The association of Adherence to the Mediterranean Diet with Mortality in the EPIC-Spain cohort using Flexible Parametric Survival Models

Yovaninna Alarcón Soto

Advisors: Catalina Bonet - Klaus Langohr

Master in Statistics and Operational Research
Universitat Politècnica de Catalunya - Universitat de Barcelona

Barcelona, May, 2016
Outline

1. Flexible Parametric Proportional-Hazards Models
2. Aim of the study
3. Application: EPIC-Spain cohort
4. Main Results
5. Conclusions, Discussion and Further Research
Consider a Weibull survival curve

\[ S(t) = \exp(-\lambda t^\gamma) \]

If we transform to the log cumulative hazard scale

\[
\ln[H(t)] = \ln[-\ln(S(t))] \\
\ln[H(t)] = \ln(\lambda t^\gamma) = \ln(\lambda) + \gamma \ln(t)
\]

Introducing covariates gives

\[
\ln[H(t|x_i)] = \ln(\lambda) + \gamma \ln(t) + x_i \beta
\]
We thus model on the log cumulative hazard scale

\[ \ln[H(t|\mathbf{x}_i)] = \ln[H_0(t)] + \mathbf{x}_i\beta \]

Restricted cubic splines with knots, \( k_0 \), are used to model the log baseline cumulative hazard

\[ \ln[H(t|\mathbf{x}_i)] = s(\ln(t)|\gamma, k_0) + \mathbf{x}_i\beta \]

For example with 3 knots we can write

\[ \ln[H(t|\mathbf{x}_i)] = \gamma_0 + \gamma_1 r_{1i} + \gamma_2 r_{2i} + \mathbf{x}_i\beta \]

- log baseline cumulative hazard
- log hazard ratios
Cubic splines

Source: www.le.ac.uk/hs/pl4/spline_continuity/spline_continuity.html
Cubic splines

Source: www.le.ac.uk/hs/pl4/spline_continuity/spline_continuity.html
Cubic splines

Source: www.le.ac.uk/hs/pl4/spline_continuity/spline_continuity.html
Cubic splines

Source: www.le.ac.uk/hs/pl4/spline_continuity/spline_continuity.html
Restricted cubic splines

Source: www.le.ac.uk/hs/pl4/spline_eg/spline_eg.html
Restricted cubic splines

Source: www.le.ac.uk/hs/pl4/spline_eg/spline_eg.html
A proportional cumulative hazards model can be written

$$\ln[H(t|x_i)] = s(\ln(t)|\gamma, k_0) + x_i\beta$$

- New set of spline variables for each stratum variable.
- If there are $D$ stratum variables then

$$\ln[H(t|x_i)] = s(\ln(t)|\gamma, k_0) + \sum_{j=1}^{D} s(\ln(t)|\delta_{j}, k_j) + x_i\beta$$

- $k_j$ is the set of number of knots for each stratum variable.
- $l_j, j = 1, \ldots, d_j$ is the number of categories for each stratum variable.
Aim of the study

The main goal of this study is to make a comparison between the Flexible Parametric PH Model and the traditional Cox model.

Steps

- Select the proper index of Adherence to the MD.
- Fit a Cox Model using this index.
- Fit Flexible Parametric PH model with the same index.
Application: EPIC-Spain cohort

EPIC-Spain

- European Prospective Investigation into Cancer and Nutrition.
- 5 Regions: Asturias, Murcia, Granada, Gipuzkoa and Navarra.
- Recruited began in 1992-96 (mean of follow-up: 18.5 years).

Population

- Initial sample: 41437 individuals (3676 deaths).
- Exclusions: 246 individuals (30 deaths).
  - 8 individuals leave the study.
  - 238 with implausible diet.

**Evaluable population: 41191 individuals (25612 females).**

- Until September 2014, 3646 deaths.
The model considers:

- **Adherence to the Mediterranean Diet Score (mdscore).**
- Mortality as the primary endpoint.
- **Stratum variables:**
  - Sex (2),
  - Center (5) and
  - Age at Recruitment (4).
- **Other covariates:**
  - Smoke status (10, Never Smoke),
  - BMI (3, < 25 kg/m²),
  - Physical activity (4, Moderately active)
  - Energy intake
  - Waist circumference (2, Normal) and
  - Educational level (6, Primary).

- Age was used as time scale, so data is left truncated.
- Data is also right censored.
Adherence to the Mediterranean Diet

- Four different versions based on the original Mediterranean Diet score (Trichopoulou et al., 1995). Based on intake of 9 key components:

<table>
<thead>
<tr>
<th>Presumed to fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$: fruits, $X_2$: vegetables, $X_3$: legumes, $X_4$: cereals, $X_5$: fresh fish, $X_6$: olive oil and $X_7$: wine.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Not to fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_8$: total meat and $X_9$: dairy products.</td>
</tr>
</tbody>
</table>

- Dietary questionnaire was applied to each individual.
- Food composition table (Slimani et al., 2007).
- $X_i$, $i = 1, \ldots, 9$ corresponds to density of intake for every component.
### Different versions of mdscore

<table>
<thead>
<tr>
<th>Score</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{mdscore}<em>{\text{sum}} = \sum</em>{i=1}^{9} \alpha_i X_i$</td>
<td>Consists in the sum of the 9 components.</td>
</tr>
<tr>
<td>$\text{mdscore}<em>{\text{cdf}} = \sum</em>{i=1}^{9} \alpha_i X_{\text{cdf}_i}$</td>
<td>By participant, the cdf was considered for each component.</td>
</tr>
<tr>
<td>$\text{mdscore}<em>{\text{sd}} = \sum</em>{i=1}^{9} \alpha_i X_{\text{sd}_i}$</td>
<td>Mean and sd from Granada and Murcia. Categorized by</td>
</tr>
</tbody>
</table>
|                     | $X_{\text{sd}} = \begin{cases} 
-2 & \text{if } x < -2 \\
-1 & \text{if } -2 \leq x \text{ and } x < -1 \\
0 & \text{if } -1 \leq x \text{ and } x \leq 1 \\
1 & \text{if } 1 < x \text{ and } x \leq 2 \\
2 & \text{if } x > 2 
\end{cases}$ |
| $\text{mdscore}_{\text{ter}} = \sum_{i=1}^{9} X_{\text{ter}_i}$ | Each of the 9 components were divided into tertiles (values 0, 1 and 2) and then summed. |

Where

$$\alpha_i = \begin{cases} 
1 & \text{if component fits to the MD} \\
-1 & \text{if component does not fit} 
\end{cases}$$
Choosing the best version of the Adherence to the Mediterranean Diet Score

<table>
<thead>
<tr>
<th>Score</th>
<th>min ; max</th>
<th>mean (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mdscore_sum</td>
<td>-0.79 ; 2.78</td>
<td>0.91 (0.73)</td>
</tr>
<tr>
<td>mdscore_cdf</td>
<td>-0.91 ; 6.06</td>
<td>2.63 (0.92)</td>
</tr>
<tr>
<td>mdscore_sd</td>
<td>-9 ; 12</td>
<td>0.25 (2.08)</td>
</tr>
<tr>
<td>mdscore_ter</td>
<td>0 ; 17</td>
<td>8.67 (2.58)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>HR (95% IC)(^{(1)})</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>p-value(^{(2)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>m_0</td>
<td></td>
<td>21</td>
<td>46781.35</td>
<td>46962.50</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>mdscore_sum</td>
<td>0.92 (0.87 - 0.96)</td>
<td>22</td>
<td>46771.13</td>
<td>46960.90</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>mdscore_cdf</td>
<td>0.90 (0.86 - 0.93)</td>
<td>22</td>
<td>46749.96</td>
<td>46939.73</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>mdscore_sd</td>
<td>0.90 (0.86 - 0.93)</td>
<td>22</td>
<td>46747.05</td>
<td>46936.82</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>mdscore_ter</td>
<td>0.97 (0.96 - 0.98)</td>
<td>22</td>
<td>46760.93</td>
<td>46950.70</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

\(^{(1)}\) All p-values for Cox models < 0.001  
\(^{(2)}\) p-values corresponding to Partial Likelihood Ratio Test.
Implementation in Stata

```
stset
  . stset Age_Exit, failure(cens==1) id(epic_id) enter(Age_Recr)

  . xi: stpm2 mdscore_cdf i.Smoke_Status i.Bmi i.Physical Energy i.Waist i.L_School, tvc(i.Center i.Sex i.Age) scale(hazard) df(4) dftvc(2)
```

- Patrick Royston wrote `stpm` in 2001.
- Lambert and Royston wrote `stpm2` in 2011.
## Knots combinations for Flexible Parametric Models

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>dftvc</th>
<th>AIC</th>
<th>BIC</th>
<th>df</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>11122.7</td>
<td>11476.37</td>
<td>41$^{(1)}$</td>
<td>100 iterations</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>11117.67</td>
<td>11488.59</td>
<td>43*</td>
<td>convergence achieved</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
<td>11111.92</td>
<td>11482.84</td>
<td>43</td>
<td>convergence achieved</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>2</td>
<td>11109.76</td>
<td>11497.93</td>
<td>45*</td>
<td>convergence achieved</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>11271.49</td>
<td>11676.91</td>
<td>47*</td>
<td>100 iterations</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>11277.56</td>
<td>11674.36</td>
<td>46*</td>
<td>100 iterations</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>11263.51</td>
<td>11694.81</td>
<td>50*</td>
<td>100 iterations</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>3</td>
<td>11274.65</td>
<td>11705.95</td>
<td>50*</td>
<td>100 iterations</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>4</td>
<td>11292.57</td>
<td>11775.63</td>
<td>56*</td>
<td>100 iterations</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>4</td>
<td>11299.3</td>
<td>11790.98</td>
<td>57*</td>
<td>100 iterations</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>4</td>
<td>11274.45</td>
<td>11740.25</td>
<td>54$^{(2)}$</td>
<td>100 iterations</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>4</td>
<td>11279.53</td>
<td>11762.59</td>
<td>56*</td>
<td>100 iterations</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>5</td>
<td>11285.04</td>
<td>11776.72</td>
<td>57*</td>
<td>100 iterations</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>5</td>
<td>11277.03</td>
<td>11794.59</td>
<td>60*</td>
<td>100 iterations</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>5</td>
<td>11295.92</td>
<td>11.830.73</td>
<td>62*</td>
<td>100 iterations</td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>5</td>
<td>11292.01</td>
<td>11.844.07</td>
<td>64$^{(3)}$</td>
<td>100 iterations</td>
</tr>
</tbody>
</table>

(1) variance matrix is nonsymmetric or highly singular.
(2) rcs Age $\geq 60_4$ omitted because of collinearity.
(3) rcs Age $\geq 60_5$ omitted because of collinearity.
* degrees of freedom are incorrect in the output of Stata.
\[ \ln[H_i(t|\mathbf{x}_i)] = s(\ln(t)|\gamma, k_0) + \sum_{j=1}^{D} s(\ln(t)|\delta_{ij}, k_j) + \mathbf{x}_i \beta \]

**Individual Profile**
- **mdscore:** 1st quartile
- **Center:** Granada
- **Sex:** Woman
- **Age at recruitment:** \( \geq 60 \text{ years} \)
- **Smoke status:** No smoker (ref)
- **BMI:** > 30 kg/m\(^2\)
- **Physical activity:** Moderately active (ref)
- **Energy intake** (centered): -512.6188
- **Waist circumference:** Abdominal obesity
- **Educational level:** Primary (ref)
$$\ln[H_i(t|x_i)] = s(\ln(t)|\gamma, k_0) + \sum_{j=1}^{D} s(\ln(t)|\delta_{jl}, k_j) + x_i \beta$$

<table>
<thead>
<tr>
<th>Covariates</th>
<th>$\hat{\beta}$</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI $&gt; 30$ kg/m$^2$</td>
<td>0.03</td>
<td>1</td>
</tr>
<tr>
<td>Energy Intake</td>
<td>-0.00007</td>
<td>-662.59</td>
</tr>
<tr>
<td>Abdominal Obesity</td>
<td>0.10</td>
<td>1</td>
</tr>
</tbody>
</table>
\[ \ln[H_i(t|x_i)] = s(\ln(t)|\gamma, k_0) + \sum_{j=1}^{D} s(\ln(t)|\delta_{j}, k_j) + x_i \beta \]

\[ s(\ln(t)|\gamma, k_0) = \gamma_0 + \gamma_1 r_1 + \gamma_2 r_2 + \gamma_3 r_3 + \gamma_4 r_4 \]

\[ r_1 = \ln(t) = \ln(80.3) = 1.48 \]

\[ r_j = (\ln(t) - k_j)^3_+ - h_j(\ln(t) - k_{\text{min}})^3_+ - (1 - h_j)(\ln(t) - k_{\text{max}})^3_+ \]

with \( h_j = (k_{\text{max}} - k_j)/(k_{\text{max}} - k_{\text{min}}) \)

<table>
<thead>
<tr>
<th>Knots</th>
<th>Position</th>
<th>\ln(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_1 )</td>
<td>0 (min)</td>
<td>\ln(36.2)</td>
</tr>
<tr>
<td>( k_2 )</td>
<td>25</td>
<td>\ln(61.2)</td>
</tr>
<tr>
<td>( k_3 )</td>
<td>50</td>
<td>\ln(68.6)</td>
</tr>
<tr>
<td>( k_4 )</td>
<td>75</td>
<td>\ln(75.2)</td>
</tr>
<tr>
<td>( k_5 )</td>
<td>100 (max)</td>
<td>\ln(87.2)</td>
</tr>
</tbody>
</table>

- \( \hat{\gamma}_0 = -2.48, \hat{\gamma}_1 = 0.68, \)
- \( \hat{\gamma}_2 = -0.002, \hat{\gamma}_3 = -0.02 \) and
- \( \hat{\gamma}_4 = -0.02 \)

- \( r_2 = -1.12, r_3 = -0.84, \)
- \( r_4 = -0.02 \)
\[
\ln[H_i(t|x_i)] = s(\ln(t)|\gamma, k_0) + \sum_{j=1}^{D} s(\ln(t)|\delta_{jl}, k_j) + x_i\beta
\]

\[
\sum_{j=1}^{D} s(\ln(t)|\delta_{jl}, k_j) = \delta_{121} r'_{121} + \delta_{221} r'_{221} + \delta_{341} r'_{341}
\]

\[
+ \delta_{122} r'_{122} + \delta_{222} r'_{222} + \delta_{342} r'_{342}
\]

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\delta} )</th>
<th>( r' )</th>
</tr>
</thead>
<tbody>
<tr>
<td>rcs Granada(_1)</td>
<td>0.06</td>
<td>1.48</td>
</tr>
<tr>
<td>rcs Granada(_2)</td>
<td>-0.02</td>
<td>-1.21</td>
</tr>
<tr>
<td>rcs Female(_1)</td>
<td>-0.34</td>
<td>1.48</td>
</tr>
<tr>
<td>rcs Female(_2)</td>
<td>-0.07</td>
<td>-1.21</td>
</tr>
<tr>
<td>rcs (\geq 60) years(_1)</td>
<td>0.08</td>
<td>1.48</td>
</tr>
<tr>
<td>rcs (\geq 60) years(_2)</td>
<td>-0.10</td>
<td>-1.21</td>
</tr>
</tbody>
</table>

\texttt{. predict lnH, xb}

-1.32
<table>
<thead>
<tr>
<th></th>
<th>Cox Model</th>
<th>Flexible Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR (se)</td>
<td>95% CI</td>
</tr>
<tr>
<td>mdscore (ref: 1st quartile)</td>
<td>0.90 (0.02)</td>
<td>0.86 - 0.93</td>
</tr>
<tr>
<td>2nd quartile</td>
<td>0.93 (0.04)</td>
<td>0.85 - 1.02</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>0.84 (0.04)</td>
<td>0.76 - 0.92</td>
</tr>
<tr>
<td>4th quartile</td>
<td>0.76 (0.04)</td>
<td>0.69 - 0.84</td>
</tr>
</tbody>
</table>
Comparison of Baselines

Men, 40-50 years

Women, 40-50 years

Men, 50-60 years

Women, 50-60 years

Survival vs. Years
K-M, Cox, Flexible
Comparison of survival probabilities for quartiles
Comparison of survival probabilities for quartiles
Others applications of Flexible Parametric PH model

- Estimation of survival probabilities.

- Time-dependent variables.

- Excess mortality rates in the context of relative survival.

- Prediction of survival differences.
Conclusions

- Greater adherence to the Mediterranean diet reduces the risk of mortality
  - 3rd quartile HR: 0.85 (0.77 - 0.93)
  - 4th quartile HR: 0.78 (0.71 - 0.86)

- The estimates we get from flexible parametric survival models are very similar to those obtained from a Cox model.
  - Stratified Cox Model HR: 0.90 (0.86 - 0.93)
  - Flexible Parametric HR: 0.91 (0.87 - 0.94)

- Also the baseline curves are very similar between models and with Kaplan-Meier curves.

- However stcox with basesurv() are step-functions.
Conclusions, Discussion and Further Research

Discussion

- There are some bugs in the presentation estat ic in Stata.
- Problem with initial values for convergence.
- Time-dependent variables.

Further Research

- Implementation in \( R \) for left-truncated data.
- Residual analysis implementation.
Thank you for your attention!

Questions and comments are welcome.