Survival Models for the Social and Political Sciences
Week 6: More on Cox Regression

JEFF GILL
Professor of Political Science
Professor of Biostatistics
Professor of Surgery (Public Health Sciences)
Washington University, St. Louis

Time Varying Covariates

- These are explanatory variables that change during the time of the study.
- Obviously it is not very realistic to assume that people, nations, etc. have all fixed right-hand-side variables.
- And the longer the period of study the more likely to have changes in these.
- So we need a way to accomodate these changes in the survival models studies so far.
- Also called *time dependent covariates*.

Time Varying Covariates, Typology

- Kalbfleisch and Prentice (1980) distinguish between:
  - internal: RHS variables at the individual level that can only be measured when the case is alive/active in the data. In medicine: blood pressure, CD4 count, cholesterol level, etc. In political science: bills debated in a parliament under a specific government, voter support for the president, etc.
  - external: RHS variables in 3 categories (Lancaster 1990):
    - fixed: a time-*independent* RHS variable with a value known in advance that does not change (race, sex, geographic location of a country, language)
    - defined: a time-*dependent* RHS variable that changes over time but these changes are exactly known (person's age, election cycle, time itself)
    - ancillary: stochastic but not affected by the event history being studied (solar flares, weather, politics in Namibia).

Time Varying Covariates, Typology

- Another typological distinction (Lancaster 1990) is:
  - ▷ exogenous: a RHS variable determined by forces outside the system being studied.
  - ▷ endogenous: the opposite.
- Exogenous explanatory TVC are relatively easy to handle, but require more elaborate data handling (as we will see here).
- True exogeneity here means that the causal path goes in only one direction: TVC to OV.
- A formal definition:

$$p(X(t, t + \Delta t)|T \geq t + \Delta t, X(t)) = p(X(t, t + \Delta t)|X(t)),$$

  which means that the extra time on the RHS of the condition in the left term does not affect values of X in the specified interval.
- This is really a substantive question, not a statistical one.

Time Varying Components, Exogenous Survival Function

- Suppose we can make the Lancaster (1990) assumption and we have discrete time data.
- This leads to the "jump process" (Peterson 1995) whereby the TVC "jumps" values from time t_j to t_{j+1}.
- The survivor function is created by piece-wise contributions from different eras for the case with a TVC:

$$S(t_k) = \prod_{j=1}^{k} p(T > t_j | T \geq t_{j-1})$$

where each of the RHS terms is gven by

$$p(T > t_j | T \geq t_{j-1}) = \exp\left(- \int_{t_{j-1}}^{t_j} h(u|\mathbf{x}_j) du \right)$$

and where $h(u|\mathbf{x}_j)$ is the hazard rate conditional on the value $\mathbf{x}_j$.

- Note also that

$$\int_{t_{j-1}}^{t_j} h(u|\mathbf{x}_j) du = H(t_k)$$

is the integrated or cumulative hazard rate in the integratable interval from t_{j-1} to t_j.

Time Varying Components, Exogenous Survival Function

- Once we have this survival function we have all of the remaining important quantities:

$$S(t_k) = \exp(-H(t_k)) \qquad \text{where} \qquad H(t_k) = \int_0^{t_k} h(x)dx \quad \longrightarrow \quad H(t_k) = -\log(S(t_k))$$

plus the derivative of the survival function gives:

$$-\frac{d}{dt_k}S(t_k) = -\frac{d}{dt_k}[1 - F(t_k)] = f(t_k)$$

and we can rewrite the hazard function as:

$$h(t_k) = -\frac{d}{dt_k}\log(S(t_k)) \qquad \text{since} \qquad -\frac{d}{dt_k}\log(S(t_k)) = -\frac{1}{S(t_k)}\frac{d}{dt_k}(-F(t_k)) = -(-1)\frac{f(t_k)}{S(t_k)}.$$

- Thus we have a recipe for all of the piece-wise functions.

Time Varying Covariates Example

- Consider modeling with piecewise constant functions over defined sub-periods.
- For example, civil status:

```
library(eha)
data(oldmort)
table(oldmort$civ)
unmarried   married     widow
      557      3638      2300

data(infants)
table(infants$civst)
married unmarried
     87        18
```

- Start with an original record that is married at time T:

$$(t_0, t, d, x(s)), \quad \text{where: } t_0 < s < t,\ t_o < T < t,$$

for start time t_0, some future time of interest t, death or right truncation status d, and time period s.

Time Varying Covariates

- Define an indicator function for marriage:

$$x(s) = \begin{cases} 0, & s \leq T \\ 1, & s > T. \end{cases}$$

  where $x(s)$ is marital status that occurs at time T.

- Use this to create two new records for this one person:
  - the first "person" is unmarried and right censored at time T:

$$(t_0, T, 0, x(s) = 0)$$

    because they get married at time T.
  - the second "person" is married and left truncated at time T:

$$(T, t, d, x(s) = 1)$$

  where person 1 becomes person 2 at exactly time T.

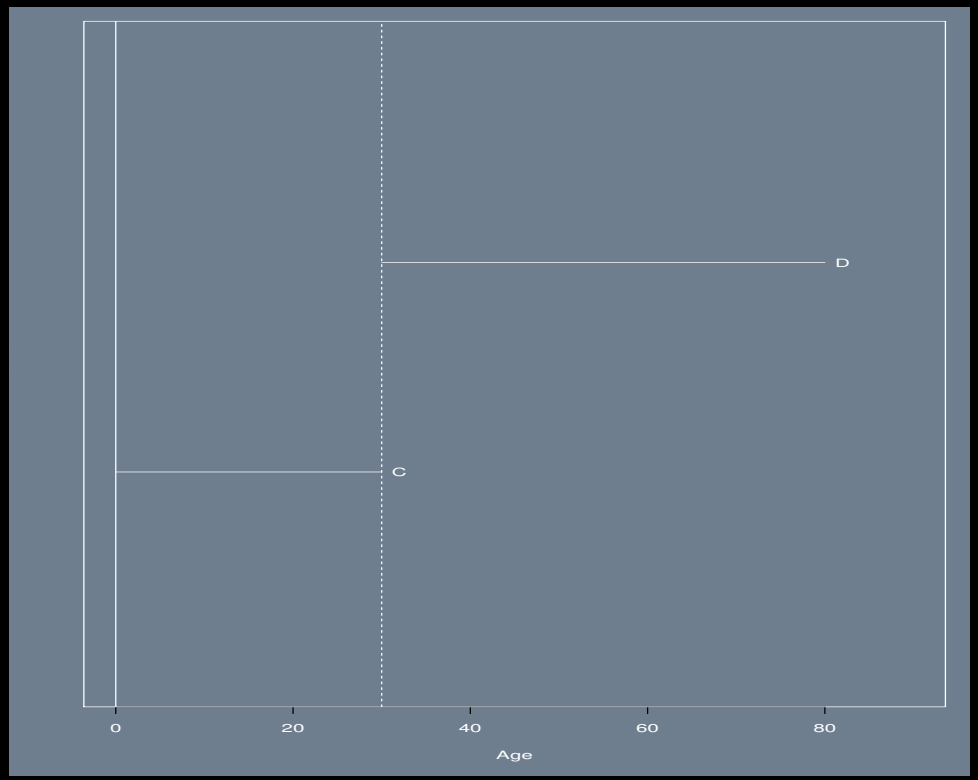
Time Varying Covariates

- An example from the book is:

id	enter	exit	event	civil status
23	0	30	0	0
23	30	80	1	1

Person 23 gets married at age 30 and dies at age 80 still married.

- This is graphically depicted:



Word of Warning

- Cox regression assumes that the risk of death is conditional only on individual characteristics.
- However, people get married (etc.) for obvious social reasons, which are time-dependent (cannot marry before a certain age, very unhealthy people unlikely to marry): "reverse causality" in the sense that event time-dependent variables can affect an explanatory variable.
- Thus time-varying parameters can possibly violate the Cox condition.
- Note that the `oldmort` dataset does not give the time of the marriage so we cannot go on further with that.

Communal Covariates

- Communal covariates are variables shared across individuals and are therefore common to these individuals at the times that they occur, also called hierarchical, grouping, or clustering variables in different contexts.
- Data: old age life histories from 1 January 1813 to 31 december 1894, with parts of life histories above age 50.
- Did food prices affect health in 19th Century Southern Sweden?
- Use the *log of the deviation from the trend* in annual rye prices.

Communal Covariates, Food Prices Example

- The Scania data dataset in the eha package:

```
data(scania)
names(scania)
[1] "id"         "enter"      "exit"       "event"      "birthdate" "sex"
[7] "parish"     "ses"        "immigrant"
> dim(scania)
[1] 1931    9
```

- The separate Log-Rye dataset in the same package:

```
dim(logrye)
[1] 94  2
summary(logrye)
      year        foodprices
Min.   :1801   Min.   :-0.37800
1st Qu.:1824   1st Qu.:-0.12425
Median :1848   Median :-0.00450
Mean   :1848   Mean   : 0.00246
3rd Qu.:1871   3rd Qu.: 0.12550
Max.   :1894   Max.   : 0.42300
```

Communal Covariates, Food Prices Example

- Looking at the data:

```
head(scania)
    id enter   exit event birthdate    sex parish   ses immigrant
29   1    50 59.242     1    1781.5   male      1 lower        no
66   2    50 53.539     0    1821.3   male      1 lower       yes
70   3    50 62.155     1    1788.9   male      1 lower       yes
83   4    50 65.142     1    1796.2 female      1 lower       yes
100  5    50 51.280     0    1822.7   male      1 lower       yes
132  6    50 70.776     1    1813.5 female      1 lower       yes

tail(scania)
        id enter   exit event birthdate    sex parish   ses immigrant
39363 1926    50 51.981     0    1842.8   male      5 lower       yes
39382 1927    50 52.674     0    1842.2   male      5 lower       yes
39404 1928    50 50.519     0    1842.3 female      5 lower       yes
39411 1929    50 53.708     0    1841.1 female      5 upper       yes
39422 1930    50 50.995     0    1843.8 female      5 upper       yes
39434 1931    50 52.027     0    1842.8   male      5 upper       yes
```

Communal Covariates, Food Prices Example

- The `eha` package has a function `make.communal` that replicates each case with different values of the communal variable *over time*, for times in which that case is observed.
- Now apply the function:

```
scania <- make.communal(scania, logrye[,2, drop=FALSE], start=1801.75)
dim(scania)
[1] 29198    10
scania[29196:29198,]
       enter   exit event birthdate   id  sex parish   ses immigrant foodprices
29196 50.000 50.946     0    1842.8 1931 male      5 upper       yes      0.117
29197 50.946 51.946     0    1842.8 1931 male      5 upper       yes      0.007
29198 51.946 52.027     0    1842.8 1931 male      5 upper       yes      0.006
```

so this is an individual measured over 3 years only and right censored out of the sample.

Communal Covariates, Food Prices Example

- Here is the first person:

```
scania[scania$id == 1,]
    enter   exit event birthdate id  sex parish   ses immigrant foodprices
1  50.000 50.296     0    1781.5  1 male      1 lower        no      0.126
2  50.296 51.296     0    1781.5  1 male      1 lower        no      0.423
3  51.296 52.296     0    1781.5  1 male      1 lower        no     -0.019
4  52.296 53.296     0    1781.5  1 male      1 lower        no     -0.156
5  53.296 54.296     0    1781.5  1 male      1 lower        no     -0.177
6  54.296 55.296     0    1781.5  1 male      1 lower        no     -0.085
7  55.296 56.296     0    1781.5  1 male      1 lower        no     -0.007
8  56.296 57.296     0    1781.5  1 male      1 lower        no      0.104
9  57.296 58.296     0    1781.5  1 male      1 lower        no      0.118
10 58.296 59.242     1    1781.5  1 male      1 lower        no     -0.197
```

so this individual was measured over 10 years and died at age 59.242.

Communal Covariates, Food Prices Example

- Now run a Cox PH model with the communal variable:

```
rye.fit <- coxreg(Surv(enter,exit,event) ~ ses + sex + foodprices, data=scania)
summary(rye.fit)
Covariate           Mean       Coef     Rel.Risk   S.E.    Wald p
ses
          upper    0.202     0          1 (reference)
          lower    0.798    -0.030     0.970     0.075     0.686
sex
           male    0.504     0          1 (reference)
         female    0.496     0.041     1.042     0.061     0.500
foodprices         0.002     0.321     1.378     0.182     0.077

Events                    1086
Total time at risk         26979
Max. log. likelihood      -7184.1
LR test statistic         3.70
Degrees of freedom        3
Overall p-value           0.295547
```

Cox Regression Output Values

- The number of `1` values in the analyzed data:

  ```
  Events                    1086
  ```

  same as `table(scania$event)[2]`.

- The sum of: years times people at risk during that year:

  ```
  Total time at risk         26979
  ```

  same as `sum(scania$exit - scania$enter)`.

- The log-likelihood function at its maximum value:

  ```
  Max. log. likelihood      -7184.1
  ```

- A likelihood ratio test for superiority over the null model

  ```
  LR test statistic         3.70
  Degrees of freedom        3
  Overall p-value           0.295547
  ```

Counting Processes and Duration Data with TVCs, Setup

- Consider an event that can re-occur: "start stop data."
- An individual enters the risk period (starts) the process at time t_0 and is observed to time t: failing or surviving or being right-censored.
- Let δ denote whether or not the case has failed or become right-censored.
- Then the observed survival information is fully described by (t_0, t, δ).
- Therefore the case is observed through the period $(t_0, t]$ and either fails ($\delta = 1$) or survives ($\delta = 1$).

Counting Processes and Duration Data with TVCs

- The TVC has some value $X = x_1$ at time t_0, and changes at time t to $X = x_2$.
- It then stays constant in the interval $(t, t_k]$.
- In both intervals δ status is observed.
- This setup can also accomodate discontinuous intervals where a case is in the observation set but is not at risk for the event.
- Also: time dependent strata, left truncation, alternative time scales, repeated events, and more.

Cox Model Likelihood Function with TVCs, Likelihood Function

- The Cox model is easy to adapt for TVCs since the partial likelihood is a function of ordered failure times not actual durations.
- So contributions to the likelihood function can come from the same individual with different covariates.
- Assuming no ties, for individual i:

$$L(\boldsymbol{\beta}|\mathbf{x}_i) = \prod_{j=1}^{p} \left[\frac{\exp(\mathbf{x}_{ji}(t_i)\boldsymbol{\beta}_j}{\sum_{k \in R(t_i)} \exp(\mathbf{x}_{ki}(t_i)\boldsymbol{\beta}_j)} \right]^{\delta_i}$$

and

$$\ell(\boldsymbol{\beta}|\mathbf{x}_i) = \sum_{j=1}^{p} \delta_i \left[\mathbf{x}_{ji}(t_i)\boldsymbol{\beta}_j - \log \sum_{k \in R(t_i)} \exp(\mathbf{x}_{ki}(t_i)\boldsymbol{\beta}_j) \right].$$

where p is the number of explanatory variables, and δ_i is again the event indicator with 0 for censored survival time and 1 otherwise.

- The $\boldsymbol{\beta}$ estimates are the change in the log-hazard ratio for a one unit change in the corresponding covariate at time t compared to the same covariate for the other observations/cases in the risk set at time t.

Cox Model Hazard with TVCs, Hazard Function

- Consider $i = 1, \ldots, n$ cases where x_{ji} is the baseline category of the jth disecrete explanatory variable: $X_j,\ j = 1 \ldots, p$.
- The hazard function for individual i at time t is given by:

$$h_i(t) = h_0(t) \exp\left[\sum_{j=1}^{p} x_{ji}\boldsymbol{\beta}_j\right]$$

  and $h_0(t)$ is the baseline hazard function.
- Now designate $x_{ji}(t)$ as the value of the jth TVC at time t.
- The hazard function becomes:

$$h_i(t) = h_0(t) \exp\left[\sum_{j=1}^{p} x_{ji}(t)\boldsymbol{\beta}_j\right]$$

- The baseline hazard function is the hazard for a case where all explanatory variables are 0 at time t_0 and stay that way to time t.

Cox Model Hazard with TVCs, Cumulative and Baseline Hazard Functions

- TVC values need to be incorporated for all cases in the risk set in a given period, $t_k \leq t < t_{k+1}$.
- Assume no ties and r total events.
- The Nelson-Aalen estimate of the cumulative hazard is:

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = \sum_{j=1}^{k} \frac{d_j}{\sum_{\ell \in R(t_j)} \exp(\mathbf{x}_\ell(t)\hat{\boldsymbol{\beta}})}$$

  where d_j are the number of deaths at time t_j (limited to one for Cox), and $k = 1, \ldots, r-1$.

- The baseline hazard function estimate is similar:

$$\hat{h}_0(t) = \frac{d_j}{(t_{j+1} - t_j) \sum_{\ell \in R(t_j)} \exp(\mathbf{x}_\ell(t)\hat{\boldsymbol{\beta}})}.$$

- And the estimate of the baseline survival function is simply:

$$\hat{S}_0(t) = \exp(-\hat{H}_0(t)).$$

Cox Model Hazard with TVCs, Individual Survival Function

- The individual survival function for individual i at time t is given by:

$$\hat{S}_i(t) = \exp\left[-\int_0^t \exp\left(\sum_{j=1}^{p} \mathbf{x}_{ji}(u)\boldsymbol{\beta}_j\right) h_0(u)du\right].$$

- The problem with this calculation is that it depends on future values of TVCs, which will not be known until the end of the study.

Cox Model Hazard with TVCs

- Since the values of x_{ji} change with time, so does the hazard ratio: $h_i(t)/h_0(t)$ for this case.
- So the hazard at time t is not proportional to the baseline hazard anymore, disqualifying this specification as proportional hazards model.
- Suppose we have two cases, k and ℓ, who have the exact same covariate values except for the $x_j$values, which differ by 1.
- The log ratio of hazards for these cases at time t is:

$$\begin{aligned}\log\left[\frac{h_k(t)}{h_\ell(t)}\right] &= \boldsymbol{\beta}_1(x_{k1}(t) - x_{\ell 1}(t)) + \cdots + \boldsymbol{\beta}_j(x_{kj}(t) - x_{\ell j}(t)) + \cdots + \boldsymbol{\beta}_p(x_{kp}(t) - x_{\ell p}(t)) + \cdots \\ &= \boldsymbol{\beta}_j(x_{kj}(t) - x_{\ell j}(t)) \\ &= \boldsymbol{\beta}_j\end{aligned}$$

meaning that $\boldsymbol{\beta}_j$ *is* the log hazard ratio.

Cox Model Hazard with TVCs

- Suppose now we have three individuals/cases labelled k, ℓ, and m.
- Case k dies at time t_k, case ℓ dies at time t_ℓ, case m right-censors out at the end of the study t_m such that $t_k < t_\ell < t_m$.
- Recall the log likelihood (now for all cases) at time t_i:

$$\ell(\boldsymbol{\beta}|\mathbf{x}) = \sum_{i=1}^{n} \left[\sum_{j=1}^{p} \delta_i \left(\mathbf{x}_{ji}(t_i)\boldsymbol{\beta}_j - \log \sum_{j \in R(t_i)} \exp(\mathbf{x}_{ji}(t_i)\boldsymbol{\beta}) \right) \right]$$

- Stipulate only one explanatory variable for simplicity.
- So at time t_k this is just:

$$\ell(\boldsymbol{\beta}|\mathbf{x}) = \mathbf{x}_k(t_k)\beta - \log \sum_{j \in R(t_k)} \exp(\mathbf{x}_{ji}(t_k)\boldsymbol{\beta})$$

  where the x are the TVC values for the individuals at time t_k.
- The time t_k simplifies because only one case is allowed to fail at a time.

Cox Model Hazard with TVCs

- So at exactly t_k, our small example is:

$$\ell(\boldsymbol{\beta}|\mathbf{x}) = \mathbf{x}_k(t_k)\beta - \log\left[\exp(\mathbf{x}_k(t_k)\boldsymbol{\beta}) + \exp(\mathbf{x}_\ell(t_k)\boldsymbol{\beta}) + \exp(\mathbf{x}_m(t_k)\boldsymbol{\beta})\right]$$

  which highlights that the TVC values for each case in the risk set is necessary in the calculation of the log-likelihood at the time that k dies.

- Moving forward in time to when ℓ dies:

$$\ell(\boldsymbol{\beta}|\mathbf{x}) = \mathbf{x}_\ell(t_\ell)\beta - \log\left[\exp(\mathbf{x}_\ell(t_\ell)\boldsymbol{\beta}) + \exp(\mathbf{x}_m(t_\ell)\boldsymbol{\beta})\right]$$

  we see that the TVC value for both remaining cases is needed here.

- Trivially, the next period looks like:

$$\ell(\boldsymbol{\beta}|\mathbf{x}) = \mathbf{x}_m(t_m)\beta - \log\left[\exp(\mathbf{x}_m(t_\ell)\boldsymbol{\beta})\right] = 0.$$

Cox Model Hazard with TVCs, Complications

- This dependency on the TVC values for all cases in the risk at the time of an individual can provide some challenges.
- It is assumed that the non-terminating TVC values are measured at the time of termination.
- So what about continuous TVCs in this issue: it could be possible that the other cases are not measured at that exact momement.
- Consider blood pressure measurement in cardiovascular disease when patient k in the study dies in the hospital.
- It is likely that we have a good measurement on k just before t_k, but the staff are unlikely to rush to the bedsides of the other patients so statisticians can have better data.
- What can we do?
  - interpolation
  - last value carried forward
  - imputation.

Parametric Models with TVCs

- The likelihood function is constructed from k successive intervals, which are actual time now.
- The hazard rate for the Weibull model is now:

$$h[t\mathbf{x}(t^-] = \exp(\mathbf{x}^-\boldsymbol{\beta})p[\exp(\mathbf{x}^-\boldsymbol{\beta})t]^{p-1}$$

  where t^- denotes where an $\mathbf{x}$ change to $\mathbf{x}^-$occurs prior to time t.

- Abbreviate $\lambda = \exp(\mathbf{x}^-\boldsymbol{\beta})$, so that the survival function is:

$$S[t\mathbf{x}(t^-)] = \exp[-(\lambda t)^p]$$

  and the density function is:

$$f[t\mathbf{x}(t^-)] = \lambda p(\lambda t)^{p-1}\exp[-(\lambda t)^p].$$

- This leads to the likelihood for t duration times for k intervals:

$$L(\boldsymbol{\beta}|\mathbf{x}) = \prod_{i=1}^{K}\left[\lambda p(\lambda t)^{p-1}\exp[-(\lambda t)^p]\right]^{\delta_i}\left[\exp[-(\lambda t)^p]\right]^{1-\delta_i}.$$

Discrete Time Models with TVCs

- Consider a logit model in discrete time.
- The outcome variable is equivalent to the binary cesnoring variable discussed in BSJ.
- When the indicator is 0, then the observation is assumed to be at risk.
- When the indicator is 1, then the observation is assumed to have failed.
- Again each observation can contribute multiple records, with different $\mathbf{x}$ values.
- For the logit model the estimated $\boldsymbol{\beta}$ for the TVC $\mathbf{x}$ gives how much the log-odds of an event occuring changes for a one-unit change in the TVC.
- Also $\boldsymbol{\beta}$ gives change in hazard probability for a one unit change in the TVC.

Discrete Time Models with TVCs, Example

- This example uses the running BSJ example of congressional career paths, where t is the number of 2 years terms in the House of Representatives.
- There are 5 dichotomous TVCs:
  - Redistricting: whether the member's district was substantially redistricted
  - leadership: whether the member has a leadership position
  - Scandal: whether or not the member was involved in a scandal
  - Prior Margin: the percentage of votes the member received in their prior election
  - Party identification: 0 for Democrat, 1 for Republican.
- BSJ specify a logit model and duration dependency is accounted for using lowess.

Discrete Time Models with TVCs, Example

- Logit model of House Careers:

Variable	Estimate (SE)	exp[coef]
Party Identfication	-0.15 (0.11)	3.158
Redistricting	1.15 (0.30)	0.861
Leadership	-0.91 (0.49)	0.403
Scandal	2.65 (0.13)	14.15
Prior Margin	-0.04 (0.005)	0.961
Duration Dependency	5.61 (2.92)	
N	5399	
Log-Likelihood	-1169.09	

- Coefficient estimates here give the change in log-odds of an event for a one-unit change in the corresponding covariate.
- We can convert these to effects on odds ratios by exponentiating them.

Tied Event Times, Broström Chapter 5

- Tied events can occur with continuous data due to the way that events are recorded.
- Too many tied events lead to biased model results.
- General strategy: permute all tied events in each risk set, the *exact method*.
- This can be *really* slow, so there are some good shortcut approximations:
  - Efron's method, usual default
  - Breslow's method, also common
  - ML: purely discrete with one parameter per observation
  - MPPL: mixes Efron and discrete approaches

  where all of these are modifications of the likelihood and score functions.

New Function, Broström Chapter 5

- The function `risksets` takes a survival object that you define with `Surv` and returns:
  - `antrs`: the number of risk sets in each stratum
  - `risktimes`: ordered distinct failure time points.
  - `eventset`: if 'members' is TRUE, a vector of pointers to events in each risk set, else NULL.
  - `riskset`: if 'members' is TRUE, a vector of pointers to the members of the risk sets, in order.
  - `size`: the sizes of the risk sets.
  - `n.events`: the number of events in each risk set.

```
risksets(Surv(fert$next.ivl, fert$event))$antrs     [1] 1881
risksets(Surv(fert$next.ivl, fert$event))$risktimes # GIVES LONG VECTOR
risksets(Surv(fert$next.ivl, fert$event))$eventset # GIVES LONG VECTOR
risksets(Surv(fert$next.ivl, fert$event))$riskset # GIVES LONG VECTOR
risksets(Surv(fert$next.ivl, fert$event))$size # GIVES LONG VECTOR
risksets(Surv(fert$next.ivl, fert$event))$n.events
[1]  3  2  1  2  2  3  3  3  3  2  1  1  3  2  2  1  1  1
:
[1873]  1  1  1  1  1  1  1  1  1
```

Tied Event Times, Broström Chapter 5 Example

- Return to the fertility dataset `fert`.
- The variable `next.ivl` (length of the coming time interval) has many ties:

```
table(risksets(Surv(fert$next.ivl, fert$event))$n.events)
  1   2   3   4   5   6   7   8   9  10  11
443 251 173 137 114  95 108  88  99  74  67
 12  13  14  15  16  17  18  19  20  21 23
 60  46  36  28  19  17   7  12   3   3  1
```

- First subset the `fert` dataset (12169×9) to make the problem manageable:

```
fert1 <- fert[fert$parity == 1,]    # PREVIOUS BIRTH WAS FIRST CHILD
dim(fert1)
[1] 1840    9
table(risksets(Surv(fert1$next.ivl, fert1$event))$n.events)
  1   2   3   4   5   6   7   9
369 199 133  60  27  14   2   2
```

using `parity`: the order of the previous birth is set to 1.

Tied Event Times, Broström Chapter 5 Example

- Now use each method in a loop:

```
meth <- c("efron","breslow","mppl","ml")
for (i in 1:length(meth))  {
    fit1 <- coxreg(Surv(next.ivl,event) ~ year + age, data=fert1, method=meth[i])
    print(sprintf("%s %.5f %.5f", meth[i], coef(fit1)[1], coef(fit1)[2]))
}
[1] "efron 0.00166 -0.03905"
[1] "breslow 0.00166 -0.03905"
[1] "mppl 0.00166 -0.03905"
[1] "ml 0.00166 -0.03905"
```

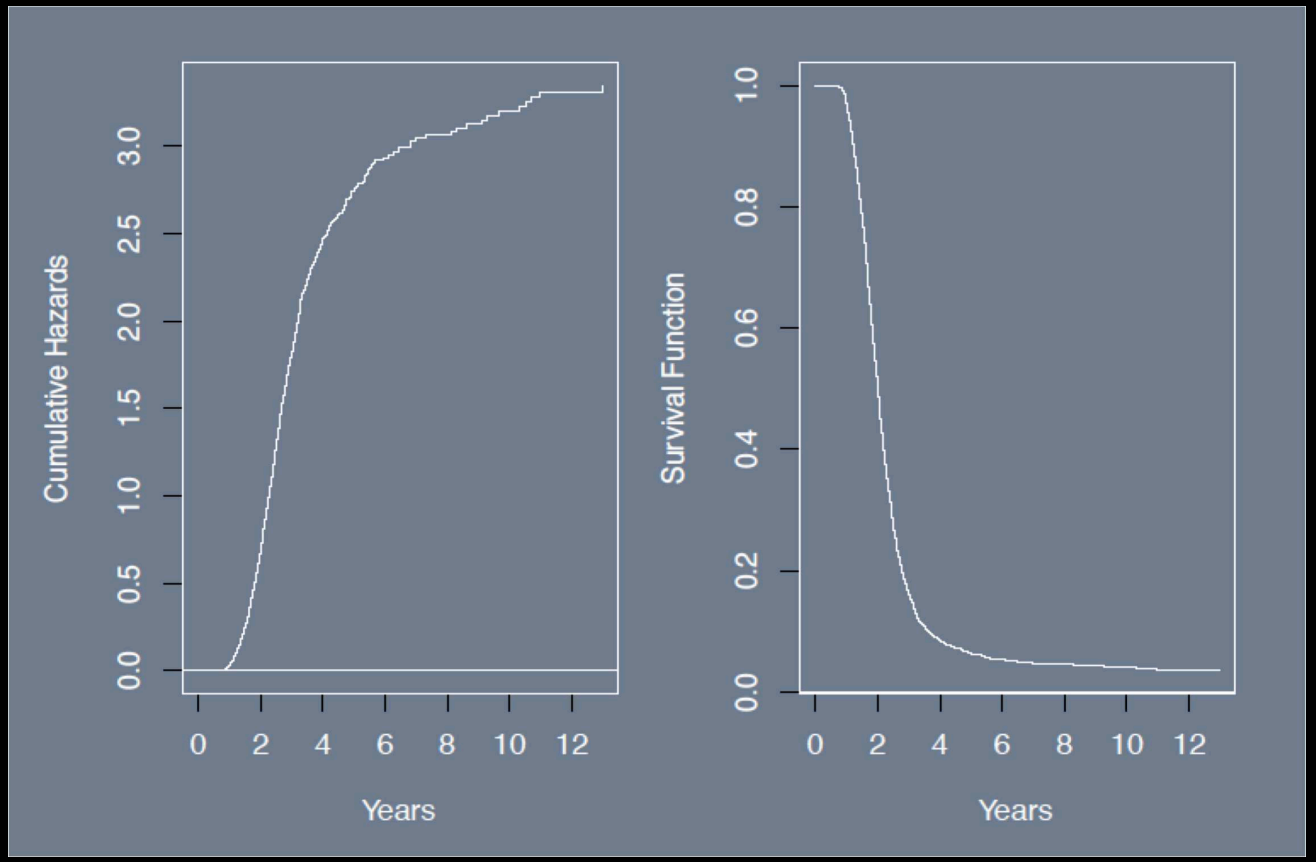
- These are the coefficient estimates for `year` and `age`, but we could have looked at the whole output.

Fertility Data, Nelson-Aalen and Survivor Plots

- A convenient way to make graphs of cumulative hazards and survival plots with the `risksets` looking at the next interval.

```
fert1.rs <- risksets(Surv(fert1$next.ivl, fert1$event))
par(mar=c(4,4,1,1),oma=c(1,1,1,1),mfrow=c(1,2),col.axis="white",
    col.lab="white",col.sub="white",col="white", bg="slategray")
plot(fert1.rs$risktimes, cumsum(fert1.rs$n.events/fert1.rs$size),
    type="s",xlab="Years",ylab="Cumulative Hazards")
abline(h=0)
plot(fert1.rs$risktimes, exp(-cumsum(fert1.rs$n.events/fert1.rs$size)),
    type="s",xlab="Years",ylab="Survival Function")
abline(h=0)
```

Fertility Data, Nelson-Aalen and Survivor Plots



Stratification

- This allows a factor to be adjusted for without having to estimate its effect directly, like groups in a multilevel model.
- Stratification involves segmenting a dataset along the categories of an explanatory variables.
- Sometimes this is done because the strata are theoretically important and hypothesis tests are performed concerning whether the group distinctions matter.
- Sometimes this is done because the proportional hazards assumption does not hold otherwise for a variable in un-stratified form.
- Stratifying complicates likelihood estimation since separate partial likelihoods are required for each strata with a final product thereafter.

Stratification

- Return to the fertility dataset and stratify along the `ses` data in a Cox PH model:

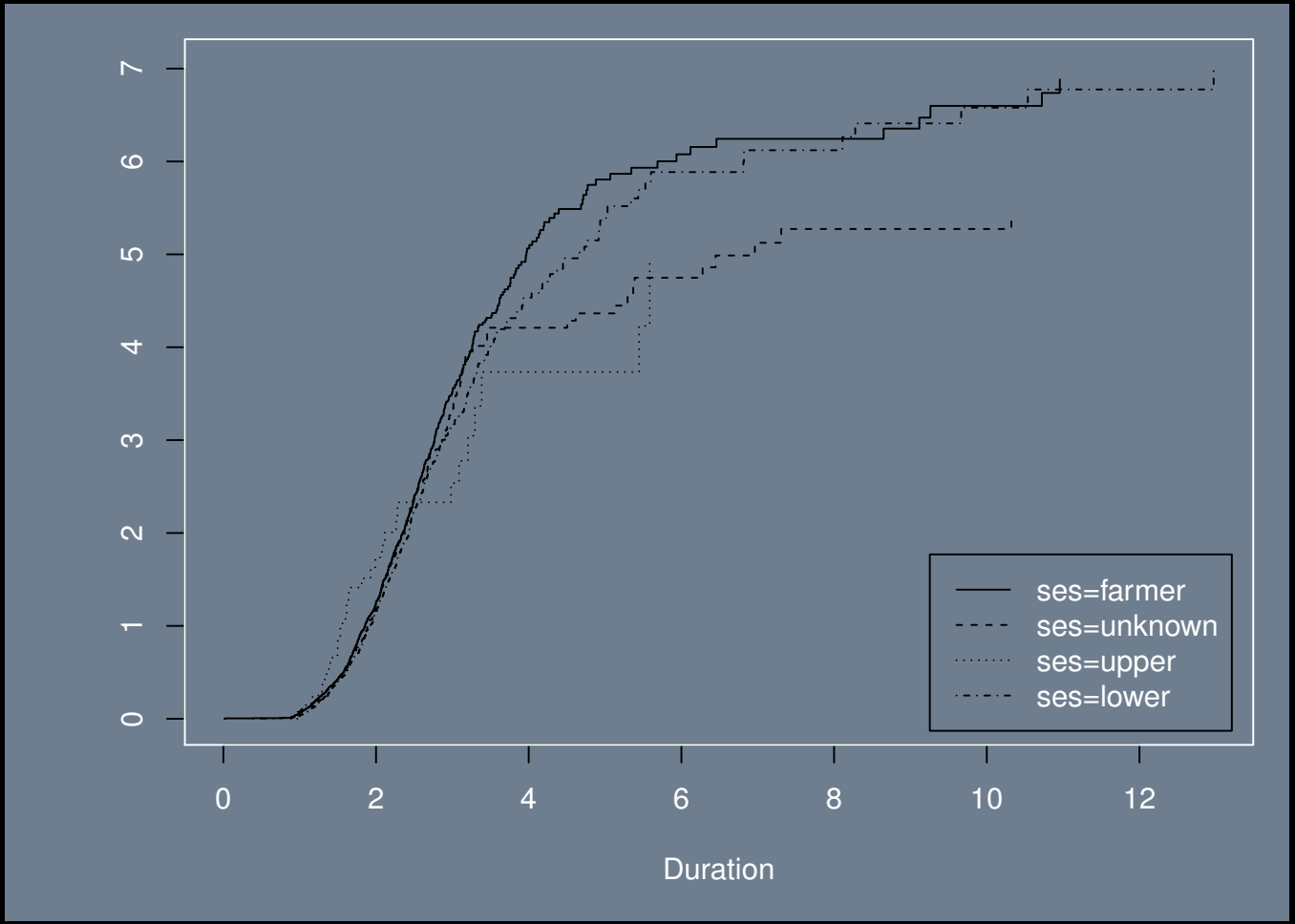
```
strat.fit <- coxreg(Surv(next.ivl, event) ~ strata(ses) + age + year + prev.ivl
                + parish, data=fert1)
summary(strat.fit)
Covariate           Mean       Coef     Rel.Risk   S.E.    Wald p
age                27.151    -0.023     0.978     0.006     0.000
year             1858.664    -0.001     0.999     0.002     0.750
prev.ivl            1.309    -0.225     0.799     0.029     0.000
parish
             JRN    0.010     0         1 (reference)
             NOR    0.053    -0.433     0.649     0.233     0.064
             SKL    0.937    -0.620     0.538     0.214     0.004

Events                    1657
Total time at risk        4500.5
Max. log. likelihood      -9047.3
LR test statistic         155.08
Degrees of freedom        5
Overall p-value           0
```

Stratification

- We cannot directly see the effects of stratification other than comparing coefficients to a non-stratified model.
- So graph the results:

```
fert1.rs <- risksets(Surv(fert1$next.ivl, fert1$event))
par(mar=c(5,5,1,1),mfrow=c(1,1),col.axis="white",
    col.lab="white",col.sub="white",col="white", bg="slategray")
plot(strat.fit)
```

Fertility Data Stratified, Nelson-Aalen and Survivor Plots



Stratification Example: VA Lung Cancer

- The `survival` package has the Veterans' Administration Lung Cancer study that is a frequent example in books and lectures:

```
data(veteran); names(veteran)
[1] "trt" "celltype" "time" "status" "karno" "diagtime" "age" "prior"
```

- The variables are:
  - `trt`: control versus treatment.
  - `celltype`: tumor cell type: small, large, squamous, adenocarcinom.
  - `time`: survival time
  - `status`: censoring status
  - `karno`: Karnofsky score indicating the patients overall baseline status, 80-100 fairly normal functioning, 50-80 unable to work but able to live at home and care for most personal needs, 1-50 cannot care for own self, 0 dead.
  - `diagtime`: time in months from diagnosis to randomization.
  - `age`: age in years at randomization.
  - `prior`: 0/10 for previous therapy of some type.

Stratification Example: VA Lung Cancer

- Fit and summarize a model stratifying on Prior Treatment:

```
strat.fit2 <- coxreg(Surv(time,status) ~ trt + factor(celltype) + karno + age
                + strata(prior), data=veteran)
summary(strat.fit2)
```

Covariate	Mean	Coef	Rel.Risk	S.E.	Wald p
trt	1.523	0.308	1.361	0.206	0.135
factor(celltype)					
squamous	0.421	0	1 (reference)		
smallcell	0.206	0.829	2.292	0.271	0.002
adeno	0.104	1.139	3.124	0.298	0.000
large	0.269	0.367	1.443	0.283	0.195
karno	68.419	-0.032	0.969	0.005	0.000
age	57.379	-0.008	0.992	0.009	0.406

Events	128	Total time at risk	16663
Max. log. likelihood	-401.87	LR test statistic	58.52
Degrees of freedom	6	Overall p-value	9.00667e-11

Rare Events

- What if there were very few events in a large dataset?
- This can be computationally expensive without much information gained from the large number of right censored cases.
- The `eha` package has the function `max.survs` for this purpose, which sets the upper limit of the number of survivors in each risk set and samples up to this limit.
- Comparing the full model and a model from sampling survivors for each risk set follows.

Rare Events, Regular Model

```
strat.fit2 <- coxreg(Surv(next.ivl, event) ~ strata(ses) + age + year + prev.ivl
                + parish, data=fert1)
summary(strat.fit2)
Covariate           Mean       Coef     Rel.Risk   S.E.    Wald p
age                27.151    -0.023     0.978     0.006     0.000
year             1858.664    -0.001     0.999     0.002     0.750
prev.ivl            1.309    -0.225     0.799     0.029     0.000
parish
    JRN    0.010     0          1 (reference)
    NOR    0.053    -0.433     0.649     0.233     0.064
    SKL    0.937    -0.620     0.538     0.214     0.004

Events                    1657
Total time at risk        4500.5
Max. log. likelihood      -9047.3
LR test statistic         155.08
Degrees of freedom        5
Overall p-value           0
```

Rare Events, Shortened Model

```
strat.fit3 <- coxreg(Surv(next.ivl, event) ~ strata(ses) + age + year + prev.ivl
                + parish, data=fert1, max.survs=10)
summary(strat.fit3)
Covariate           Mean       Coef     Rel.Risk   S.E.    Wald p
age                27.151    -0.023     0.977     0.006     0.000
year             1858.664    -0.001     0.999     0.002     0.800
prev.ivl            1.309    -0.216     0.806     0.030     0.000
parish
    JRN    0.010     0          1 (reference)
    NOR    0.053    -0.372     0.690     0.249     0.136
    SKL    0.937    -0.551     0.576     0.229     0.016

Events                    1657
Total time at risk        4500.5
Max. log. likelihood      -3948.8
LR test statistic         134.21
Degrees of freedom        5
Overall p-value           0
```

Analyzing Residuals, Cox-Snell

- Unfortunately residuals analysis is not as straightforward as with linear or even generalized linear models.
- Consider survival time T and $S(t)$ is a survivor function, then:

$$U(t) = S(t) = \int_0^t s(t)dt$$

  is uniformly distributed on $[0:1]$ by the *probability integral transformation*.
- Therefore $E(t) = -\log(U(t))$ is exponentially distributed with parameter 1.
- So given a set of survival times, $T_1, \ldots, T_n$, the Cox-Snell residuals are defined as:

$$r_{C_i} = \exp(\mathbf{X}_i\boldsymbol{\beta})\hat{H}_0(T_i|\mathbf{X}_i)$$

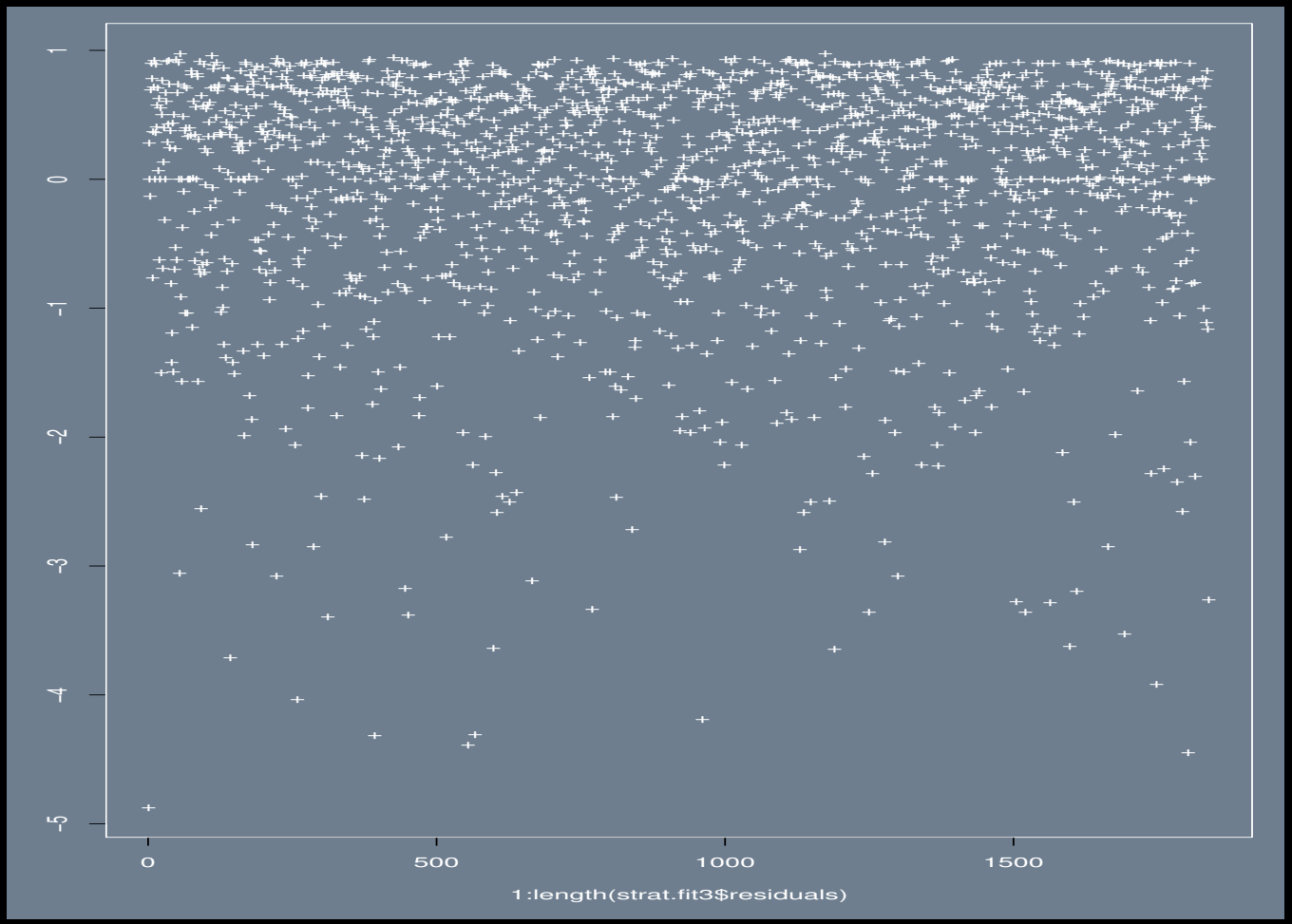
  where $\hat{H}_0(T_i|\mathbf{X}_i)$ is an estimate of the baseline cumulative hazard, and usually this is estimated with Nelson-Aalen: $\hat{H}_0(t) = \log \hat{S}_0(t) = \sum_{j=1}^{k} \frac{d_j}{\sum_{\ell \in R(t_j)} \exp(\mathbf{x}_\ell(t)\hat{\boldsymbol{\beta}})}$.
- Ideally these should look like a sample from $\mathcal{E}(1)$.

Analyzing Residuals, Martingale

- The Cox-Snell residuals include censored values so in practice Martingale residuals (r_{M_i}) are used.
- Essentially: the martingale residual = excess observed events = observed events − (expected events|the model fit).
- Instructions:
  - ▷ add 1 to the censored residuals making them look like uncensored residuals
  - ▷ subtract 1 from all of the residuals
  - ▷ multiply by -1.
- The residuals object from running `coxreg` defaults to a set of Martingale residuals.
- Sometimes these are hard to interpete graphically, and they are more commonly used in model diagnostics.

```
par(mar=c(5,3,1,1),mfrow=c(1,1),col.axis="white",
    col.lab="white",col.sub="white",col="white", bg="slategray")
plot(1:length(strat.fit3$residuals),strat.fit3$residuals,pch="+")
```

Analyzing Residuals



Analyzing Residuals, Deviance

- Although Martingale residuals have some nice properties, such as summing to zero in large samples and mean zero for uncensored observations, they can be difficult to analyze in practice due to skewness.
- A modification called deviance residuals is an attempt to make martingale residuals more symmetric:

$$r_{D_i} = \text{sign}(r_{M_i}) \left[-2(r_{M_i} + \delta_i \log(\delta_i - r_{M_i}))\right]^{\frac{1}{2}}$$

  and δ_i is again the event indicator with 0 for censored survival time and 1 otherwise.
- The model deviance (contrasted with the saturated deviance and the null deviance) is simply:

$$D = \sum_{i=1}^{n} r_{D_i}^2$$

  which leads directly to model comparison.
- If the fitted model is "correct" then the deviance residuals are symmetric about zero (although they do not necessarily sum to zero).

Analyzing Residuals, Schoenfeld Residuals

- One disadvantage in using the three residual types discussed is that the all depend on an estimate of the cumulative hazard function.
- The Schoenfeld residual for the ith individual and the jth explanatory variable is:

$$r_{S_{ji}} = \delta_i(x_{ji} - \hat{a}_{ji}),$$

where:

$$\hat{a}_{ji} = \frac{\sum_{\ell \in R(t_i)} \exp(\mathbf{x}_{j\ell}\hat{\boldsymbol{\beta}})}{\sum_{\ell \in R(t_i)} \exp(\mathbf{x}_{\ell}\hat{\boldsymbol{\beta}})}$$

and $\ell \in R(t_j)$ is the riskset at time t_i.

- So there is a residual for explanatory variable for each case.
- Non-zero values only occur with uncensored observations.
- If the largest T is uncensored then $\hat{a}_{ji} = x_{ji}$, so $r_{S_{ji}} = 0$.
- Since these are not zero in the residual sense, they are usually given `NA`.

Analyzing Residuals, Schoenfeld Residuals

- A more useful version for detecting outliers is the scaled (or weighted) Schoenfeld residual, in vector form:

$$r^*_{S_i} = r\text{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{r}_{S_i}$$

  where:

  - r is the number of events (deaths)
  - $\text{Var}(\hat{\boldsymbol{\beta}})$ is the variance-covariance matrix of the estimated coefficients from the Cox model
  - $\boldsymbol{r}_{S_i}$ is the vector of Schoenfeld residuals.

- For more details see Grambsch and Therneau (1994).

Analyzing Residuals, Score Residuals

- The score residual is also obtained from the first derivative of the log of the partial likelihood function with respect to $\boldsymbol{\beta}$.
- Using the same terminology as the Schoenfeld residual for time t_i, define:

$$\frac{\partial \log L(\boldsymbol{\beta}|\mathbf{X})}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{n} \left[\delta_r (x_{ji} - a_{ji}) + \exp(\mathbf{X}_i \boldsymbol{\beta}) \sum_{t_r < t_i} \frac{(a_{jr} - x_{ji})\delta_r}{\sum_{\ell \in R(t_r)} \exp(\mathbf{x}_\ell \boldsymbol{\beta})} \right]$$

where a_{jr} does not have a hat because this is not yet based on estimated quantities:

$$a_{ji} = \frac{\sum_{\ell \in R(t_i)} \exp(\mathbf{x}_{j\ell} \boldsymbol{\beta})}{\sum_{\ell \in R(t_i)} \exp(\mathbf{x}_\ell \boldsymbol{\beta})}$$

- Notice that the second summation accounts for all periods prior to t_i.

Analyzing Residuals, Score Residuals

- The score residual for the ith individual and the jth explanatory variable is:

$$r_{SC_{ji}} = \delta_r(x_{ji} - \hat{a}_{ji}) + \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}) \sum_{t_r<t_i} \frac{(a_{jr} - x_{ji})\delta_r}{\sum_{\ell\in R(t_r)} \exp(\mathbf{x}_\ell\hat{\boldsymbol{\beta}})}$$

- Like the Schoenfeld residuals these sum to zero in large enough samples, except for data with censoring.

- In fact, the first part *is* the Schoenfeld residual, $\delta(x_{ji} - \hat{a}_{ji})$, so these could also be considered a modified Schoenfeld residual like the scaled version.

Analyzing Residuals, Comparison

```
postscript("Class.Survival/Images/residuals.comparison.ps")
strat.fit3 <- coxreg(Surv(next.ivl, event) ~ strata(ses) + age + year + prev.ivl
                  + parish, data=fert1, max.survs=10)
par(mar=c(7,4,0,1),oma=c(2,1,0,0),mfrow=c(2,2),col.axis="white",
    col.lab="white",col.sub="white",col="white", bg="slategray")
hist(residuals(strat.fit3,type="martingale"),main="",
    xlab="Martingale Residuals",col="gold2",border="white")
hist(residuals(strat.fit3,type="deviance"),main="",
    xlab="Deviance Residuals",col="gold2",border="white")
hist(residuals(strat.fit3,type="score"),main="",
    xlab="Score Residuals",col="gold2",border="white")
hist(residuals(strat.fit3,type="schoenfeld",weighted=TRUE),main="",
    xlab="Schoenfeld Residuals",col="gold2",border="white")
dev.off()
```

Analyzing Residuals, Comparison