive hazards model in R

Currently, the Additive hazards model is only implemented in the software package R. Therefore the benefits described in this paper require a basic knowledge of R. In this appendix we detail the required R commands from data loading to final estimates (as presented in Table 2 of the main text).

**Step 1: Data loading.**
We assume that data is saved as comma separated values in a plain text file (called SICdata.csv) with column headings in the first row. The data is loaded into the R variable SIC by the command:

```
SIC <- read.csv("c:/.../SICdata.csv")
```

Note that R is case sensitive. To verify that data has been loaded correctly the first six rows of data can be printed by the command:

```
head(SIC)
```

**Step 2: Estimating the Additive hazards model**
The functions required for estimating the Additive hazards model are contained in the package timereg. Before using the package for the first time it must be installed using the menu Packages->Install packages.... Once installed the package can be loaded by the command:

```
library(timereg)
```

The model can now be fitted using the `aalen` function. The name “aalen” refers to a specific version of the additive hazards model (namely the one where all effects are assumed age-dependent) introduced by Aalen,\(^1\) but the R function includes the whole class of Additive hazards models. The output from the function (i.e. the model fit) must be saved to an R variable from which for instance parameter estimates can be extracted. Therefore we fit the model using this command

```
fitAalen <- aalen(Surv(inAge, outAge, event) ~
                  const(factor(educ))*const(factor(smoker)) + const(cohort) + factor(sex),
                  data=SIC, start.time=50)
```

where the first argument to the `Surv`-function is age when entering the study, the second argument is event time, and the third argument is an indicator for event (=1) or censoring (=0). The explanatory variables are listed to the right of the “~” symbol. The wrapper `const` in front of right hand side variables instructs R to estimate an age-invariant effect of that variable, which is analogues to assuming a constant hazards ratio in a Cox model. The wrapper `factor` instructs R that the variable is categorical (it is not needed for cohort since this is a string variable and therefore categorical by nature). The argument `start.time=50` indicates that only person-time after the age of 50 should be included in the analysis.
For further details about estimating and interpreting Additive hazards models with age varying effects see the book by Martinussen and Scheike.²

Next parameter estimates can be extracted by the command

\[
\text{summary(fitAalen)}
\]

which produces the following output

Additive Aalen Model
Test for nonparametric terms

Test for non-significant effects
Supremum-test of significance p-value H₀: B(t)=0

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>SE</th>
<th>Robust SE</th>
<th>z</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>9.78</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>factor(sex)1</td>
<td>4.17</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test for time invariant effects
Kolmogorov-Smirnov test p-value H₀:constant effect

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>SE</th>
<th>Robust SE</th>
<th>z</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.0263</td>
<td>0.001</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>factor(sex)1</td>
<td>0.0308</td>
<td>0.033</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cramer von Mises test p-value H₀:constant effect

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>SE</th>
<th>Robust SE</th>
<th>z</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.0113</td>
<td>0.000</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>factor(sex)1</td>
<td>0.0146</td>
<td>0.009</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parametric terms:

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>SE</th>
<th>Robust SE</th>
<th>z</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>const(factor(educ))1</td>
<td>-4.79e-05</td>
<td>7.25e-05</td>
<td>7.31e-05</td>
<td>-0.6550</td>
<td>5.12e-01</td>
</tr>
<tr>
<td>const(factor(smoker))1</td>
<td>2.42e-03</td>
<td>1.15e-04</td>
<td>1.14e-04</td>
<td>21.3000</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>const(factor(Koh_us))45 år, 1936-kohorten</td>
<td>-1.71e-04</td>
<td>2.72e-04</td>
<td>2.82e-04</td>
<td>-0.6050</td>
<td>5.45e-01</td>
</tr>
<tr>
<td>const(factor(Koh_us))Inter99 baseline</td>
<td>2.60e-04</td>
<td>1.83e-04</td>
<td>1.89e-04</td>
<td>1.3800</td>
<td>1.69e-01</td>
</tr>
<tr>
<td>const(factor(Koh_us))Monica1</td>
<td>-9.73e-05</td>
<td>2.05e-04</td>
<td>2.01e-04</td>
<td>-0.4830</td>
<td>6.29e-01</td>
</tr>
<tr>
<td>const(factor(Koh_us))Monica2</td>
<td>1.85e-04</td>
<td>3.55e-04</td>
<td>3.62e-04</td>
<td>0.5110</td>
<td>6.09e-01</td>
</tr>
<tr>
<td>const(factor(Koh_us))Monica3</td>
<td>2.36e-05</td>
<td>3.59e-04</td>
<td>3.71e-04</td>
<td>0.0637</td>
<td>9.49e-01</td>
</tr>
<tr>
<td>const(factor(Koh_us))Øbus2</td>
<td>2.34e-04</td>
<td>1.61e-04</td>
<td>1.62e-04</td>
<td>1.4500</td>
<td>1.47e-01</td>
</tr>
<tr>
<td>const(factor(educ))1:const(factor(smoker))1</td>
<td>1.85e-03</td>
<td>2.22e-04</td>
<td>2.29e-04</td>
<td>8.0800</td>
<td>6.66e-16</td>
</tr>
</tbody>
</table>

The first section of output contains test for significance and age-invariance of baseline rate and the effect of sex. It is seen that the effect of sex is age-varying and significant. The second part of the output (starting from “Parametric terms”) contains estimates, standard errors, and p-values for the age-invariant effects, which are the variables with the const-wraper. It is for instance seen that smoker=1, which corresponds to smokers, is associated with 24.2 extra cases pr. 10,000 person years. Confidence intervals are available as estimates plus/minus 1.96 times the robust standard error.

All estimates and confidence intervals in Table 2 can be computed from this output along with the covariance matrix of parameter estimates obtained by the function \text{vcov(fitAalen)}. However, it is easier to simply change the reference category as below:
fitAalen <- aalen(Surv(inAge, outAge, event) ~
const(factor(I(1-educ)))*const(factor(smoker)) + const(cohort)
+ factor(sex), data=SIC, start.time=50)

where $I(1-educ)$ computes the difference, and thereby change the reference level. The last part of Table 2 can be obtained by computing a new variable as the interaction of education and smoking status and then including this new variable in the fit of the additive hazards model:

SIC$interact <- interaction(SIC$educ, SIC$smoker)

fitAalen <- aalen(Surv(inAge, outAge, event) ~
const(interact) + const(cohort) + factor(sex), data=SIC,
start.time=50)

References


eAppendix 2: The mathematical details of the additive model

The additive hazard model used in the main text is a special case of the semi-parametric additive hazard model first introduced in McKeague & Sasieni (1994). This paper considered the semi-parametric additive intensity model given by

$$\rho_i(t) = Y_i(t)\{X_i^T(t)\beta(t) + Z_i^T(t)\gamma\},$$

where $Y_i(t)$ is the at risk indicator, $X(t)$ and $Z(t)$ are predictable locally bounded covariate vectors, $\beta(t)$ is a locally integrable function, and $\gamma$ is a regression vector. The special case of the model used in the main text is obtained by imposing that the non-parametric part (that is $X_i^T(t)\beta(t)$) only consists of the baseline hazard, all covariates are included in the $Z$-vector, and none of the covariates are time-dependent. There exist explicit estimation formulas for the parameter vector $\gamma$, which is our object of interest, and under the usual assumption in survival analysis of independent censoring and a technical regularity condition (Condition 5.2 of Martinussen & Scheike) the estimator of $\gamma$ is asymptotically normal (Theorem 5.3.1 of Martinussen & Scheike).

The validity of the assumption of non-time dependent effect of the covariates can be assessed by initially including all covariates in the $X$-vector in the equation above. From this model supremum type tests can be used to guide which covariates can be moved to the $Z$-vector; that is which variables have non-time dependent effects. The limiting distribution of the supremum test statistics is non-standard, but $p$-values can be computed through simulation. The mathematical result underpinning this procedure is Theorem 5.4.1 of Martinussen & Scheike.

An individual element of the regression vector $\gamma$ assesses the (adjusted) effect of that covariate. Thus effect sizes are measured on the hazard scale. As long as there are not too many events over one unit of time the effects given by that element of $\gamma$ can be interpreted as the expected number of additional cases per unit of time associated with that covariate.

Both the fully flexible semi-parametric model given in the equation above and the simpler additive model used in the main text of the paper can be analyzed using the “timereg” package in the R software. The “timereg” package includes functionality for estimating, testing, and assessing goodness-of-fit. In addition graphical output such as survival functions can also be obtained. A thorough explanation of both the package and the underlying mathematics can be found in Martinussen & Scheike Chapter 5.

References