Some Topics in Survival Analysis

Padua University, June 2014

Donald A. Pierce

Oregon Health Sciences University

Course materials at
www.science.oregonstate.edu/~piercedo/Padua_Course/

Some useful (highly selected) references:

Kalbfleisch and Prentice, *The Statistical Analysis of Failure Time Data*, 2nd Ed, Wiley

Therneau and Grambsch, *Modeling Survival Data: Extending the Cox Model*, Springer

Andersen, Borgan, Gill and Keiding, *Statistical Models Based on Counting Processes*, Springer-Verlag

Breslow and Day, *Statistical Methods in Cancer Research Vol II, The Design and Analysis of Cohort Studies*, Int'l Agency for Research on Cancer, Lyon

Fleming and Harrington, *Counting Processes and Survival Data*, Wiley

Hosmer and Lemeshow, *Applied Survival Analysis: Regression Modeling of Time-to_Event Data*, Wiley

One further reference

Klein and Moeschberger, *Survival Analysis: Techniques for censored and truncated data*. Springer

Motivating example, illustrating several important points --- being theme of course

A Danish population of 1500 diabetics was followed up for about 10 years, in a given calendar period. One interesting issue is the relation between diagnosis age and subsequent age-specific death rates

It is problematic to estimate the probability distribution of age-at-death from these data, since subjects are of different ages starting follow-up

Towards that aim, it is crucial to understand that the information on each person is conditional on his being alive at the start of follow-up

That they survived to this age is not "follow-up" information, since had they not, they would not have been in the study

However, for analysis of rates rather than "response times" none of this is problematic as rates do not depend on the time origin

In STATA one can start to investigate this as follows

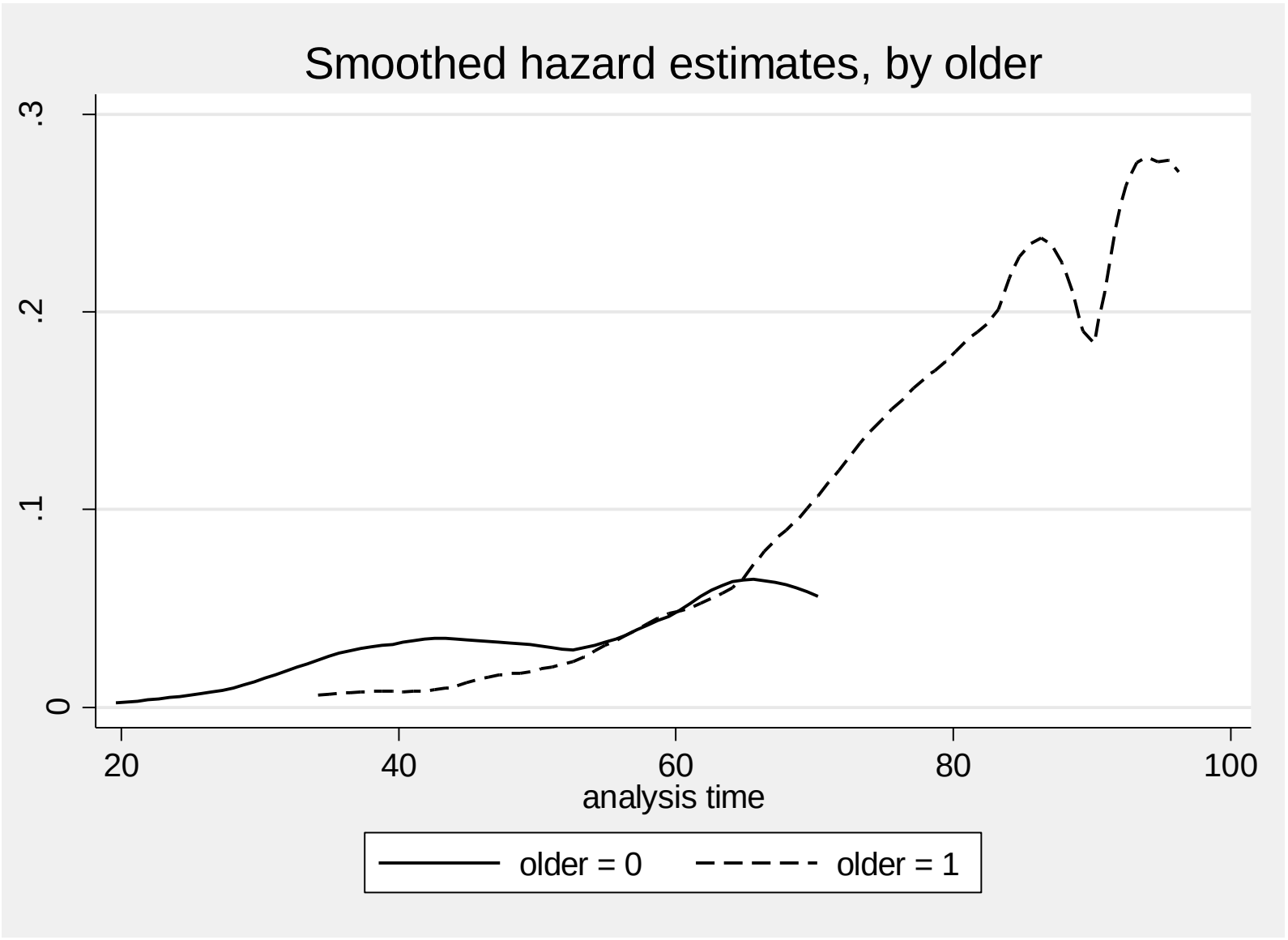
```
use diabetes
gen exitage = entryage + futime/365.25
stset exitage , failure(status==1) enter(time entryage)
```

Standard Cox regression (p. 22 here), showing that those over 20 at diagnosis have considerably lower death rate during the follow-up. Note that the "intercept" is aliased with the baseline hazard.

```
gen older = dxage >20
stcox older
 Log likelihood  =   -2324.0326
         _t   |   Haz. Ratio   Std. Err.      z    P>|z|
 -------------------------------------------------------
       older |    .606          .102       -2.97   0.003
 -------------------------------------------------------
```

There is much more to any reasonable analysis, and I recommend starting with

```
sts graph, by(older) hazard
```


Smoothed hazard estimates, by older
.3
.2
.1
0
20
40
60
80
100
analysis time
older = 0
older = 1

Continuing with the theory now

The main reason survival analysis is so special is not the usual nonnormality, or the censoring, but that rates are so important

For continuous time the failure rate or hazard function is

$$\lambda(t) = \lim h \downarrow 0 \; pr(t \le T < t+h \,|\, T \ge t)/h$$
$$= \frac{f(t)}{S(t)} = \frac{d}{dt}\left[-\log\{S(t)\}\right]$$

where $f(t)$ is the density and $S(t) = pr(T > t)$ is the survival function

Hence we have the important relation

$$S(t) = \exp\left\{-\int_0^t \lambda(u)du\right\} = \exp\{-\Lambda(t)\}$$

where $\Lambda(t) = \int_0^t \lambda(u)du$ is the cumulative hazard

More generally, and importantly,

$$P(T > t \mid T > s) = \exp\{\int_s^t \lambda(u)du\}$$

Of course this can be written as $\exp\{\Lambda(t) - \Lambda(s)\}$, but inferentially one may estimate the conditional probability without the unconditional ones.

It is often the case that from given data one can estimate only the conditional probabilities as above, and then any presentation involving $\Lambda(t)$ or the unconditional survival function $S(t)$ will be misleading.

The melanoma data is a good example of this.

Most texts start with failure times T, defining from this the rate or hazard function $\lambda(t)$ as on slide 6. To an extent I will take the rate as the more fundamental object for inference, and will try to show you why I do this.

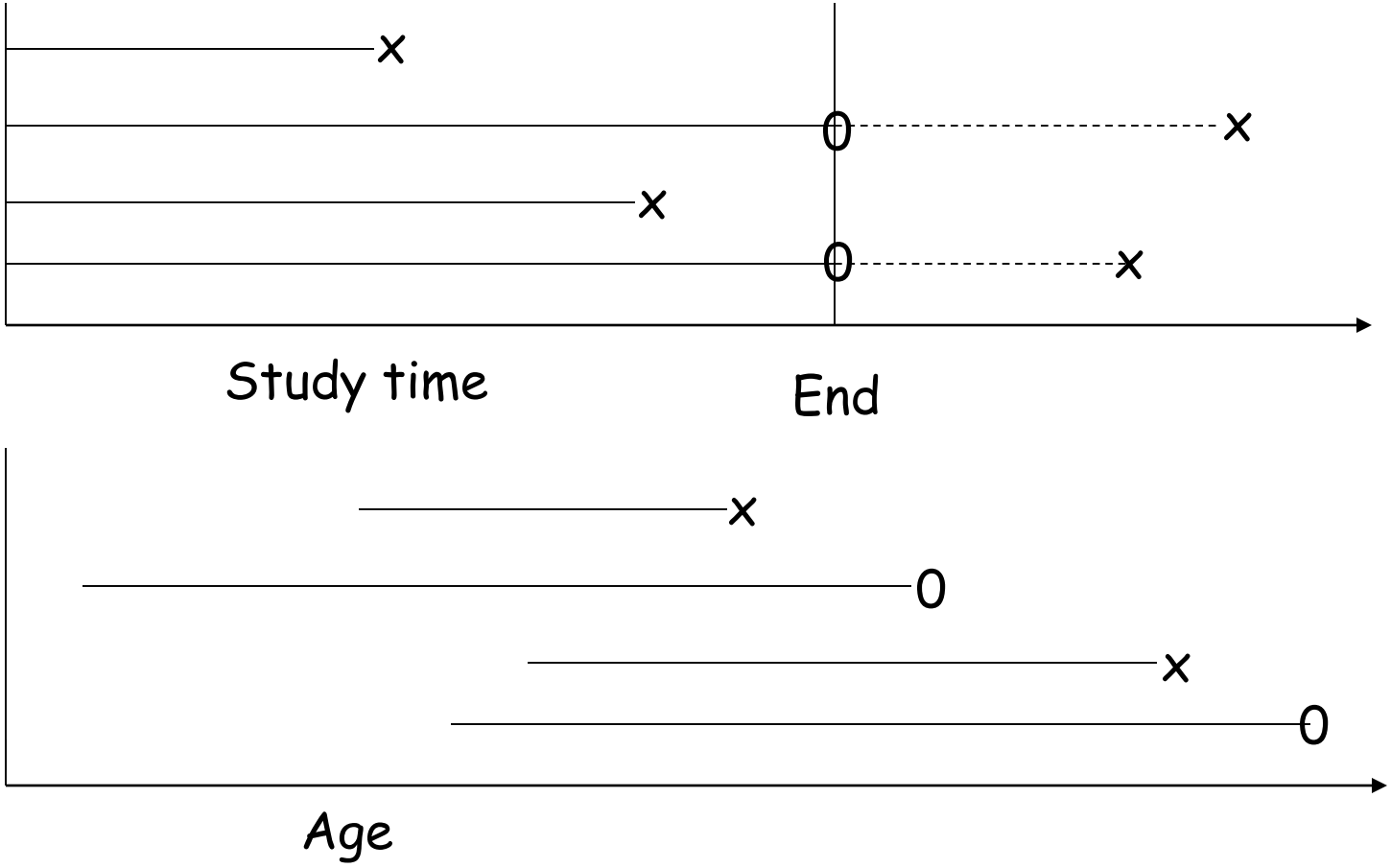
Also, one should as early as possible focus on the dependence of these on covariables. In applied statistics one is seldom concerned with i.i.d. observations, but some kind of comparisons.

A key aspect of survival data is censoring. That is, some or most individuals are not observed until they fail, but only until they are for some reason removed from observation.

The most common model for this is that for each individual there are two latent endpoints: failure time T and censoring time W, with the observation being the smaller of these along with identification of which it is. This model is not without problems, but suffices for many needs.

Censoring (right) and left-truncation (delayed entry)

Consider a study of people who enter at different ages



The two most commonly-used "time scales" in medical studies are: (a) time from diagnosis, treatment, etc. and (b) age of subject

The most common reason for censoring is the end of study, although subjects are often "lost" for other reasons

In terms of time scales (a) and (b) even censoring at the end of study leads to different censoring "times" for individuals, since they usually enter the study at different calendar times, or at different ages.

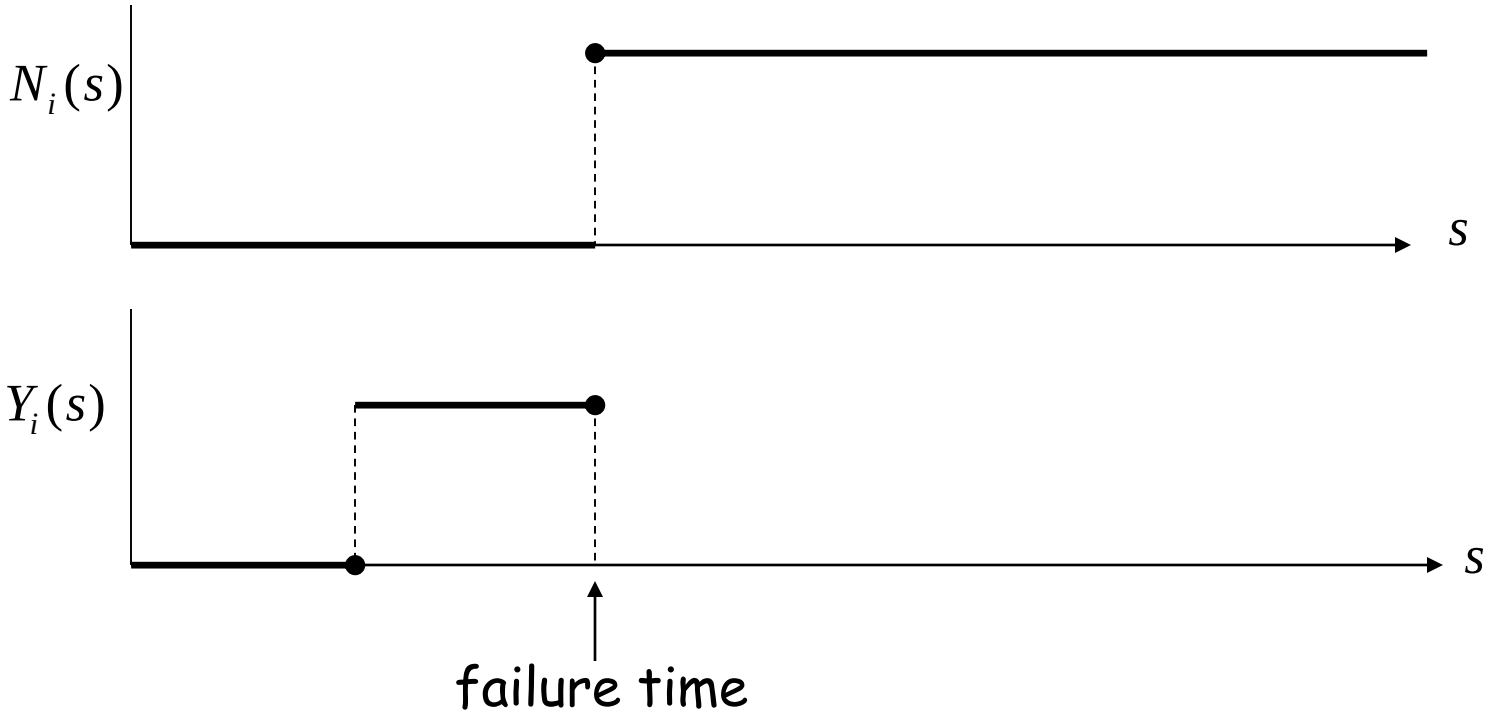
Note that these "time scales" really differ in regard to the origin, not a nonlinear sense, and we shall see that rates are not much affected, since they are so totally conditional

Simple methods require that censoring is "uninformative", e.g. that a censoring is not "indicative" of an impending failure, Kalb & Prent Sect. 3.2

Elements of counting process, for case where each subject can respond only once. Consider first only the i^{th} individual

$N_i(s)$: 0-1 indicator of failure in [0,s] --- right continuous

$Y_i(s)$: 0-1 indicator of being at risk at (s-) --- left continuous



Kaplan-Meier estimator: This is motivated by considering some fixed time intervals and using the argument considered above, namely that

$$\hat{S}(t) = pr(\text{surv } 1st \text{ interval}) \times pr(\text{surv } 2nd \mid \text{surv } 1st) \times \cdots$$

That is,

$$\hat{S}(t) = \prod_{\text{intervals up to } t} \left\{ 1 - \frac{\text{number of failures in interval}}{\text{number at risk starting the interval}} \right\}$$

As the intervals become arbitrarily narrow, this becomes the "product limit" estimator

$$\hat{S}(t) = \prod_{j:t_j \le t} \left\{ 1 - \frac{dN_{\cdot}(t_j)}{Y_{\cdot}(t_j)} \right\}$$

Where $N_{\cdot}(s) = \sum_{i=1}^{n} N_i(s)$ etc.

Understand that $dN_{\cdot}(t_j)$ is the number of failures at time t_j, and $Y_{\cdot}(t_j)$ is the number at risk just before that time

Nelson–Aalen estimator: Thinking again of time intervals, for each interval the estimate of failure rate is the number of failures in the interval divided by the number at risk at the start of the interval

As the intervals become arbitrarily narrow, we have that

$$\hat{\lambda}(t) = \frac{\sum_{i=1}^{n} dN_i(t)}{\sum_{i=1}^{n} Y_i(t)} = \frac{dN_\cdot(t)}{Y_\cdot(t)}$$

This is nonzero only at failure times, and at those values it is (with no ties) 1 over the number at risk just before the failure. As with densities, it usually must be smoothed to be useful.

The Nelson-Aalen estimator of the cumulative or integrated hazard sums these estimates over failures up to time t

$$\hat{\Lambda}(t) = \int_0^t \frac{1}{Y_\cdot(s)} dN_\cdot(s)$$

These both pertain to discrete distributions, requiring the following alterations.

For discrete distributions the hazard is defined as

$$\lambda(t) = \frac{pr(T = t)}{pr(T \geq t)} = \frac{p(t)}{S(t-)}$$

Then at the k^{th} ordered value of the random variable

$$\begin{aligned} S(t_k) &= pr(T > t_1)\ pr(T > t_2 \,|\, T > t_1) \cdots pr(T > t_k \,|\, T > t_{k-1}) \\ &= \{1-\lambda(t_1)\}\{1-\lambda(t_2)\}\cdots\{1-\lambda(t_k)\} \\ &= \prod_1^k \{1-\lambda(t_j)\} \end{aligned}$$

and in connection with the previous relation $S(t) = \exp\{-\lambda(u)du\}$

$$S(t_k) = \prod_{j\leq k}\{1-\lambda(t_j)\} \doteq \prod_{j\leq k}\exp\{-\lambda(t_j)\} = \exp\left\{-\sum_{j\leq k}\lambda(t_k)\right\}$$

Many think of the K-M and N-A estimators as alternatives, but in fact they provide exactly the same discrete distribution and thus are just different representations of the same thing

The correct relationship is the one for discrete distributions

$$\hat{S}_{KM}(t_{(i)}) = \prod_{j \le i} \left\{ 1 - d\hat{\Lambda}_{NA}(t_{(j)}) \right\}$$

It is true, though, then when thinking of these as estimators for continuous distributions, they do provide different estimates since then one expects to have

$$\hat{S}(t) = \exp\left\{ -\hat{\Lambda}(t) \right\}$$

but in fact

$$\hat{S}_{KM}(t) \neq \exp\left\{ -\hat{\Lambda}_{NA}(t) \right\}$$

They are very nearly equal, though, for time ranges where the jumps of the N-A estimator are small

Variances of these estimators

Greenwood's formula for K-M estimator:

$$\operatorname{var}\left\{-\log \hat{S}_{KM}(t)\right\} \doteq \int_0^t \frac{1}{Y_\cdot(s)\left\{Y_\cdot(s) - dN_\cdot(s)\right\}} dN_\cdot(s)$$

Martingale-based estimator for N-A estimator

$$\operatorname{var}\left\{\hat{\Lambda}_{NA}(t)\right\} \doteq \int_0^t \frac{1}{\left\{Y_\cdot(s)\right\}^2} dN_\cdot(s)$$

Distinction is essentially that of Binomial and Poisson distributions. The latter is based on the exact relation from the martingale theory

$$\operatorname{var}\left\{\hat{\Lambda}(t)\right\} = E\left\{\int_0^t \frac{1}{Y_\cdot(s)} \lambda(s) ds\right\}$$

Validity of the estimators requires that a censoring is not "indicative" of an impending failure, Kalb & Prent Sect. 3.2; perhaps main issue is "uninformative" censoring

What is referred to as "independent censoring" does not literally mean censoring times are stochastically independent of failure times, but something along the lines above

Under these conditions the <u>likelihood function</u> is the product of terms

$$L_i = \begin{Bmatrix} f(t_i) & \text{for failures} \\ S(t_i) & \text{for censored obsvns} \end{Bmatrix}$$

There are also important issues regarding "delayed entry" and "competing risks" that will be emphasized throughout these lectures

For staggered delayed entry as shown on the earlier slide there are problems. The contribution to survival estimation from each individual is valid only when conditioning on their being "alive" at age of entry. When the ages are staggered the overall survival function estimator has little or no meaning.

Note that $pr\{T < t \mid T > s\} = \exp\{-\int_s^t \lambda(u)du\}$ so the conditioning changes the lower limit from zero. Thus issues related to the integrated (cumulative) hazard require modification for such conditioning

It is important that the rate estimate $\hat{\lambda}(t)$ presents no problems at all, and the problem is with interpretation of the integrated rate. In many situations the desired inferences pertain to rates much more than to integrated rates

A related issue arises when censoring is due to competing risks. Again, this causes no problem with rate estimates, but must be accounted for for integrated rates and survival

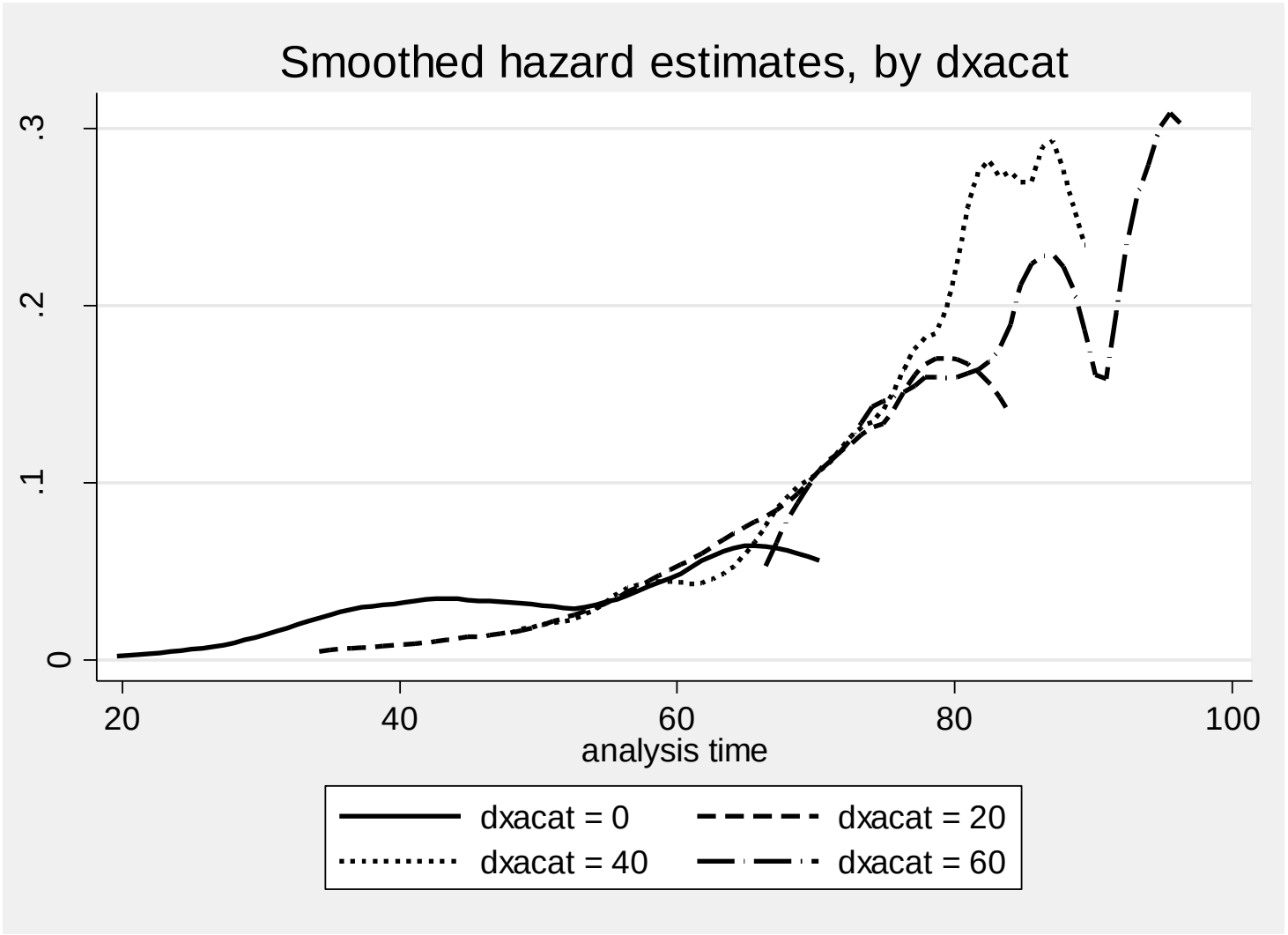
A Danish collection of diabetics was followed up for about 10 years, in a given calendar period

Although their age, and age at diagnosis were determined, there was effectively "no follow-up" until the study began

One interesting issue is the relation between diagnosis age and subsequent age-specific death rates

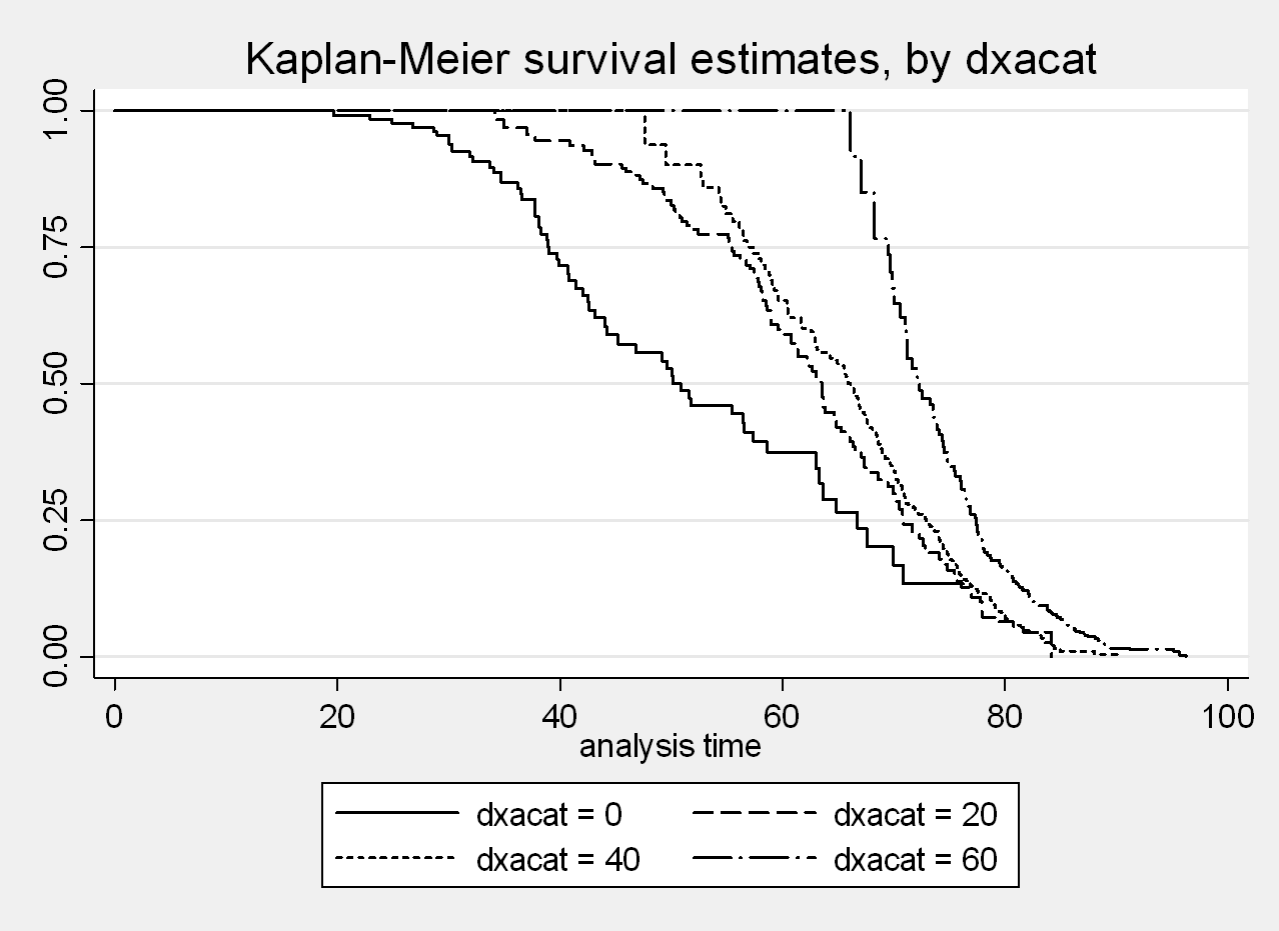
In STATA one can start to investigate this as follows

```
use diabetes
egen dxacat = cut(dxage) , at(0,20,40,60,120)
gen exitage = entryage + futime/365.25
stset exitage , failure(status==1) enter(time entryage)
sts graph, by(dxacat) hazard
```



Note that "analysis time" here is age

Note that the survival estimates below are totally useless. The contribution from each person is conditional on their being alive at entry age (not closely related to diagnosis age)



For data such as these, it is important to keep in mind that although information at a broad range of ages is available, that from each person comes from an age range of only about 10 years

This is common in medical studies, whenever it is important to consider a primary time scale as age as opposed to time since the study was started. However, the most important time scale for clinical studies is often time since diagnosis.

The general point is that none of this (choice of the origin for time scales) interferes seriously with inferences about rates but it must be carefully taken into account for inferences about survival or cumulative rates

Some useful further analysis using relative risk (Cox) regression

Recall that for this, the hazard for individual i is represented in terms of covariables z_i as

$$\lambda_i(t;\beta) = \lambda_0(t)\exp(z_i'\beta) = \lambda_0(t)RR(z_i;\beta)$$

where $\lambda_0(t)$ is left as totally unspecified

For groups of data $j = 0, 1, \ldots, G-1$ this reduces to

$$\lambda_j(t)/\lambda_0(t) = \exp(\beta_j) \quad \text{for all} \quad t \; ; j = 1, \cdots, G-1$$

Since it will not be true that the ratio is exactly independent of t, the parameter estimates are best thought of as estimating an *average* over t

Some further analysis is *always* required regarding the extent to which such averages are useful

First, some basics of Cox regression: consider a 2-sample problem

For each of the ordered failure times $t_{(j)}$ consider a Bernoulli observation $y_j = 0,1$ indicating whether the failure is in sample *0* or sample *1* , with associated probability

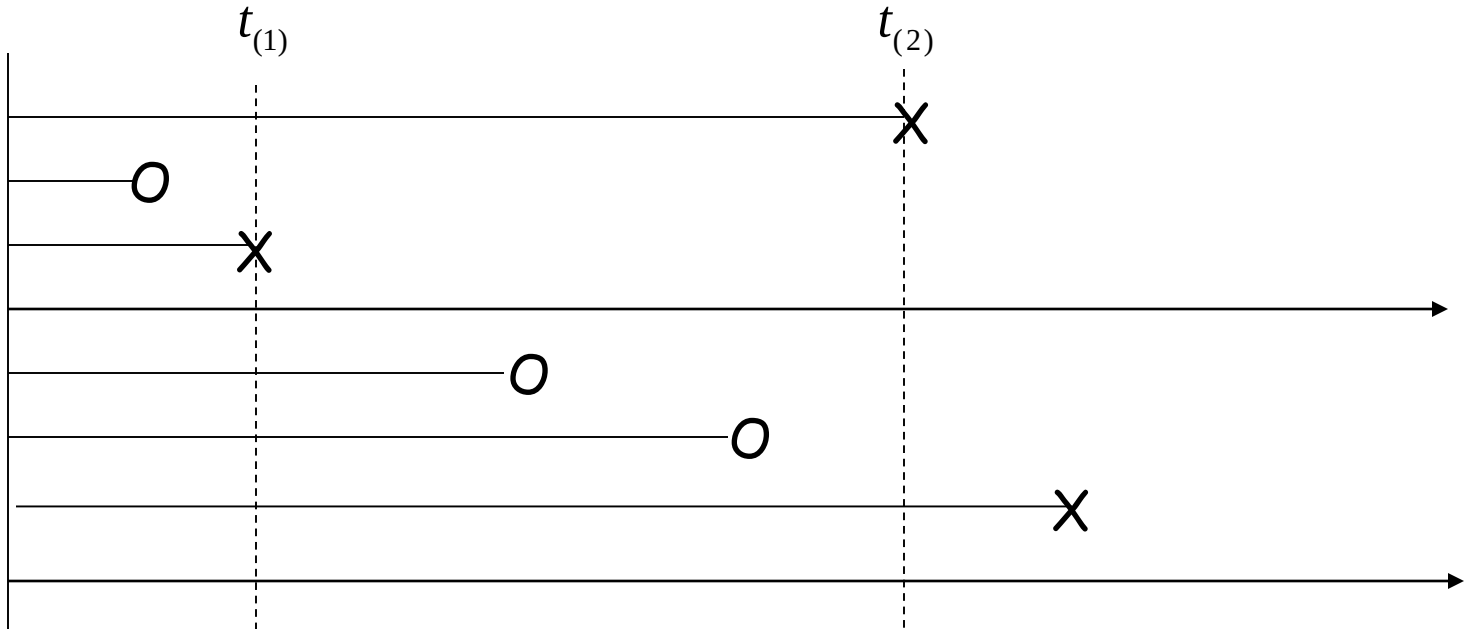
$$p_j = \frac{n_{1j} e^{\beta}}{n_{0j} + n_{1j} e^{\beta}}$$

where n_{0j}, n_{1j} are the numbers currently at risk from each sample

The Cox regression estimator is then the MLE in this binomial regression, pooling strata defined by failure times

For more general covariates, including several, this generalizes directly by replacing the binomial distribution with a multinomial for sample size unity

COX REGRESSION: Hazards of form $\lambda(t;z,\beta)=\mu(t)e^{z\beta}$, with $\mu(\cdot)$ unspecified. Interest parameter a scalar function of β with remaining coordinates as nuisance parameters



Risk set $\mathbf{R}_i$: those alive at failure time $t_{(i)}$

Multinomial likelihood contribution $L_i(\beta)=e^{z_{(i)}\beta}/\sum_{j\varepsilon\mathbf{R}_i}e^{z_j\beta}$, the probability that it is individual (i) among these that fails.

Partial likelihood $L(\beta)=\Pi_i L_i(\beta)$

In general, for covariate vectors z_i , the estimation "balances", over all the failure-time risk sets, the chance of obtaining the observed covariate value $z_{(j)}$ compared to the average z-value of those who might have failed at that time, using weights proportional to $\exp(z_k\beta)$

In particular, the vector of estimating equations is of form

$$\sum_{\text{Risk sets } R_j} \left\{ z_{(j)} - E_{\hat{\beta}}(z:\ over\ z_k\text{'s } in\ R_j \right\} = 0$$

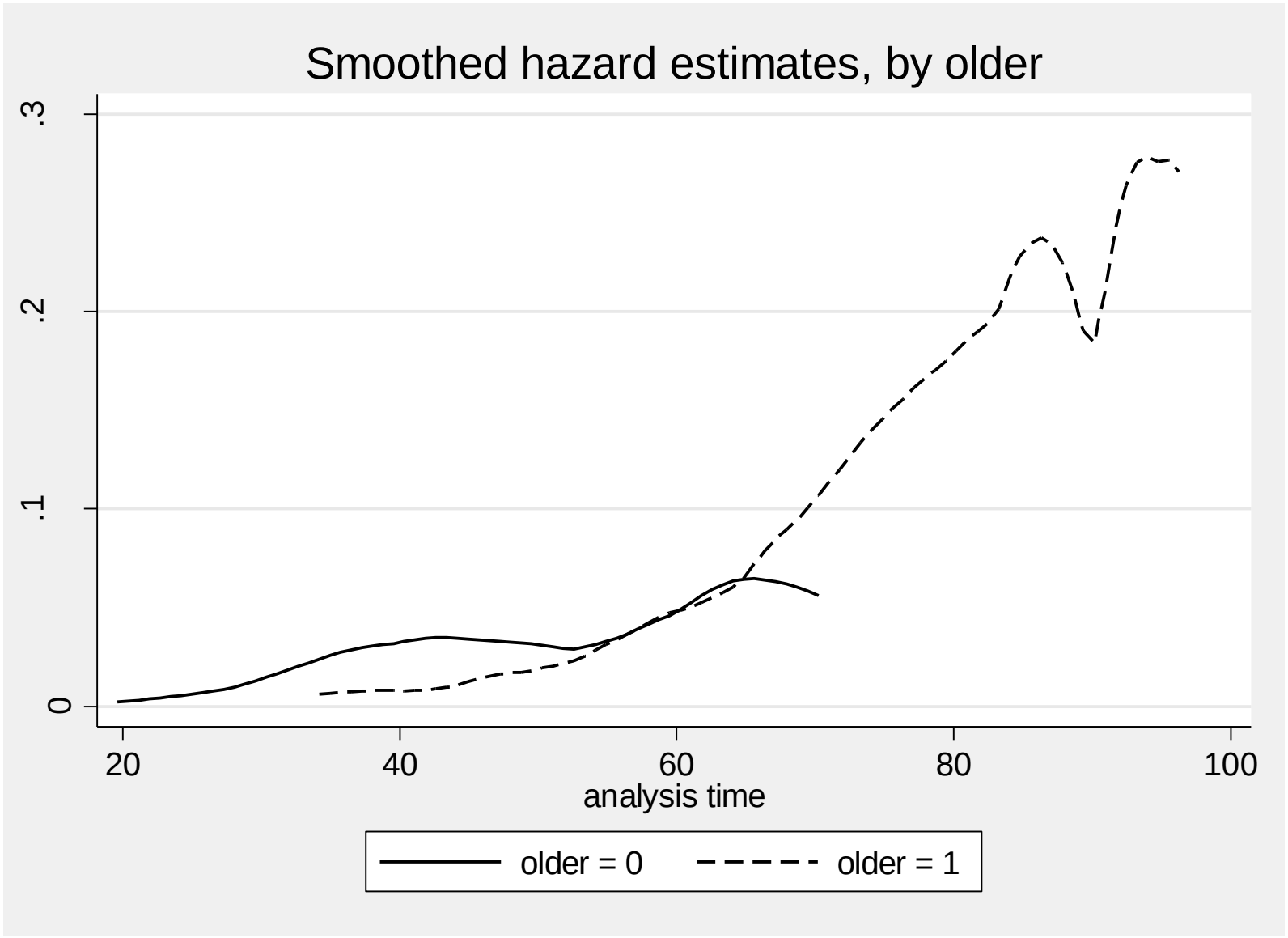
where the expectation is taken with weights proportional to $\exp(z_k\beta)$. Note that nonlinear transformations of the time scale would have no effect on this

Through comparison of covariate values for failures, rather than failures themselves, this has a retrospective flavor that has close connections with case-control studies

Application to previous example: first with 4 groups and then 2

```
xi: stcox i.dxacat
Log likelihood  =   -2323.5229
          _t |     Haz. Ratio  Std. Err.   z      P>|z|
-------------------------------------------------------------
 _Idxacat_20 |     .591        .105       -2.94     0.003
 _Idxacat_40 |     .633        .120       -2.42     0.016
 _Idxacat_60 |     .552        .124       -2.64     0.008
-------------------------------------------------------------
estimates store fullmod
gen older = dxage >20
stcox older
 Log likelihood  =   -2324.0326
          _t |   Haz. Ratio   Std. Err.      z    P>|z|
 ------------------------------------------------------------
       older |       .606        .102       -2.97   0.003
 ------------------------------------------------------------
lrtest fullmod .
  likelihood-ratio test                 LR chi2(2)  =      1.02
 (Assumption: . nested in fullmod)    Prob > chi2 =    0.6006
```

The under 20 group has greater hazard than the others, and there may not be much difference within the others


Smoothed hazard estimates, by older
.3
.2
.1
0
20
40
60
80
100
analysis time
older = 0
older = 1

Although the hazard ratio, averaged over age, is significantly different from unity, we must consider whether this summary is adequate

Following are two ways to do this:

1. Allowing the ratio to vary on some specified intervals of age
2. Investigating a smooth age trend in the RR

The first method uses a "splitting" of the follow-up into intervals of the time scale, which is important in both S-Plus (R) and Stata (called the "counting process form" of Cox regression)

```
stsplit agecat , at(0,30,40,50,55,60,80)
      (913 observations (episodes) created)
gen cov1 = older * (agecat==30)
gen cov2 = older * (agecat==40)
    ...
gen cov6 = older * (agecat==80)
```

For example, subject *id = 15* had *entryage* = 79.98, *exitage* = 81.65, and died at that age

This command splits this into 2 records, the first with *_t0 = 79.98* and *_t = 80* and the second with *_t0 = 80* and *_t = 81.56*
(*entryage* and *exitage* are not changed, but in analysis *_t0* and *_t* created by *stset* or *stsplit* are always used for this purpose)

The first of these is labeled as censored and the second one as a death

cov5 = 1 for the first and *cov6 = 1* for the second

Now Cox regression can be used to compute a hazard ratio for each of the age intervals

```
stcox cov1-cov6
------------------------------
   _t | Haz. Ratio   Std. Err.
------+-----------------------
 cov1 |    .190       .104
 cov2 |    .440       .163
 cov3 |    .602       .321
 cov4 |   1.116       .533
 cov5 |   1.00        .364
 cov6 |   20.07178          .
------------------------------
estimates store tvarmod
quietly, stcox older
lrtest tvarmod .
  LR chi2(4)  =     10.12  Prob > chi2 =    0.0385
```

This shows that the hazard ratio is far from constant

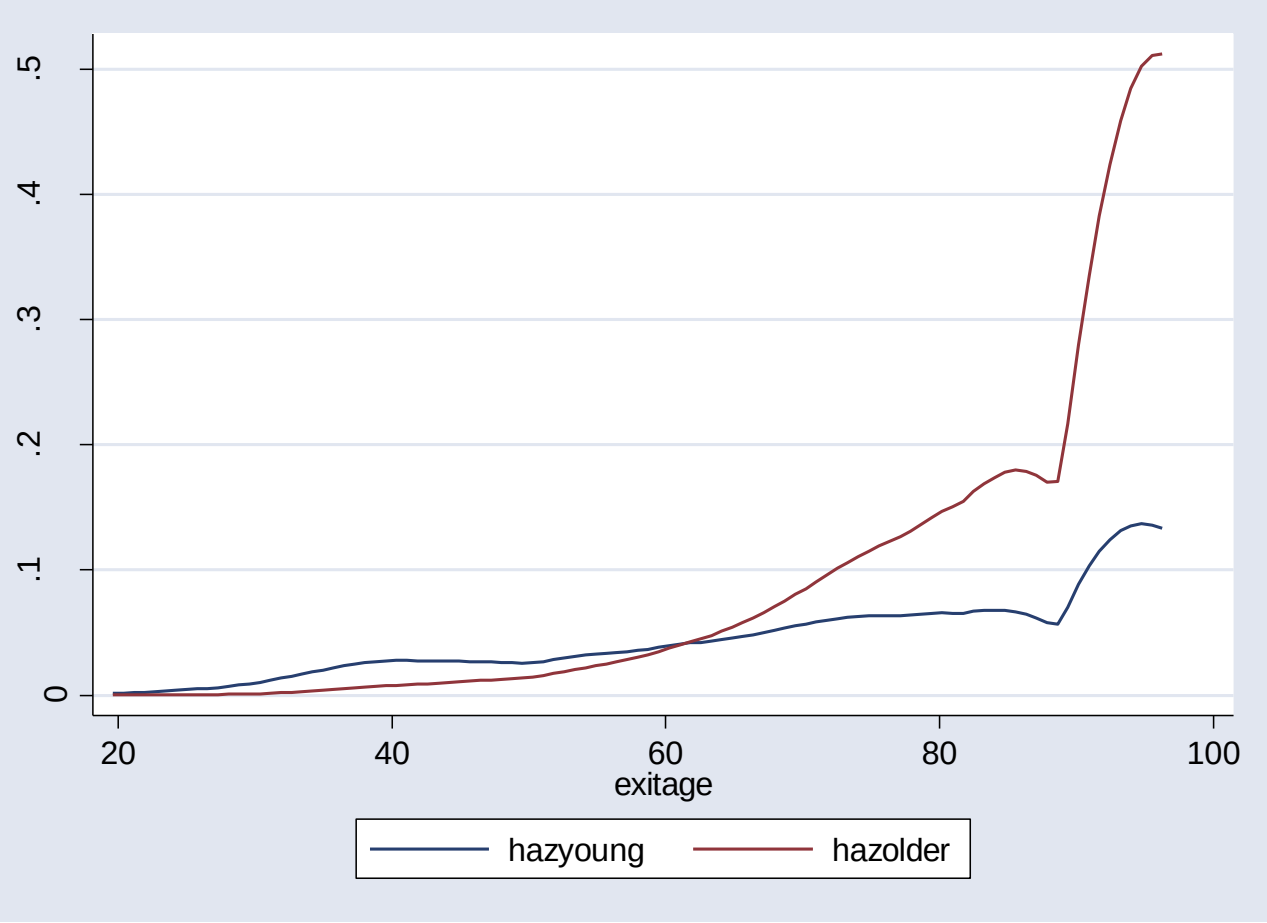
The following investigates, more simply, a trend in the RR using covariate *older * log(age/55)*

```
use diabetes, clear
gen exitage = entryage + futime/365.25
stset exitage , failure(status==1) enter(time entryage)
id(id)
gen older = dxage > 20
stcox older , tvc(older) texp(log(exitage/55))  nohr
------------------------------------------------------
          _t     |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
rh             |
       older     |  -.3217053   .2076046    -1.55   0.121
-------------+----------------------------------------
_t             |
       older     |    2.98408    .971825     3.07   0.002
------------------------------------------------------
```

This fits the model with $RR = \beta_0 \,\mathrm{cov} + \beta_1 \,\mathrm{cov} \times \log(age/55)$

Should not take this model fit very seriously --- why?

Fitted values: compare to earlier plot of the two separate hazards. Although the model is useful in testing for a trend, it should not be used in an "extrapolative" manner

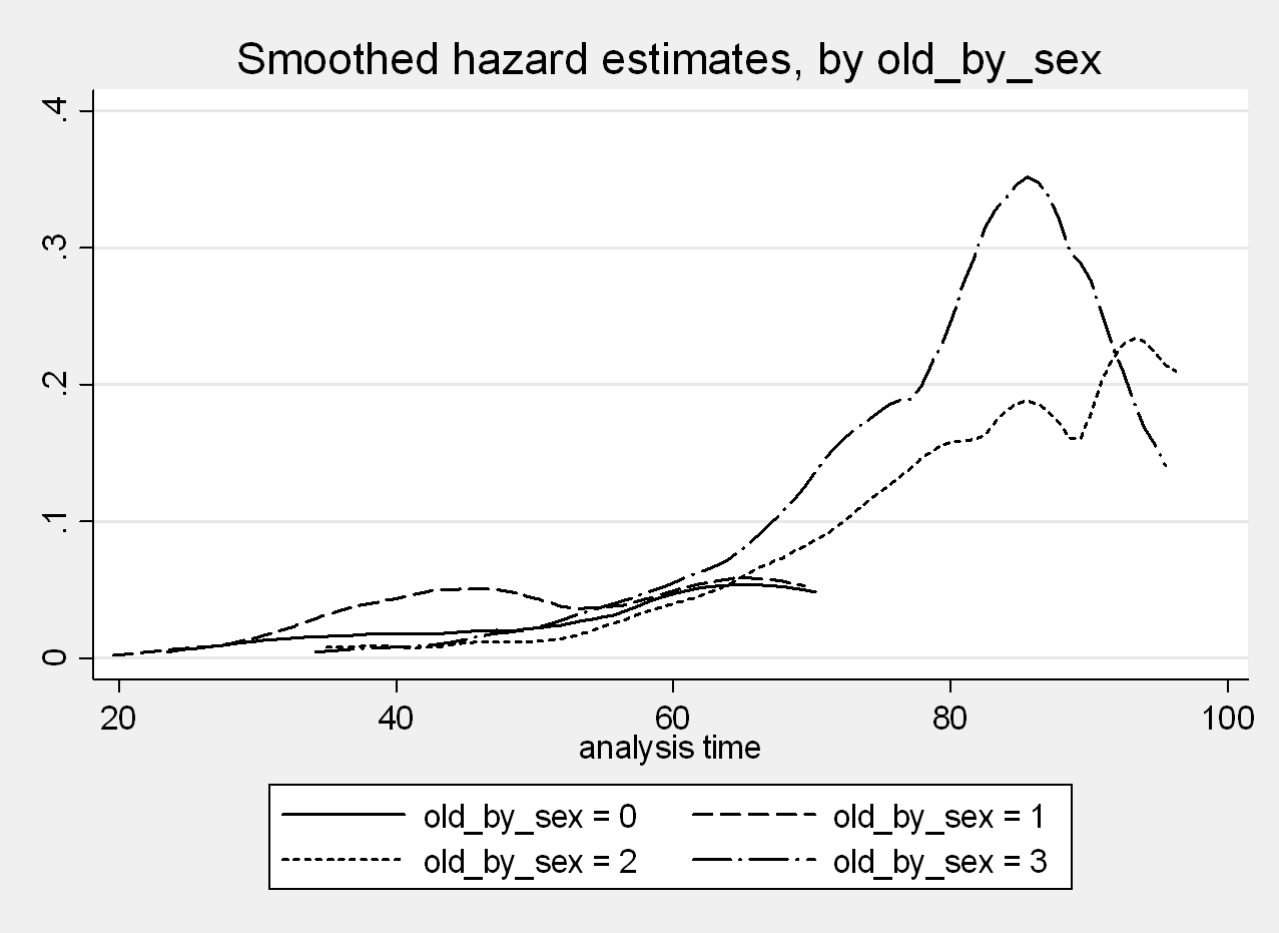


Before proceeding we should note that there is a highly significant sex effect as well. In terms of average hazard ratios over age this can be analyzed as follows --- large sex effect about the same for the two diagnosis-age groups

```
xi: stcox i.older i.sex

Log likelihood  =   -2314.6367
------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|
-------------+----------------------------------------
   _Iolder_1 |   .6008939    .100909    -3.03   0.002
     _Isex_1 |   1.497661   .1395761     4.33   0.000
------------------------------------------------------

xi: stcox i.older*i.sex

Log likelihood  =   -2314.6193
------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|
-------------+----------------------------------------
   _Iolder_1 |   .6210895   .1524749    -1.94   0.052
     _Isex_1 |   1.569123   .4198654     1.68   0.092
_IoldXsex_~1 |   .9482587   .2708617    -0.19   0.852
------------------------------------------------------
```

A plot without modeling as proportional hazards

```
gen old_by_sex = 2*older + sex
sts graph , by(old_by_sex) hazard
```



There is another way of approaching this, which is often preferable. First, though, note that it was not necessary to use the "xi" prefix since the categorical variables were binary

```
stcox older sex

Log likelihood  =   -2314.6367
------------------------------------
   _t | Haz. Ratio   Std. Err.        z      P>|z|
------------+---------------------
older |   .6008939    .100909    -3.03   0.002
  sex |   1.497661   .1395761     4.33   0.000
------------------------------------------------
gen older_by_sex = older*sex
stcox older sex older_by_sex

Log likelihood  =   -2314.6193

-------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|
-------------+-----------------------------------------
       older |   .6210895   .1524749    -1.94   0.052
         sex |   1.569123   .4198654     1.68   0.092
older_by_sex |   .9482587   .2708617    -0.19   0.852
-------------------------------------------------------
```

The alternative way is to stratify on sex, meaning that the baseline hazard is allowed to differ arbitrarily by sex, whereas the RR parameters are common to strata.

```
stcox older , strata(sex)

Log likelihood  =   -1989.5718

------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|
-------------+----------------------------------------------
       older |   .5952816   .1002846    -3.08   0.002
------------------------------------------------------------
       Stratified by sex
```

The advantage of this is that it is no longer assumed that the baseline hazards for the sexes are proportional in age. The main disadvantage is that one cannot readily see, as before, how large is the sex effect in baseline death rates.

Most applications of Cox regression do use stratification in some manner

It is still possible to see, parametrically, whether the RR parameter for older differs between the sexes

```
stcox older older_by_sex, strata(sex)

Log likelihood  =   -1989.4957

------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|
-------------+----------------------------------
       older |   .6500353   .1844718    -1.52   0.129
older_by_sex |   .8718167   .3072994    -0.39   0.697
------------------------------------------------
       Stratified by sex
```

In the example just considered, the "time scale" for Cox regression was taken as age. It certainly would not be useful to take the time scale as time-since-study-entry, and taking it as time-since-diagnosis would not be ideal.

As noted earlier the choice of "time scale" for such purposes is not really a matter of "scale" but a matter of the origin used. This choice does not really affect rates, but affects the modeling of them. In particular, which age-time aspects are taken as the primary time scale, and which are used as covariates.

In clinical studies it is common, and usually best, to take the primary time scale as time-since-diagnosis, or time-since-treatment.

Thus I will introduce such an example, leaving most of the analysis for it as an exercise in Set 1.

From some bureaucratic administrative data (often not as good as medical-study data) we have for 1612 women the time from diagnosis of cervical cancer until death or end of study. Of primary interest is the relation of death rates to whether each participated in a PAP screening program.

Although there is considerable range of age-at-diagnosis, which is important since "death" is taken as to any cause, the most important time scale is probably time from diagnosis to death. This is often the case for clinical studies.

With that choice, one must seriously consider using diagnosis age as a covariable. There is also a variable representing the stage of the cancer at diagnosis.

Results of Cox regression are rather complicated, and perhaps unexpected. In particular, even within stage there is a screening effect, whereas one might expect screening mainly to result in earlier detection.

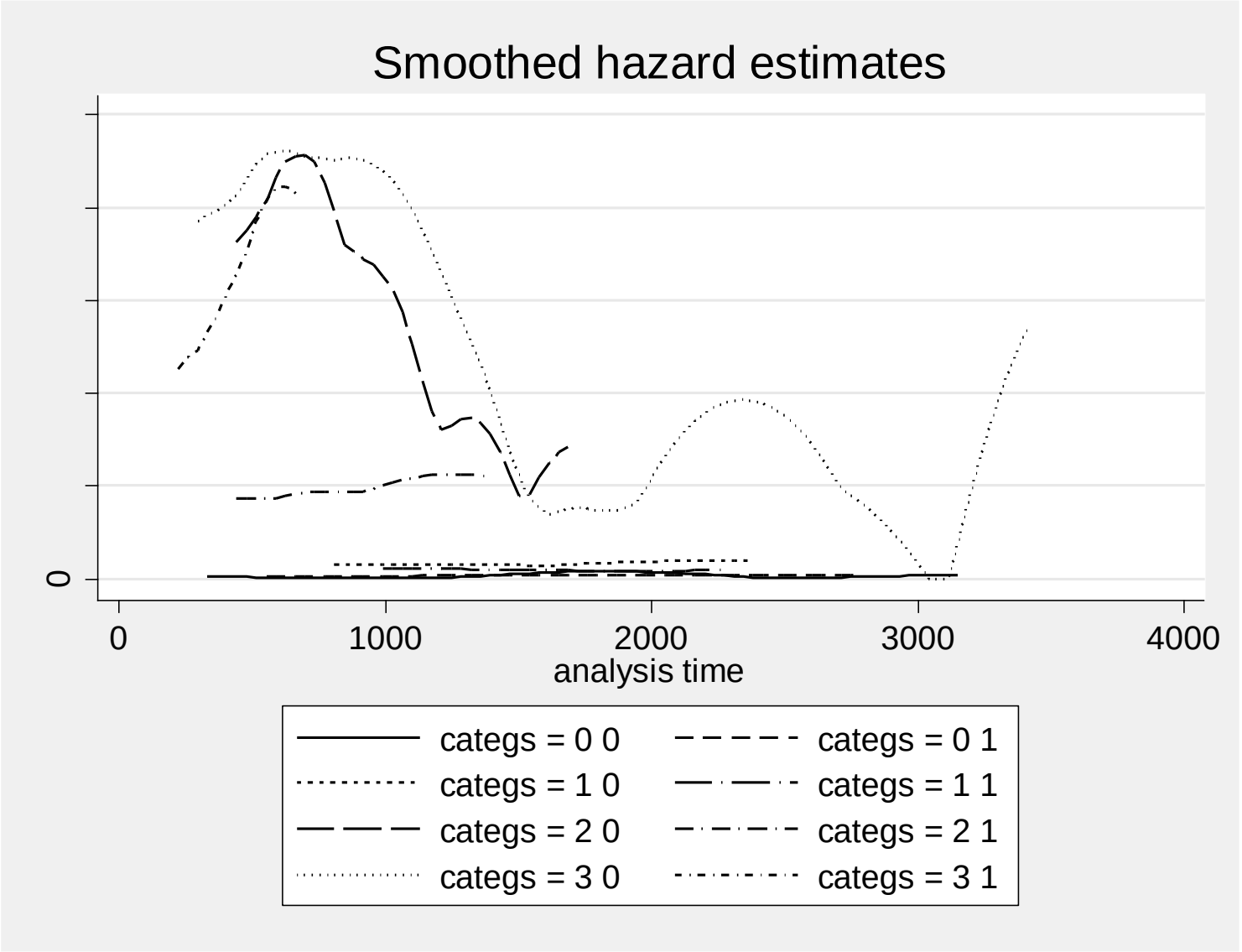
All that I will demonstrate here is some graphical methods without which the Cox regression is hopelessly confusing.

This consists of plotting death rates, as earlier here, for categories defined by stage and screening. A main point is that since the Cox regression models rates, rather than cumulative rates or survival curves, it is most helpful to plot the (smoothed) estimates of the hazard function itself.

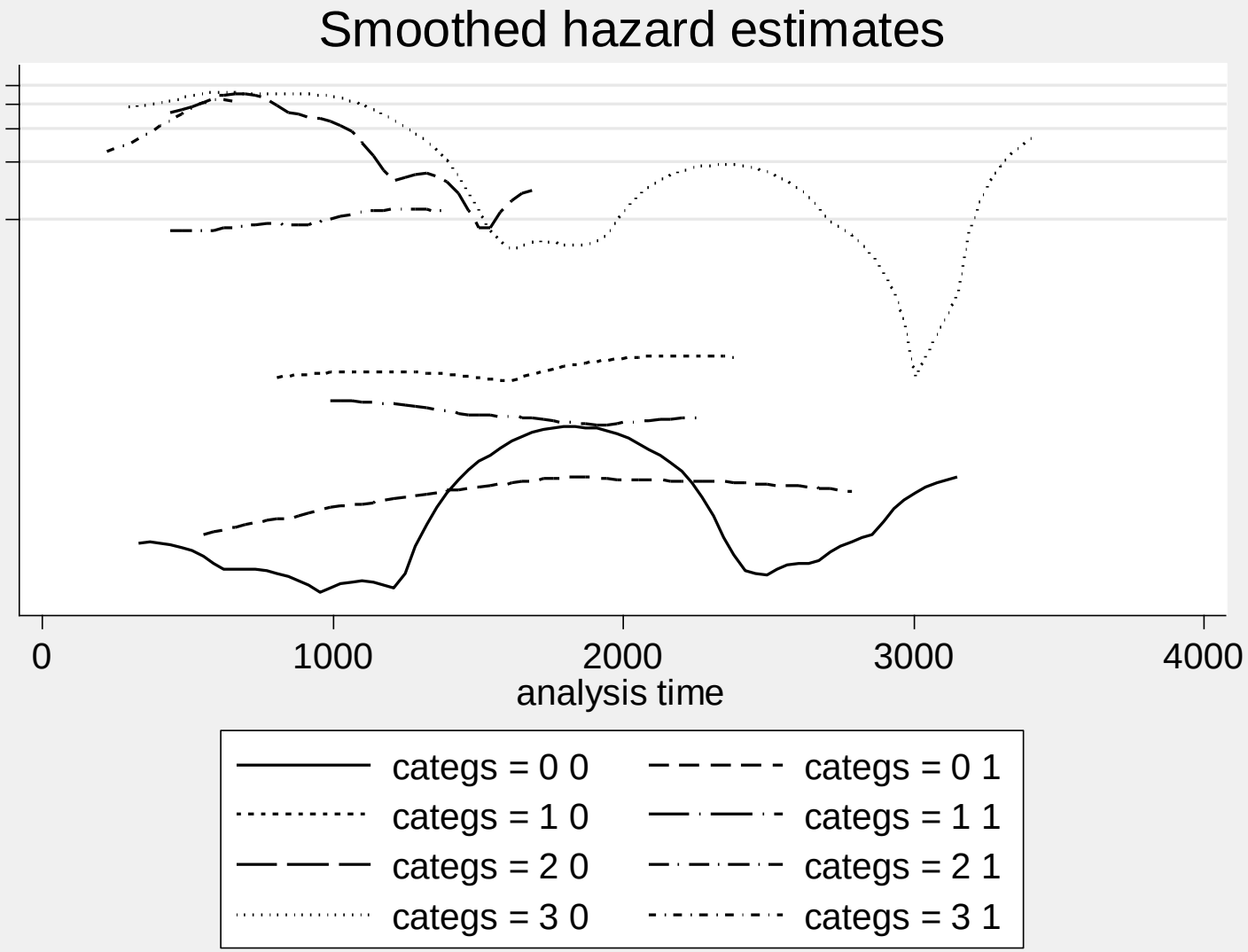
Variables are: age_diag, exit_time, dead, screened {0,1}, stage {0,1,2,3}

```
stset exit_time , fail(dead)
egen categs = group(stage screened) , label
sts graph , by(categs) hazard
```

Hard to read here, but note that some rates are very small



sts graph , by(categs) hazard yscale(log)



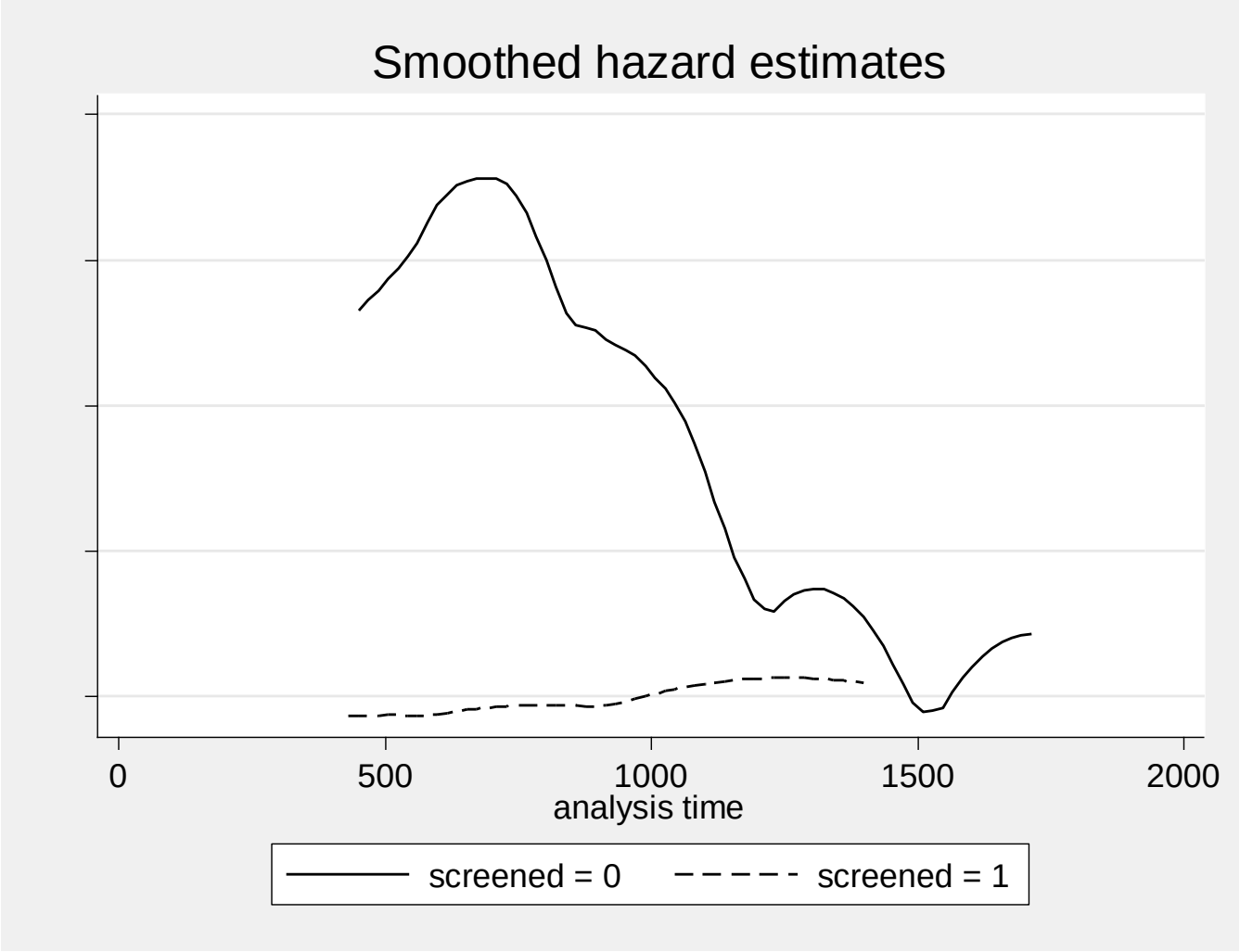
Part of the need for this type of plotting is to look into the proportional hazards assumption for the Cox regression model on p. 22.

This was considered for the diabetes example on pp. 28-32, but I want to consider it further here, as many users have a difficult time with this issue in practice.

In the exercise on this cervical cancer data, you will need to consider this for the Cox regressions. A main issue there will be the extent to which the hazards for the two levels of screening are proportional, for each diagnostic stage.

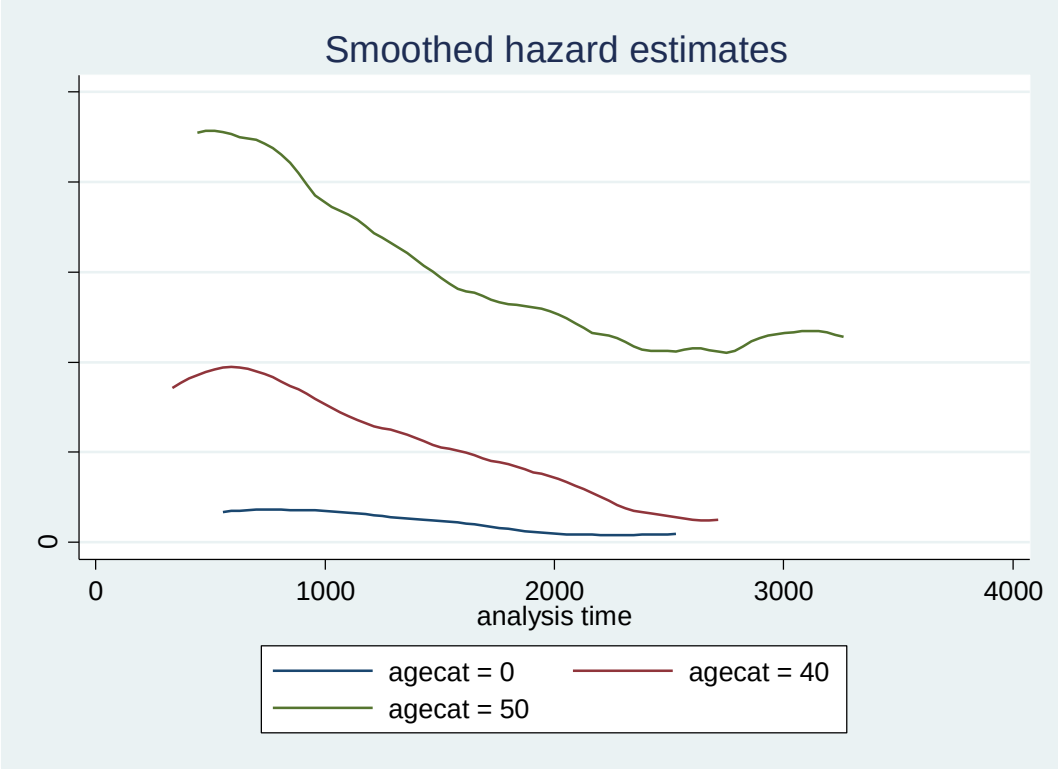
The following plot shows more clearly than above what results for diagnostic stage 2.

```
preserve
keep if stage==2
sts graph, by(screened) hazard
restore
```



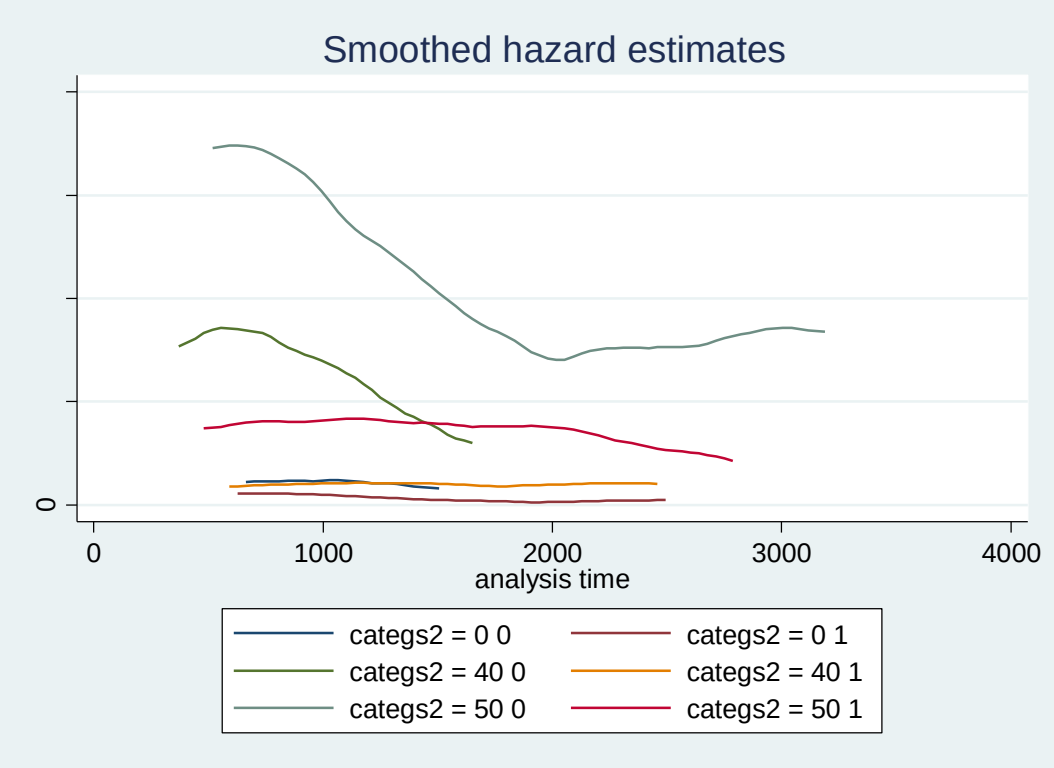
Age at diagnosis is likely to be an important covariate.

egen agecat=cut(age_diag) , at(0,40,50,100)
sts graph , by(agecat) hazard



Can we see anything this way about the joint effect of diagnosis age and screening?

egen categs2 = group(agecat screened) , label
sts graph , by(categs2) hazard



Some general issues about proportional hazards are:

1. Stata provides for tests and diagnostics regarding the assumption, but these are less useful than the following.
2. Hazard plots along lines here are usually far more useful.
3. It is less that proportional hazards are strictly required for Cox regression, than that estimates will correspond to the average hazard ratios over time --- sometimes this is quite adequate for the needs. This is likely the case for the previous page, and for other stages as well.
4. If time-averages of hazard ratios are not adequate for the need, then one can with Cox regression model the time-variation of hazard ratios using time-dependent covariables --- the proportional hazards assumptions is not strictly required for Cox regression

More on Time-Dependent Covariables: Above we used time-dependent covariables in a formal way to investigate the proportional hazards assumption. Often they are used in more fundamental ways, modeling changes in risk with follow-up that are of particular interest.

Often-used example: Stanford Heart Transplant Data

Some patients are randomized to medicine therapy --- those randomized to transplant are under observation while they await the transplant --- the "transplant" covariate for them changes at the time of transplant

Variables are:

stime: time from treatment to death or censoring
transplant: binary indicator of transplant
wait: time until transplant
age: age at start of study
year: year of entry to study
surgery: binary indicator of *prior* surgery

```
use stanford
replace wait=wait - 0.1 if stime==wait   * remove some ties
stset stime , fail(died) id(id)
stsplit aftertrans , after(wait) at(0)   * the main point here
replace aftertrans = aftertrans+1 if transplant==1  * fix labeling
```

e.g. patient 4 with *wait*=36, *stime=39, died=1* now has 2 records: one with follow-up for the first 36 days, censored at that time, and another with follow-up for 3 more days, with event "*died*" at that time

```
stcox age aftertrans surgery
Log likelihood  =   -291.13858
------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z
-------------+----------------------------------
         age |   1.033254   .0144661     2.34
  aftertrans |   1.002663   .3063415     0.01
     surgery |   .3305263   .1420819    -2.58
------------------------------------------------
```

Transplant patients have same subsequent death rate as those with medicine treatment

One should always consider effects of age

```
egen agecat = cut(age) , at(0,40,50,100)
keep if surgery==0     * just to simplify things
xi: stcox I.agecat*aftertrans

Log likelihood  =   -260.11109
------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z
-------------+----------------------------------
 _Iagecat_40 |   1.090803   .5234069     0.18
 _Iagecat_50 |   1.880217   .9320849     1.27
  aftertrans |   .4841006   .3400513    -1.03
_IageXaft~40 |   2.201334   1.738274     1.00
_IageXaft~50 |   2.399345   1.902204     1.10
------------------------------------------------
```

Young patients seem to benefit from transplant, while older ones seem to do worse with transplant than with medicine

From STATA 11 new handling of factor variables

Output clearer, making me realize I was interpreting previous fit wrongly --- figure out why the difference

```
stcox aftertrans#i.agecat

        Log likelihood  =   -260.11109

------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z
-------------+----------------------------------
  aftertrans#|
      agecat |
       0 40  |   1.090803   .5234069     0.18
       0 50  |   1.880217   .9320849     1.27
       1  0  |   .4841006   .3400513    -1.03
       1 40  |   1.162433   .5442203     0.32
       1 50  |   2.183918    .991644     1.72
------------------------------------------------
```

Are the doctors getting better with time? On the contrary

```
egen yearcat = cut(year) , at(66,69.5,71.5,75)
xi: stcox I.yearcat*aftertrans

Log likelihood  =    -262.9827
------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z
-------------+----------------------------------
 _Iyearcat_2 |   1.054558   .4384916     0.13
 _Iyearcat_3 |   .4521415   .2211328    -1.62
  aftertrans |   .9660618   .3947484    -0.08
_IyeaXafte~2 |   1.180939   .6831352     0.29
_IyeaXafte~3 |   1.721368   1.071051     0.87
------------------------------------------------
```

However, this could be due to confounding with age of subjects,but it seems from the following not to be.

```
tabulate agecat yearcat    * no remarkable correlation
           |              yearcat
    agecat |        66       69.5       71.5 |
-----------+---------------------------------+
         0 |         8          9          12 |
        40 |        20         20          26 |
        50 |        27          8          18 |
-----------+---------------------------------+
```

Again, stratification can be important here. Takes underlying hazard as unspecified within while covariables have common effect over strata

In the age analysis above we took the age effect on the underlying hazard as simply 3 multiplicative parameters --- stratification relaxes that assumption enormously

```
xi: stcox I.agecat|aftertrans , strata(agecat)

Log likelihood  =   -192.57567
------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z
-------------+----------------------------------
  aftertrans |   .6568869   .4945401    -0.56
_IageXaft~40 |   1.432418   1.277949     0.40
_IageXaft~50 |   1.874898    1.69942     0.69
------------------------------------------------
```

Age interaction is somewhat smaller this way

The following provides stronger evidence (although probably still not significant) that the year effect is not just confounding with age

```
xi: stcox I.yearcat*aftertrans , strata(agecat)

Log likelihood  =   -188.73802
------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z
-------------+----------------------------------
 _Iyearcat_2 |   1.275453   .5527938     0.56
 _Iyearcat_3 |   .4267472   .2118292    -1.72
  aftertrans |    .659119   .3022043    -0.91
_IyeaXafte~2 |   1.575471   .9737808     0.74
_IyeaXafte~3 |    2.16495   1.393578     1.20
------------------------------------------------
```

Cox regression methods are called partially nonparametric, since the underlying hazard $\lambda_0(t)$ is left totally unspecified

Survival analysis can be done fully parametrically as follows. Let $f_i(t;\theta)$ be the density (before censoring), where the dependence on i would usually be dealt with through covariates.

The fully parametric likelihood for the censored data is then

$$L(\theta) = \prod_{i=1}^{n} f_i(t_i;\theta)^{\delta_i} S_i(t_i;\theta)^{1-\delta_i}$$

$$= \prod_{i=1}^{n} f_i(t_i;\theta)^{\delta_i} \exp\{-\int_{e_i}^{t_i} \lambda_i(s;\theta) ds\}^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \lambda_i(t_i;\theta)^{\delta_i} \exp\{-\int_{e_i}^{t_i} \lambda_i(s;\theta) ds\}$$

Where δ_i is the indicator of failure and e_i is the entry time

There seem to be only 3 ways of doing this in practice.

1. Fully parametric proportional hazards models

$$\lambda_i(\theta;t) = \lambda_0(\gamma;t)RR(\mathrm{z}_i,\beta)$$

where the baseline hazard $\lambda_o(\gamma;t)$ is fully-specified by parameters, and for example $RR(\mathrm{z}_i,\beta) = \exp(\mathrm{z}_i'\beta)$

2. Accelerated failure-time models, where $f_i(t;\theta)$ corresponds to taking $\log t_i = \mathrm{z}_i'\beta + \varepsilon$, where ε has density fully-specified by a parameter γ , typically a distribution with mean zero and possibly unknown variance (often normal)

3. Poisson Regression as indicated on the next slide, and taken up again later in these notes

Approaches 1 and 2 are generally better than Cox regression for "time" inferences rather than "rate" inferences

Except for very special cases evaluating the likelihood for methods 1 and 2 requires numerical integration, which Stata will carry out for a number of parametric model forms (will return to this later)

However, it is clearly possible to do this relatively simply if the hazard function $\lambda_0(t;\gamma)$ for approach 1 is replaced by a piecewise-constant approximation, defined on fixed intervals of time

It turns out that calculations for this approach can be done through what is called "Poisson regression"

Subject to the approximation made, which involves small error if the number of time intervals used is large, this provides a very flexible approach to survival analysis

Further, the essence of Cox regression can be regained in this approach, as will be shown

I will now indicate briefly some of the main issues in the martingale theory (but that for Cox regression is different)

For individuals define $N_i(s)$, $Y_i(s)$ as earlier, and let $\lambda_i(s)$ be the hazard functions

The (mean zero) counting process martingale is then

$$M(t) = N_{\cdot}(t) - \sum_{i=1}^{n} \left\{ \int_0^t Y_i(s)\lambda_i(s)ds \right\}$$

where the second term is called the compensator of $N_{\cdot}(t)$

The martingale defining property can be stated as that

$$E\{dM(t) \mid \text{history of process up to } t-\} = 0$$

The two most important derived properties are that

(a) $$\text{var}\{dM(t) \mid \text{history to } t-\} = \sum_{i=1}^{n} Y_i(t)\lambda_i(t)dt$$

and hence $$\text{var}\{dM(t)\} = E\left\{ \sum_{i=1}^{n} Y_i(t)\lambda_i(t)dt \right\}$$

(b) Increments of $M(t)$ over disjoint intervals are uncorrelated

Property (a) holds only for continuous time processes, and means that the process is "Poisson like", with variance equal to mean (there is analogous "binomial like" theory for discrete time processes). Note: (a) can be re-expressed as

$$\begin{aligned}\mathrm{var}\{dM(t) \mid \text{history to } t-\} &= \mathrm{var}\{dN_{\bullet}(t) \mid \text{history to } t-\} \\ &= E\{dN_{\bullet}(t) \mid \text{history to } t-\}\end{aligned}$$

With the extension on the next slide, this property is very useful for finding variances of various statistics

Property (b) means that for deriving limit theorems one can take the uncorrelated "contributions" to the loglikelihood as being from successive failures, rather than from individuals (contributions from individuals are *not* uncorrelated for Cox regression)

This was used by several authors (Greenwood, Cox) before the martingale theory became popular

Let $H(t)$ be a function that is nonrandom, given the history up to $t-$, (referred to as "predictable"), and

$$L(t) = \int_0^t H(s)dM(s)$$

Then since $dM(s_1)$, $dM(s_2)$ are uncorrelated,

$$\mathrm{var}\{L(t)\} = \int_0^t H^2(s)\,\mathrm{var}\{dM(s)\}ds = E\int_0^t H^2(s)Y_\bullet(s)\lambda(s)ds$$

For the N-A estimator

$$\hat{\Lambda}(t) - \Lambda(t) = \int_0^t \frac{1}{Y_\bullet(s)}\{dN_\bullet(s) - Y_\bullet(s)\lambda(s)ds\} = \int_0^t \frac{1}{Y_\bullet(s)}dM(s)$$

so

$$\mathrm{var}\{\hat{\Lambda}(t)\} = E\int_0^t \frac{1}{Y_\bullet^2(s)}Y_\bullet(s)\lambda(s)ds = E\int_0^t \frac{1}{Y_\bullet(s)}\lambda(s)ds$$

which can be estimated by

$$\int_0^t \frac{1}{Y_\bullet^2(s)}dN_\bullet(s)$$

Likelihood scores and martingales (fully-parametric models)

The contributions to the loglikelihood are of form

$$l_i = f_i \log \lambda_i(t_i) - \Lambda_i(t_i) \qquad f_i \text{ indicator of failure}$$

$$= \int_0^\infty \{\log \lambda_i(s) dN_i(s) - Y_i(s)\lambda_i(s) ds\}$$

Contributions to the score are then

$$\partial l_i / \partial \theta = \int_0^\infty \frac{\partial}{\partial \theta} \log \lambda_i(s;\theta) \{dN_i(s) - Y_i(s)\lambda_i(s;\theta) ds\}$$

which are of form $\int H_i dM_i(s)$ and hence are martingales

Further, we have from above results that

$$\text{var}(\partial l_i / \partial \theta) = E \int \left\{ \frac{\partial}{\partial \theta} \log \lambda_i(s;\theta) \right\}^2 Y_i(s)\lambda_i(s;\theta) ds$$

Analogous results for Cox regression are derived very differently from this (Flemington & Harrington text)

Now returning to Poisson Regression: Consider grouping data into fixed time intervals $j = 1, \cdots, k$ so that for the i^{th} individual

$$c_{ij} = 0 - 1 \text{ indicator of failure in interval } j$$

$$t_{ij} = \text{ time at risk in interval } j$$

Consider a failure time model (with censoring) where the hazard functions $\lambda_i(t;\theta)$ are piecewise constant at $\lambda_{ij}(\theta)$ on the above time intervals

It is well known that the likelihood function under the survival data model is identical to that under a Poisson regression model

$$c_{ij} \sim Psn\left\{t_{ij}\, \lambda_{ij}(\theta)\right\}, \quad \text{for all } (i, j)$$

Often the covariate values are in groups and these data may be reduced by summing over the groups

This can be implemented either in the proportional hazards setting, or more generally. We will consider in detail here only the proportional hazards formulation

Suppose for example that one wishes to take the hazard function, before this piecewise-constant approximation, as

$$\lambda_0(t) = \theta_0 + \theta_1 \log(t) + \theta_2 \{\log(t)\}^2$$

rather than allowing it to be totally free. Using any other specified parametric smooth function would involve no essential changes to what follows

For our diabetes data this can be done in Stata as follows

```
use diabetes , clear
gen exitage = entryage + futime/365.25
gen older = dxage > 20
stset exitage , failure(status==1) enter(time entryage) id(id)
stsplit agecat , at (10(10)100)
gen died = _d
gen risktime = _t - _t0
collapse (rawsum) died risktime (mean) dxage exitage [w = risktime]
        , by(agecat older)
gen logage = log(exitage/55)
gen logage2 = logage*logage
poisson died logage logage2 older , exposure(risktime)
```

```
Log likelihood = -45.126123
------------------------------------------------------------------------------
       died |      Coef.   Std. Err.      z    P>|z|
------------+-----------------------------------------------------------------
    logage  |   3.764782   .2429372    15.50   0.000
    logage2 |    .975292   .4171822     2.34   0.019
      older |   -.502885   .1655882    -3.04   0.002
      _cons |  -3.395595   .1377553   -24.65   0.000
   risktime | (exposure)
------------------------------------------------------------------------------
```

Now we can add the time-dependent covariable used before, in a much more straightforward way --- one of the many advantages of the Poisson regression approach

```
gen older_by_age = older * logage
poisson died logage logage2 older older_by_age, exposure(risktime)

Log likelihood = -37.283704

------------------------------------------------------------------------------
             died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          logage |   1.950482   .5541518     3.52   0.000     .8643599    3.036604
         logage2 |  -1.725914    .915486    -1.89   0.059    -3.520235    .0684071
           older |  -.5267626    .189471    -2.78   0.005    -.8981199   -.1554053
     older_by_age |   3.706481   1.024495     3.62   0.000     1.698504    5.714458
           _cons |  -3.279235   .1481355   -22.14   0.000    -3.569575   -2.988895
         risktime |  (exposure)
------------------------------------------------------------------------------
```

We can also carry out the analogue of Cox regression with this approach, by allowing the baseline hazard to be totally free on the specified time intervals, that is

$$\lambda_{ij} = \lambda_{0j} \exp(z_i' \beta)$$

If the time intervals are narrow use of this profile likelihood is essentially equivalent to Cox regression; exactly so if they are narrow enough to isolate each failure

Further, if the λ_{0j} are evaluated at their MLEs for fixed β the resulting profile likelihood is the analog of the partial likelihood, agreeing exactly for narrow time intervals

For the 10-yr age intervals used above, the results can be obtained simply by the following command

```
xi: poisson died i.agecat older , exposure(risktime) nocon
Log likelihood = -124.31356
------------------------------------------------------------
        died |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------------
 _Iagecat_10 |  -6.612124          1    -6.61   0.000
 _Iagecat_20 |  -5.627886   .4473565   -12.58   0.000
 _Iagecat_30 |  -3.888499   .1970353   -19.74   0.000
 _Iagecat_40 |  -3.706132   .2133077   -17.37   0.000
 _Iagecat_50 |  -3.060482    .183424   -16.69   0.000
 _Iagecat_60 |  -2.317894   .1784671   -12.99   0.000
 _Iagecat_70 |  -1.541279   .1812111    -8.51   0.000
 _Iagecat_80 |  -1.166253   .2031198    -5.74   0.000
 _Iagecat_90 |  -1.317222   .4411247    -2.99   0.003
       older |  -.4135063   .1671058    -2.47   0.013
    risktime |  (exposure)
------------------------------------------------------------
```

Introducing the age-dependent covariable as before

```
gen older_by_age = older * log(exitage/55)
xi: poisson died i.agecat older older_by_age,
       exposure(risktime) nocon

Log likelihood = -53.199815

------------------------------------------
        died |      Coef.   Std. Err.      z
-------------+----------------------------
 _Iagecat_10 |  -6.612124          1    -6.61
 _Iagecat_20 |  -5.570925   .4472849   -12.45
 _Iagecat_30 |  -3.652769   .1984815   -18.40
 _Iagecat_40 |  -3.482641   .2239435   -15.55
 _Iagecat_50 |  -3.186232   .2072845   -15.37
 _Iagecat_60 |  -2.882984   .2896004    -9.96
 _Iagecat_70 |  -2.485169   .3923227    -6.33
 _Iagecat_80 |  -2.430896   .4970603    -4.89
 _Iagecat_90 |  -2.864424   .7045398    -4.07
       older |  -.4064962   .1921386    -2.12
older_by_age |   2.818797   .9743427     2.89
```

Cox regression and rank tests: Clearly the estimating equations for Cox regression have nothing to do with actual failure times --- only their *ordering* and the covariate values at failure times

$$\sum_{\text{Risk sets } R_j} \left\{ z_{(j)} - E_{\hat{\beta}}(z\text{: over } z_k\text{'s in } R_j \right\} = 0$$

This means that Cox regression is *invariant* to monotonic transformations of the time scale

Generally statistical methods with this character are referred to as rank methods, the most well-known being the Wilcoxon test

In Cox regression for the "two-sample" setting (one binary treatment), the score test of no treatment effect is an alternative to the Wilcoxon test called the *logrank test*

In the following example, the small difference between these is due to minor conventions

```
use "c:\aarhus data\melanoma.dta"
stset survtime , failure(status==1) id(id)
sts test ecells, logrank

 Log-rank test for equality of survivor functions
          |   Events         Events
 ecells   |  observed       expected
 --------+-------------------------
 absent   |        16          26.16
 present  |        41          30.84
 --------+-------------------------
 Total    |        57          57.00

              chi2(1) =       7.30
              Pr>chi2 =     0.0069

stcox ecells, iterate(0)

 Log likelihood  = -279.3844    LR chi2(1)      =      7.63
                                                      Prob >
chi2     =    0.0057
 ------------------------------------------------------------
           _t | Haz. Ratio   Std. Err.      z    P>|z|
 -------------+----------------------------------------------
       ecells |   2.174292   .6411342     2.63   0.008
 ------------------------------------------------------------
```

When there is no censoring, Cox regression corresponds exactly to regression based on ranks, i.e. reduction of data to

$$\left\{(r_j, z_{r_j}),\ j = 1 \cdots n\right\}$$

where for ordered response times $t_{(j)}$ we define $t_{r_j} = t_{(j)}$

In fact, the Cox partial likelihood, as a function of the data, is exactly the probability distribution of the ranks

With censoring, these results maintain under a censoring model where a given number are censored between successive failures (Type II progressive censoring)

These notions give some idea of why *residuals* are of so little value in Cox regression ---- one might suspect they would be of limited value for rank-based methods

Fully parametric models: proportional hazards and otherwise

In the formulation

$$\lambda_i(t;\beta) = \lambda_0(t)\exp(\mathrm{z}_i'\beta) = \lambda_0(t)RR(\mathrm{z}_i;\beta)$$

can specify a parametric model $\lambda_0(t;\theta)$ rather than leaving this totally free

Most common choices are exponential and Weibull, and all the packages handle this (with censoring and delayed entry)

For the Weibull $\log\lambda_0(t) = \theta_1 + \theta_2 \log t$

In the proportional hazards framework, the distributions for all *i* are also Weibull

As indicated above, if data are "grouped" by taking the underlying hazard as piecewise constant, then quite general parametric forms can be utilized (essentially as a GLM)

It might be said that the exponential distribution bears a similar relation to survival data as the normal distribution does to measurement data --- analysis using this model is often surprisingly adequate

The Weibull is a useful generalization, but only allows (partially) for hazard functions that are monotonic in time – often they are not

A useful way of testing the adequacy of the Weibull is to use the piecewise constant hazard formulation and fit models such as

$$\log \lambda_0(t) = \theta_1 + \theta_2 \log t + \theta_3 (\log t)^2$$

or regression splines with 2-3 knots

This indicates very nicely the usefulness of the Poisson regression approach

Accelerated failure time models: an alternative to proportional hazards (in addition to Cox regr with TDC)

With censoring concepts as usual, the general form takes the distributions of response times of form

$$\log t_i = z_i'\beta + \varepsilon$$

where ε is a r.v. with mean zero and possibly unknown variance: most important case is lognormal

That is, the effect of covariates on time is a *scale change*

There is some disagreement about whether this is more or less generally useful, and natural, than models in terms of hazard functions

The interpretation is likely to be problematic when there are more than one important time scale (e.g. age and time since trtmt), or when there is staggered delayed entry

To the *melanoma* data I will fit a Weibull model, using both the proportional hazards and accelerated failure time parametrizations. This is fairly confusing

In the PH formulation we can write the Weibull survival function as

$$S(t) = e^{-\lambda t^p e^{z\beta}}$$

We can re-express this as a scale parameter (i.e. AFT) model as

$$S(t) = e^{-\lambda\left[\frac{t}{e^{-z\beta/p}}\right]^p}$$

So if the RR is $e^{z\beta}$, then the scale parameter should be $e^{z\beta^*}$

where $\beta^* = -\beta / p$

```
use melanoma, clear
stset survtime , fail(status==1)
stcox ecells , nohr
------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
      ecells |   .7774218   .2949166     2.64   0.008
------------------------------------------------------
streg ecells, dist(weibull) nohr
Weibull regression -- log relative-hazard form
------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
      ecells |   .7931742   .2948689     2.69   0.007
       _cons |  -10.19779   1.050353    -9.71   0.000
-------------+----------------------------------------
       /ln_p |   .0905778    .118339     0.77   0.444
-------------+----------------------------------------
           p |   1.094807   .1295583
         1/p |   .9134032   .1080912
------------------------------------------------------
streg ecells, dist(weibull) time
Weibull regression -- accelerated failure-time form
------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
      ecells |  -.7244879   .2805435    -2.58   0.010
       _cons |   9.314693   .2849617    32.69   0.000
-------------+----------------------------------------
       /ln_p |   .0905778    .118339     0.77   0.444
-------------+----------------------------------------
           p |   1.094807   .1295583
         1/p |   .9134032   .1080912
------------------------------------------------------
```

In using the Cox model where

$$\lambda_i(t;\beta) = \lambda_0(t)\exp(z_i'\beta) = \lambda_0(t)RR(z_i;\beta)$$

with $\lambda_0(t)$ totally unspecified, the primary inference is about the *RR*. Although it is *possible* to compute a nonparametric estimate of the baseline survival time distribution, using this is usually unwise in practice.

If the aim is inference about the distribution of *survival times*, rather than the *RR*, then Cox regression is usually the wrong approach.

Using one of the parametric models just described can be useful, but considerable checking is required regarding the adequacy of the model.

The more flexible approach arising from the Poisson regression is a useful alternative to this.

Residuals in Cox regression: there are several definitions of these, and largely they are much less useful than in ordinary regression. Recommend referring to Therneau & Grambsch

Martingale residuals

$$\hat{M}_i = N_i(\infty) - \int_0^\infty Y_i(s) e^{z_i(s)\hat{\beta}} d\hat{\Lambda}_0(s)$$

with no TDC or delayed entry these are

$$\hat{M}_i = N_i - e^{z_i\hat{\beta}} \hat{\Lambda}_0(t_i)$$

with $1-\hat{M}_i$ being related to censored observations from a unit negative exponential distribution

These are useful neither for "checking error distribution" nor for plotting against fitted values

Although they do have some limited uses, matters are extremely subtle

Deviance residuals: an attempt to modify martingale residuals to make their distribution somewhat more normal. In practice this has not been very useful

Score residuals: an matrix of dimension *n* by *p* = dim(beta)
----- recall that the estimating equation has form

$$\sum_{Risk\ sets\ R_j} \left\{ z_{(j)} - E_{\hat{\beta}}(z:\ over\ z_k\ 's\ in\ R_j \right\} = 0$$

This can be expressed as

$$\sum_{i=1}^{n} \int_0^{t_i} \left\{ z_i - E_{\hat{\beta}}(z : \text{risk set at time } s) \right\} \left\{ dN_i(s) - e^{z_i \hat{\beta}} d\hat{\Lambda}_0(s) \right\} = 0$$

and the score residuals are the contributions to the left side from each individual

These are useful for influence analysis, and for computing "robust" variance estimates

Schoenfeld residuals: these are intended for investigating the proportional hazards specification. Recall the earlier example where we fit the model

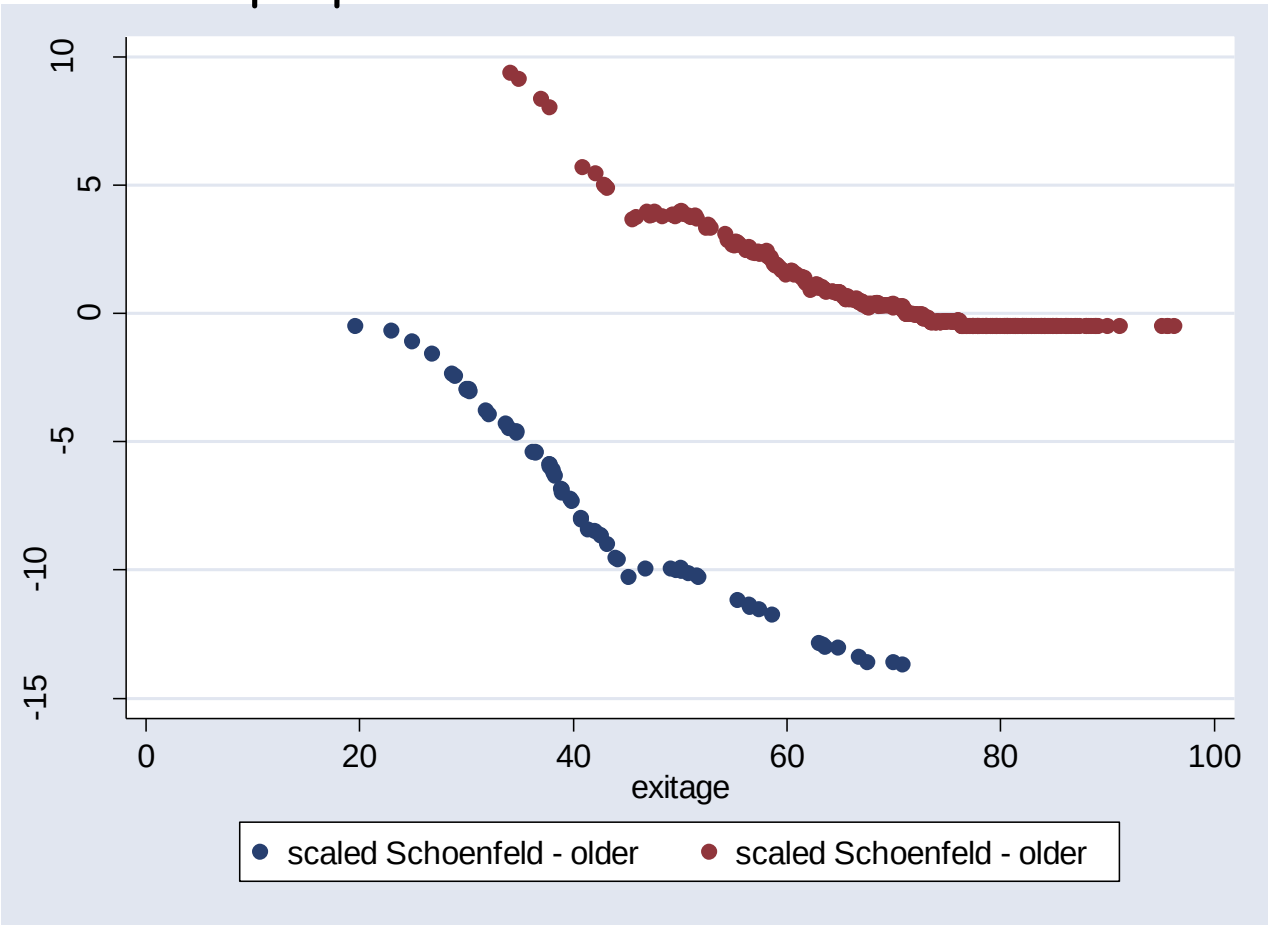
```
stcox cov1-cov6
------------------------------
   _t | Haz. Ratio   Std. Err.
------+-----------------------
 cov1 |    .190        .104
 cov2 |    .440        .163
 cov3 |    .602        .321
 cov4 |   1.116        .533
 cov5 |   1.00         .364
 cov6 |   20.07178           .
```

where *covj* was the product of the indicator of an age interval with the primary covariable *older* .

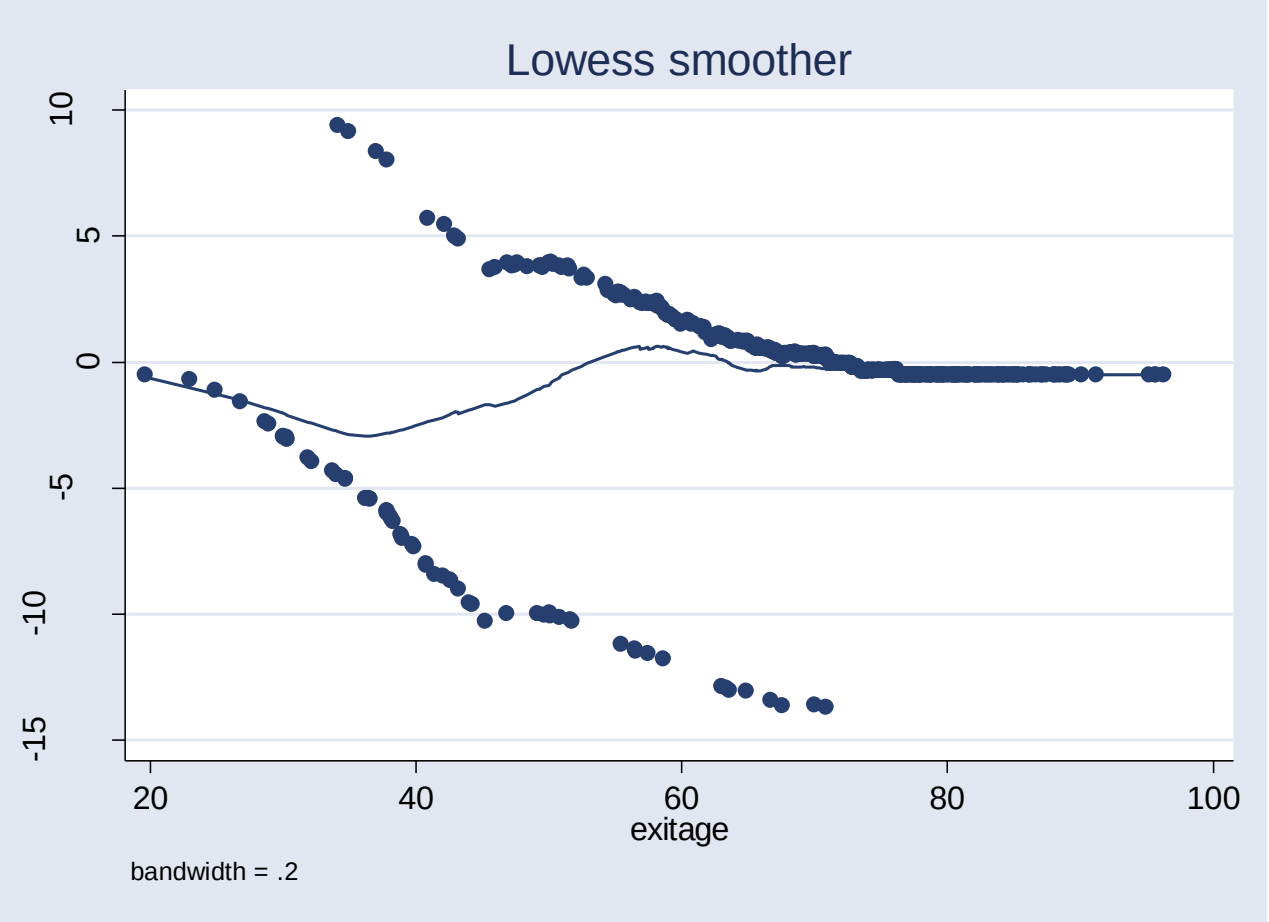
If we take the age intervals narrow enough to isolate failures, and compute the score test (of no effect) rather than fitting the model, the results are the vector of Schoenfeld residuals

This is useful, but then what was done above may often be easier to understand and explain

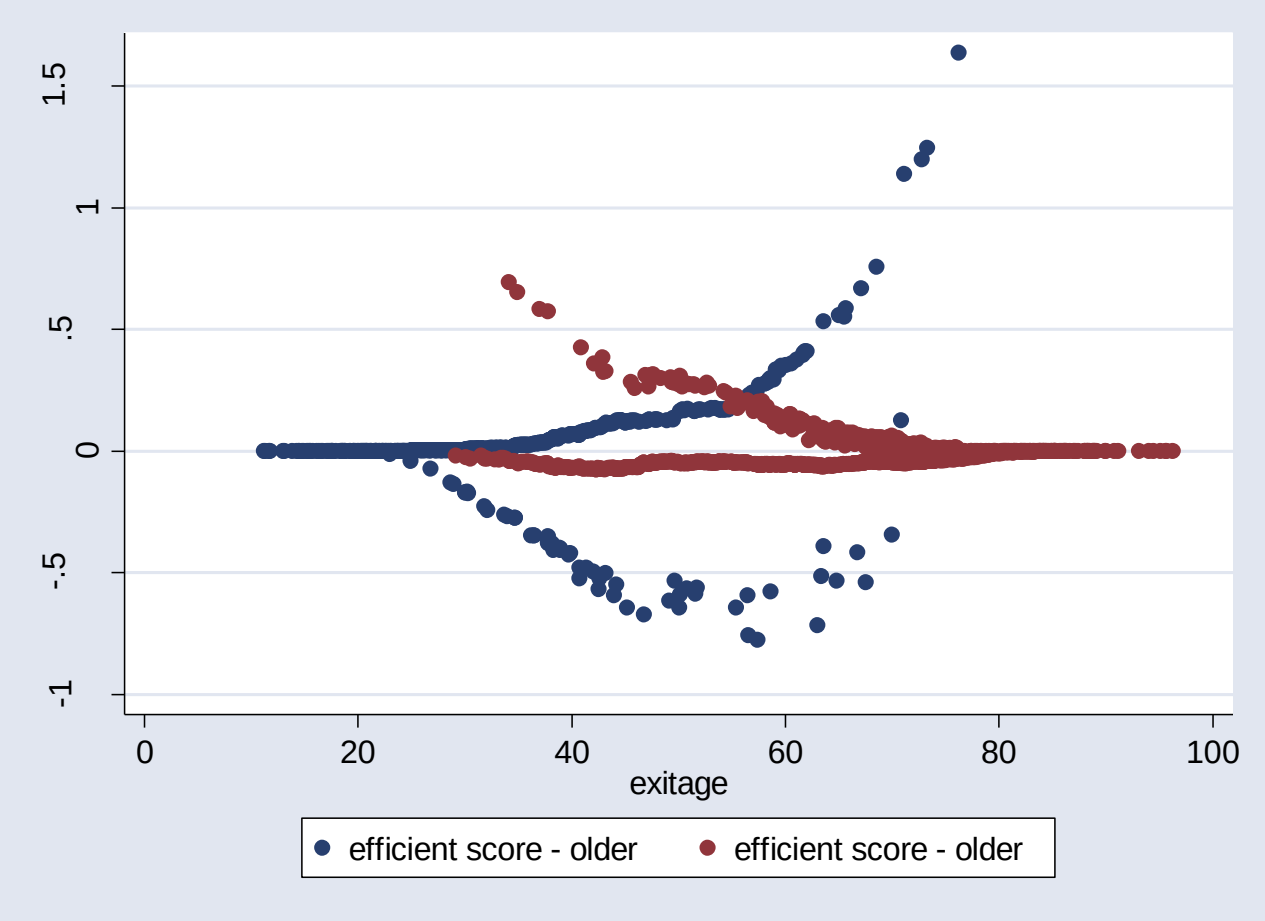
For the *diabetes* data, these are Schoenfeld residuals colored by the two levels of *older*. The point is that as *exitage* increases, points are first more dense for *older = 0* , gradually then becoming more dense for *older = 1* . This indicates, although none to clearly, departure from proportional hazards



Perhaps the best way to interpret a plot of this nature is to fit a "smooth" to the entire dataset. This indeed shows what was said about the shifting density from one set to the other.



These are the score residuals for the same dataset. I am not certain what all this means, but in fact the few largest values correspond to persons with a young diagnosis age, and who were censored at a fairly old age. They are not fitting the model so well.



Recently, I have found that "influence" diagnostics can be particularly useful in Cox regression. A type often referred to as DFBETAs was added to Stata in version 11. It turns out that these are essentially the same as the score residuals, so you can do this with Stata 10.

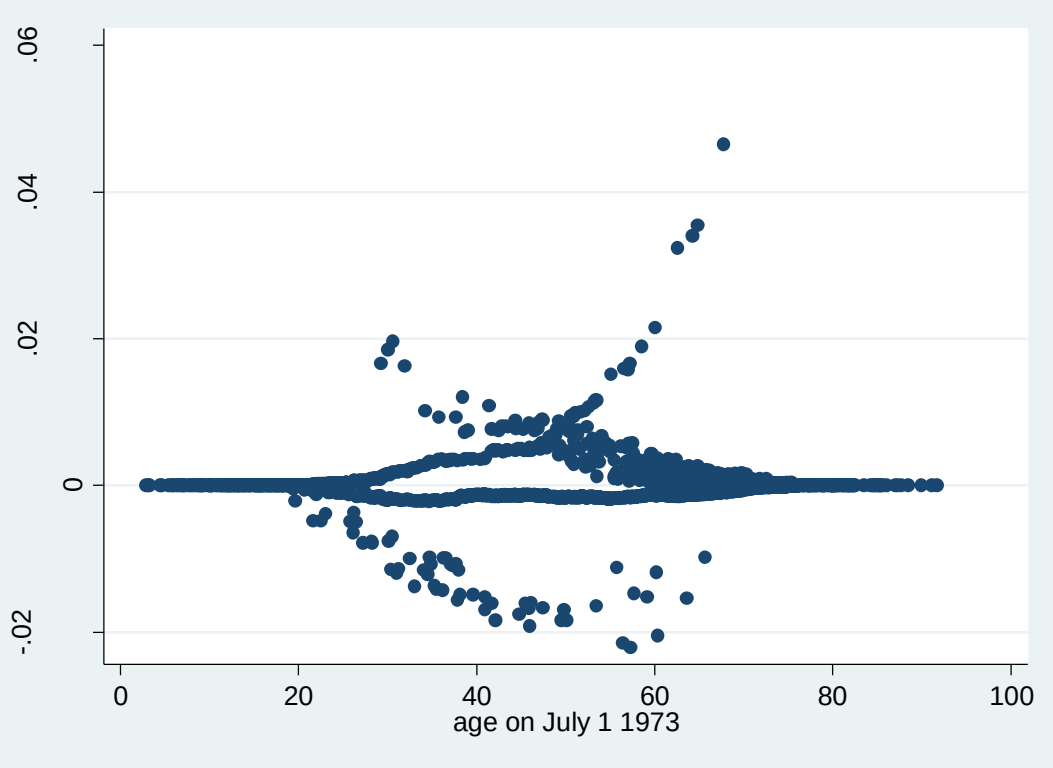
The point is that one would want to know if some small set of observations in the dataset are having inordinately large effect on the parameter estimates.

To see how this is used, I will apply it to the diabetes dataset. Setting up the data as before, we find

```
stcox older , nohr

Log likelihood  =   -2324.0326
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------------------------------
       older |  -.5008126   .1686263    -2.97   0.003
------------------------------------------------------------------------------
```

And from the commands below obtain a plot similar to that from the score residuals. In fact, DFBETA is simply SCORE_resid * cov(betahat). Plots of influence against other variables is usually more useful than simply a histogram.

predict influence , dfbeta
scatter exitage influence



The largest DFBETA is about 0.05, and this is for a man who was diagnosed at 15 years old, exiting this study alive at 76 years old. This is slightly peculiar, although the "influence" of 0.05 is not so large as to be really problematic.

When we omit that person the fit is as below, we find that the 0.05 is the approximate effect of <u>including</u> rather than <u>excluding</u> this person.

```
preserve
keep if influence < 0.04
  (1 observation deleted)
stcox older , nohr

Log likelihood  =   -2322.2933
------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------------
       older |  -.5495589   .1708115    -3.22   0.001
------------------------------------------------------------
```

I am adding this topic after recently finding it very useful in the A-bomb survivor work. There, the Cox regression for age at death to cancer included in the RR terms of form

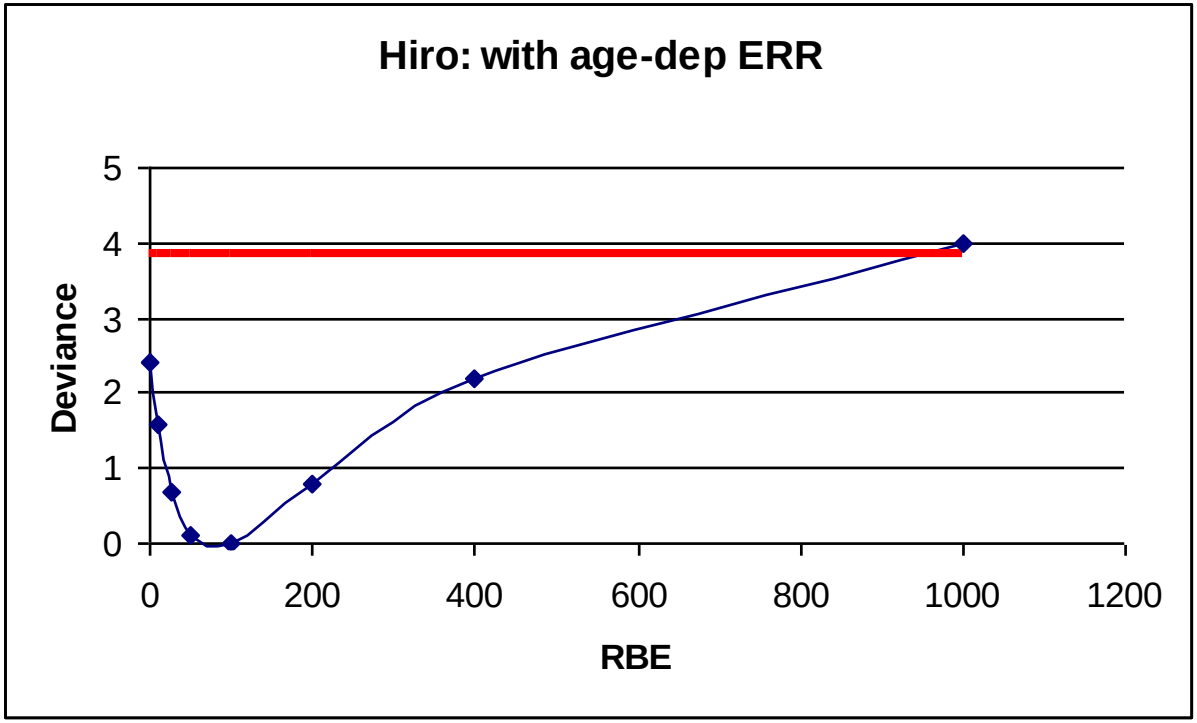
$$\beta_1 \ gamma\,ray\,dose + \beta_2 \ neutron\,dose$$

Where the neutron component is only about 1% of the total, and the two components are very highly correlated.

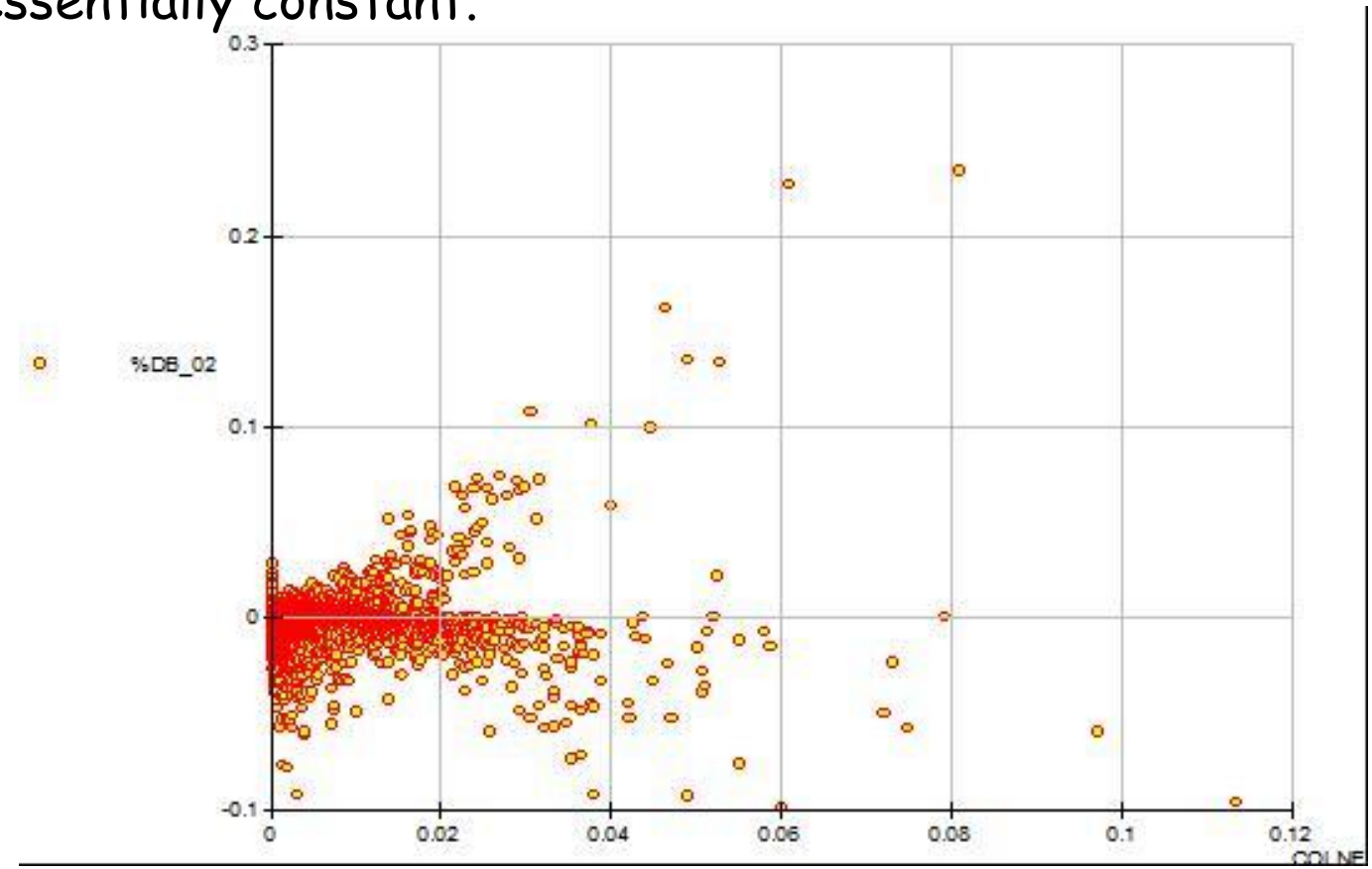
Therefore, the ratio β_2 / β_1 is usually just fixed at about 10 or 20, and our recent interest was in the feasibility of estimating that ratio.

We were greatly surprised to find that the data seemed far more informative on this than expected, and that the MLE was around 100.

Ratios where the Deviance is less than about 4 comprise a 95% confidence interval. MLE is nearly 100.



This is DFBETA vs neutron dose, for inference on the ratio β_2 / β_1 . The upper-right 3 points correspond to totally implausible dose estimates. After omitting these 3 out of 60,000 individuals, the deviance plot above becomes essentially constant.



We now want to introduce the data on excess cancer among A-bomb survivors, because this illustrates ways in which the standard approach to Cox regression requires extensions

The most important of these relates to needs is likely to be found in any dose response study

It develops that the needs cannot be met using any of the standard software packages (STATA, S-PLUS, etc), and I will briefly introduce the package that was developed specifically to meets the needs of interest here

A-bomb survivors: the main RERF dataset consists of follow-up for cancer of about 100,000 survivors with individual radiation dose estimates

We usually analyze this in grouped form, mainly because it encourages exploratory analysis with such a large dataset

Interest is in dose-response analysis, requiring methods rather different from usual survival analysis methods

This is because of need to focus on the excess risk due to radiation exposure, how it depends on dose level, and how factors such as sex and exposure age modify this

Will do a little bit in *STATA*, but mainly will demonstrate use of software that we developed for the needs (and is rather widely used elsewhere)

The usual approach would consist of modeling the cancer rates, relative to those for unexposed persons, in the form

$$RR = \exp(z'\beta)$$

where z is constructed from underlying covariables *dose, sex, exposure age, time since exposure, (possibly) attained age*

First consider only *dose* d . We might start with form

$$RR = \exp(\beta d) \doteq 1 + \beta d$$

Adding a *sex* term has no effect since the baseline risk must be allowed to depend on sex --- thus one must consider something like

$$RR = \exp(\beta_1 d + \beta_2 d \times s)$$

This is not unreasonable, but as one adds interactions with further factors the <u>multiplicative</u> form becomes quite unsatisfactory for studying the <u>excess</u> risk

What one really wants to analyze is more like how does the parameter β in models as below depend on *sex*, *exposure age*, etc

$$RR = 1 + \beta d$$

$$RR = 1 + \beta g(d)$$

where a form for $g(d)$ might be developed in exploratory analysis (including formal parameterizations)

A starting point is the generic form

$$RR = 1 + (z_1'\beta_1)\exp(z_2'\beta_2)$$

with an instance we use often

$$RR = 1 + \beta_1 d \times \exp(\beta_2 s + \beta_3 agex)$$

The class of models used in EPIWIN is much broader, but this form will do for our needs this week

In EPIWIN these are employed for Cox, Poisson, and Logistic regression

Another feature involves baseline "stratum" parameters such as those for age category above

Usually one wants far more of these --- stratifying also on other factors such as sex, calendar time, etc

Dealing with all these parameters in the ordinary fitting process would be infeasible

However, for fixed values of the RR parameters, the MLE of the stratum parameters is of closed form

The approach taken is iteratively to (a) fix the RR parameters, (b) compute the MLE of stratum parameters, (c) update the RR parameters holding the stratum parameters as fixed, then return to step (b), and so forth (a Gauss-Seidel algorithm)

The following example is for cancer mortality in the cohort, arriving at the model with age-dependent RR

$$RR = 1 + \beta_1 d \times \exp\{\beta_2 s + \beta_3 agex + \beta_4 \log(age)\}$$

For this, the baseline risk follows a stratified model using categories of: *city, sex, age, birth cohort, and calendar time*

Thus the model for cancer rates becomes

$$rate = \lambda_{c,s,a,bc,t}\left[1 + \beta_1 d \times \exp\{\beta_2 s + \beta_3 agex + \beta_4 \log(age)\}\right]$$

These data are publicly available from the website *rerf.or.jp* , and I can loan you the Epiwin program if you would like to try it (with some help from me)

```
tran agex30 = agex-30 @
tran lage60 = log(age/60) @
cases solid @
strata city sex agxcat agecat time @
                  The current model has 13260 strata
linear 1 sex*dose @
fit @
         Product additive excess model { T0 * ( 1 + T1 + T2 + ...) }
         Stratification on city sex agxcat agecat time with 1767 strata

 # Name                            Estimate     Std.Err.  P value
-- ---------------------------   ----------   ---------  --------
Linear term 1
 2 sex_1 * dose.............          0.3560     0.06542   < 0.001
 3 sex_2 * dose.............          0.6658     0.08139   < 0.001

                Deviance    12672.504

loglinear 1 agex30 @
fit @
-- ---------------------------   ----------   ---------  --------
Linear term 1
 2 sex_1 * dose.............          0.3779     0.06892   < 0.001
 3 sex_2 * dose.............          0.6478     0.08746   < 0.001

Log-linear term 1
 4 agex30...................       -0.04465    0.007761   < 0.001

                Deviance    12636.834
```

```
null @
loglinear 1 +lage60 @
fit @
 # Name                          Estimate     Std.Err.  P value
-- ---------------------------  ----------  ---------  --------
Linear term 1
 2 sex_1 * dose.............        0.3959     0.07271   < 0.001
 3 sex_2 * dose.............        0.6665     0.08975   < 0.001

Log-linear term 1
 4 agex30...................       -0.03761    0.00924   < 0.001
 5 lage60...................        -0.6537     0.5042      0.195

                Deviance    12634.934
lrt @
             LR statistic        1.900      Degrees of freedom      1
                  P value        0.168
profile 5 @
      2-sided          Bounds                exp(Bounds)
       Level       Lower       Upper       Lower        Upper

      ------------------------------------------------------------
      25.0%      -0.7990     -0.5061      0.4498       0.6028
      50.0%      -0.9599     -0.3395      0.3829       0.7121
      68.3%       -1.105     -0.1842      0.3311       0.8318
      75.0%       -1.172     -0.1115      0.3097       0.8945
      90.0%       -1.391      0.1323      0.2489        1.141
      95.0%       -1.529      0.2918      0.2168        1.339
      97.5%       -1.652      0.4375      0.1917        1.549
      99.0%       -1.798      0.6148      0.1656        1.849
      99.5%       -1.899      0.7404      0.1498        2.097
```

This includes a time-dependent covariable (log age), although in the Poisson regression this is less of a special issue than for Cox regression organized as usual

Although these data cannot be readily analyzed in STATA, I have prepared a version where some things can be done there.

The dataset *a_bomb_dat.dta* is a cross-tabulation including estimates for baseline rates that I have computed using EpiWin (given on the log scale in *lbkrate*)

You can do then some analysis by using Poisson regression with the "offset" in the linear predictor computed as
offst = log(pyr) + lbkrate

For example, useful results follow the command

poisson cases dose , offset(offst) nocon

Some Key Points of Summary

There are common settings where analysis of *rates* is much more feasible than analysis of *response times* (or survival functions), esp in settings involving delayed entry and competing risks

Cox regression does not require proportional hazards (even though that is important), since non-proportionality can be modeled in terms of time-varying covariables

If interest is primarily on *rate ratios* then Cox regression is usually the method of choice, but when interest is on *survival times* then Cox regression with *nonparametric* estimation of the baseline hazard is often not so useful

Although one *might* turn to (usual) "fully parametric" models, it will often be much better to use the Poisson regression approach with *exploratory parametric modeling* of the baseline hazard