Choice of time-scale in the Cox model for epidemiologic cohort studies where entry has no direct biological relevance

Paul Dickman and Anna Johansson
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
October 28, 2005

Slides available at http://www.pauldickman.com/teaching/

• Time since entry is not of direct biological interest.
• The choice of variables to adjust for in a statistical model should be based, first and foremost, on biological and clinical considerations; we should only adjust for time-since entry if it has direct biological relevance.
• How do we, technically, adjust for the fact that a single exposure variable can assume multiple values for a single individual?
• One approach is to 'split' the person-time for each individual into bands, creating a data set containing multiple observations for each individual.
• This is what we do with Poisson regression; can adjust for two (but not three) time-varying explanatory variables.

Time-varying exposure in epidemiological cohort studies

• Study 1: We wish to examine the association between exposure to radioactive iodine and incidence of thyroid cancer among survivors of the Chernobyl accident. From a biological perspective it is important to consider
  – age at exposure
  – time since exposure
  – (attained age)
• Time since exposure is a ‘time-varying’ explanatory variable (the value changes with time) whereas age at exposure is fixed for each individual.
• Study 2: Invite women from the general population to participate in a cohort study; follow-up to assess the association between diet and incidence of breast cancer.
• From a biological perspective it is important to consider age at time of follow-up (attained age), a ‘time-varying’ explanatory variable.

The Cox proportional hazards model

• The ‘intercept’ in the Cox model, the hazard (event rate) for individuals with all covariates \( z \) at the reference level, is an arbitrary function of time\(^1\), often called the baseline hazard and denoted by \( \lambda_0(t) \).
• The hazard at time \( t \) for individual with other covariate values is a multiple of the baseline
  \[
  \lambda(t; z) = \lambda_0(t) \exp(\beta'z).
  \]
• Can extend the model to a ‘stratified Cox model’ which has separate baseline hazards for each level of some factor \( j = 1, \ldots, J \)
  \[
  \lambda(t; j, z) = \lambda_{0,j}(t) \exp(\beta'z).
  \]

Example: survival of patients diagnosed with colon carcinoma

• Patients diagnosed with colon carcinoma in Finland 1984–95. Potential follow-up to end of 1995; censored after 10 years.
• Outcome is death due to colon carcinoma.
• Time-scale \( t \) is time-since-diagnosis in years.
• Interested in the effect of clinical stage at diagnosis (distant metastases vs no distant metastases).

\(^1\)Time \( t \) can be defined in many ways, e.g., attained age, time-on-study, calendar time, etc.

Smoothed empirical hazards (cancer–specific mortality rates)
sts graph, by(distant) hazard

Smoothed empirical hazards on log scale
sts graph, by(distant) hazard yscale(log)
Fit a Cox model

```
. stcox distant, basehc(base)
```

failure _d: status == 1
analysis time _t: (exit-origin)/365.25
origin: time dx

| _t  | Haz. Ratio | Std. Err. | z    | P>|z| [95% Conf. Interval] |
|-----|------------|-----------|------|------------------------|
| distant | 7.190404  | .1833347  | 77.37| 0.000 6.839905 7.558863 |

```
. stcurve, hazard at1(distant=0) at2(distant=1)
```

Fitted hazards from Cox model on log scale

```
. stcox distant old, basehc(base)
```

failure _d: status == 1
analysis time _t: (exit-origin)/365.25
origin: time dx

| _t  | Haz. Ratio | Std. Err. | z    | P>|z| [95% Conf. Interval] |
|-----|------------|-----------|------|------------------------|
| distant | 7.252431  | .185139  | 77.61| 0.000 6.898494 7.624528 |
| old   | 1.57537   | .0384735 | 18.61| 0.000 1.50174 1.652611 |

```
. stcurve, hazard at1(distant=0) at2(distant=1) yscale(log)
```

Fitted hazards from Cox model

Possible models

- Let $a_0$ be the age at entry and $t$ the time-on-study.
- Using time-on-study as the time scale and adjusting for age at entry we have
  $$
  \lambda(t|a_0, z) = \lambda_0(t) \exp(\xi a_0 + \gamma z)
  $$
  (Korn model 4).
- Using attained age as the time scale we have
  $$
  \lambda(a_0, z) = \lambda_0(a) \exp(\beta z)
  $$
  (Korn model 5).
- Model (4) is appropriate for the cancer survival data but not for epidemiological cohort studies where time-on-study has no direct relevance.

Nevertheless, this model is commonly applied in epidemiology.
- Model (5) is appropriate for epidemiological cohort studies (provided there are no cohort or period effects).
- Korn et al. [1] argue for the model with age as the time-scale and stratified on birth cohorts $B_j$
  $$
  \lambda(a|b_0 \in B_j, z) = \lambda_0(b) \exp(\beta z)
  $$
  (Korn model 3)
  that is, separate baseline hazards for each birth cohort.
- We will focus on a comparison of models (4) and (5), those most commonly applied in epidemiology.
- In particular, we will study conditions under which model (5) is correct but model (4) provides estimates without large bias.
• Assume model (5) is appropriate (hazard depends on attained age and there are no period or birth cohort effects).

• We also assume, for the moment, that the exposure of interest does not vary over time.

• Korn et al. suggested two conditions under which the γ’s estimated from model (4) are similar to the β’s estimated from model (5) is
  1. the baseline hazard \( \lambda_0(a) = c \exp(\psi a) \) for some \( c > 0 \) and \( \psi \); or
  2. the baseline ages, \( a_0 \), are independent of the covariates 1.

• Thiébaut and Bénichou (2004) [2] performed simulations and observed bias even when the second condition was met.

• First condition can be written as \[ \ln(\lambda_0(a)) = \ln(c) + \psi a \];

\[ \text{we require the log hazard to be a linear function of (attained) age}. \]

---

**Example 1 from Korn et al.: condition 1 is satisfied**

---

**Example 2 from Korn et al.: condition 1 is not satisfied**

---

**Simulation study of Thiébaut and Bénichou (2004)**

• Designed to simulate risk of breast cancer in the E3N cohort, 100k French women aged 40-65 years at recruitment (1989/90).

• Exposure of interest is menopausal status at recruitment (time-fixed) and menopausal status (time-varying).

• Table I: Covariate independent of age at entry

• Table II: Covariate dependent on age at entry; \( \beta = 0 \)

• Table III: Covariate dependent on age at entry; \( \beta = \ln(5) \)
Mortality in relation to snus use; a cohort of Swedish men

- Randomly selected men (n=9976) aged 14–99 (at entry) living in Uppsala county 1973
- Participants were:
  - Invited to oral examination (at dentist)
  - Questionnaire on snus use (plus tobacco use & lifestyle factors)
  - Follow-up for cancer and death 1973-2002 via population registers
- 1427 (14%) were snus users; 8408 (84%) non-users at baseline
- If cumulative dose is the underlying exposure of interest and we model exposure (hours/day) at baseline as a fixed covariate then age-at-entry may approximate cumulative dose at entry and time-since entry may approximate cumulative dose during follow-up.

### Results from fitting various models (outcome is death)

<table>
<thead>
<tr>
<th>Time-scale</th>
<th>adj. for age entry</th>
<th>adj. for attained age</th>
<th>adj. for time-on-study</th>
<th>Stratified age entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-6 hrs/day</td>
<td>1.0 (0-1.0)</td>
<td>1.0 (0-1.0)</td>
<td>1.0 (0-1.0)</td>
<td>1.0 (0-1.0)</td>
</tr>
<tr>
<td>7-15 hrs/day</td>
<td>1.0 (0-1.0)</td>
<td>1.0 (0-1.0)</td>
<td>1.0 (0-1.0)</td>
<td>1.0 (0-1.0)</td>
</tr>
</tbody>
</table>

### Categories used
- Age-at-entry: 0-19, 20-24, 25-29, 30-34, ..., 90-94, 95+ yrs
- Attained age: 0-19, 20-24, 25-29, 30-34, ..., 90-94, 95+ yrs
- Attained follow-up: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29 yrs
- Results are preliminary and not adjusted for potential confounders (smoking, alcohol, etc.). Such adjustment will be performed after data have been cleaned.

### References