Fourth Italian Stata Users Group Meeting

September 24–25, 2007 - Rome

Arjas plots
with Stata

Enzo Coviello

Outline of the talk

· Graphical checks of the proportional hazards
  assumption

· Brief digression on the hazard plot

· Arjas plots by examples

· -starjas-

Graphical checks of the proportional hazards assumption

- The validity of Cox's regression analysis relies on the assumption of proportionality of the hazard rates of individuals with distinct values of covariates.

- There are several graphical techniques for checking this assumption. Some of them are readily available by using official Stata commands or can be easily produced by just a few instructions.

- The effect of two categorical covariates are introduced as examples:

  - Distant metastases at diagnosis in Finnish colon cancer (available from the Paul Dickman web site net get http://www.pauldickman.com/rsmodel/stata_colon/strs), where hazard rates are not proportional

  - Detection method of diagnosis in a sample of Italian breast cancer screening study, where the proportionality assumption holds

Distant Metastases in Colon Cancer

```
. stcox distant, sch(sch) sca(sca) nolog

. estat phtest

    Test of proportional-hazards assumption
    Time:   Time
    --------------------------------------------------------------
                    |            chi2       df       Prob>chi2
    ----------------+---------------------------------------------
      global test   |           80.60       1          0.0000
    --------------------------------------------------------------
```

Invited to breast cancer screening test
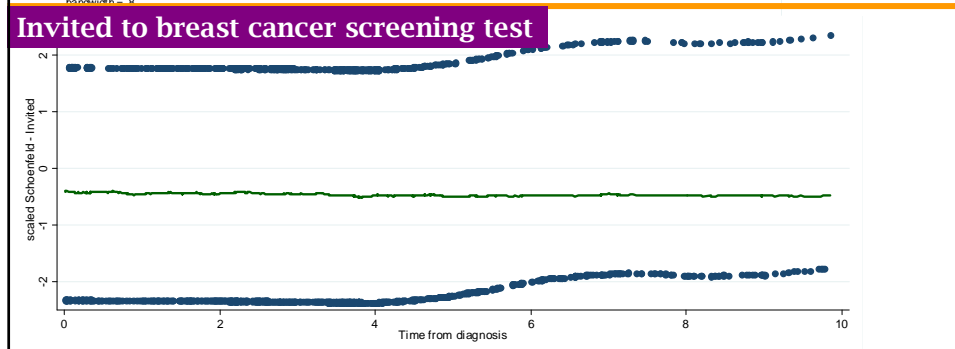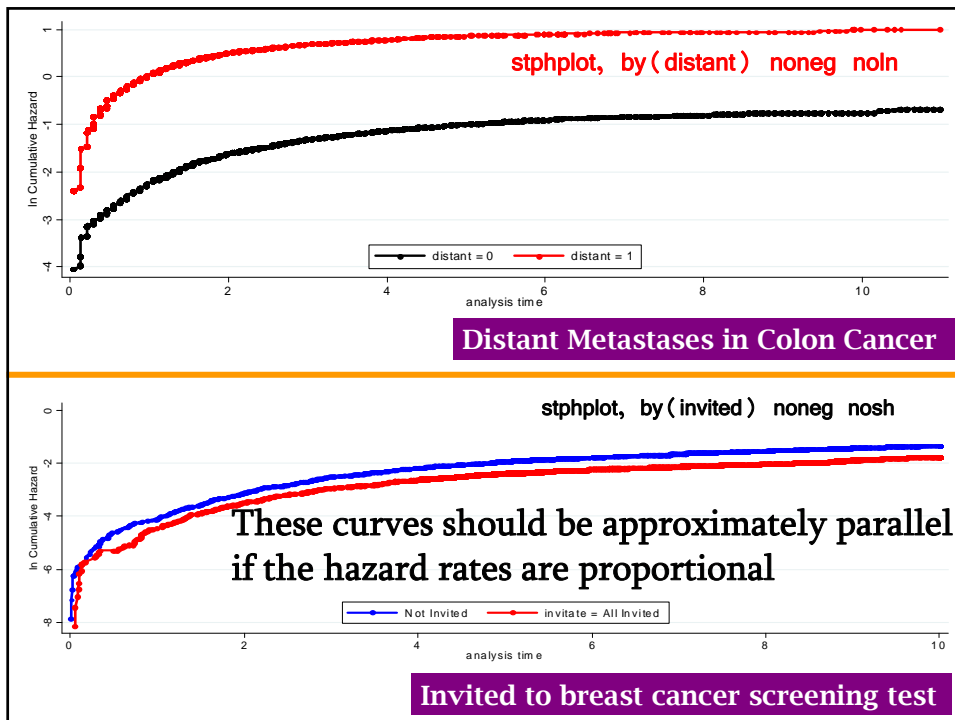
```
. stcox invited, sca(sca) sch(sch)
. estat phtest

    Test of proportional-hazards assumption
    Time:   Time
    --------------------------------------------------------------
                    |            chi2       df       Prob>chi2
    ----------------+---------------------------------------------
      global test   |            0.05       1          0.8281
    --------------------------------------------------------------
```

stphplot, by（distant） noneg noln

Distant Metastases in Colon Cancer

stphplot, by（invited） noneg nosh

These curves should be approximately parallel if the hazard rates are proportional

Invited to breast cancer screening test

An alternative to the previous approach is to plot log cumulative hazard difference. If the proportionality holds, **we should see a roughly horizontal line** that is easier to assess than comparing parallel curves.

Distant Metastases in Colon Cancer

Invited to breast cancer screening test

Andersen Plot

· Another graphical check based on cumulative hazard is the so-called Andersen plot.

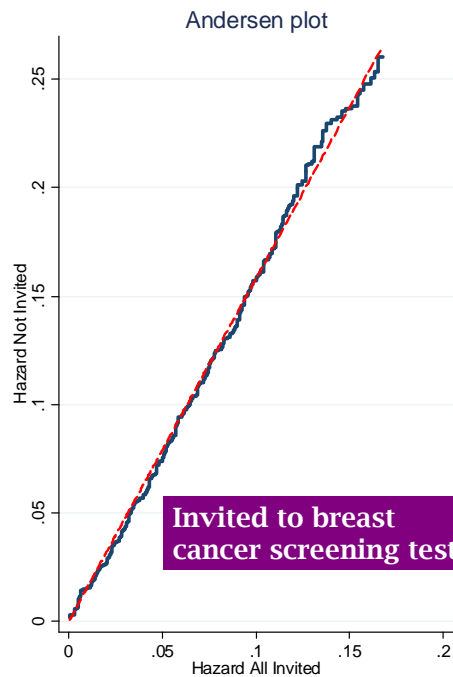· For a binary covariate as "distant" (or "invited") we simply plot:

$$H_{(distant=1)} \quad \textcolor{red}{versus} \quad H_{(distant=0)}$$
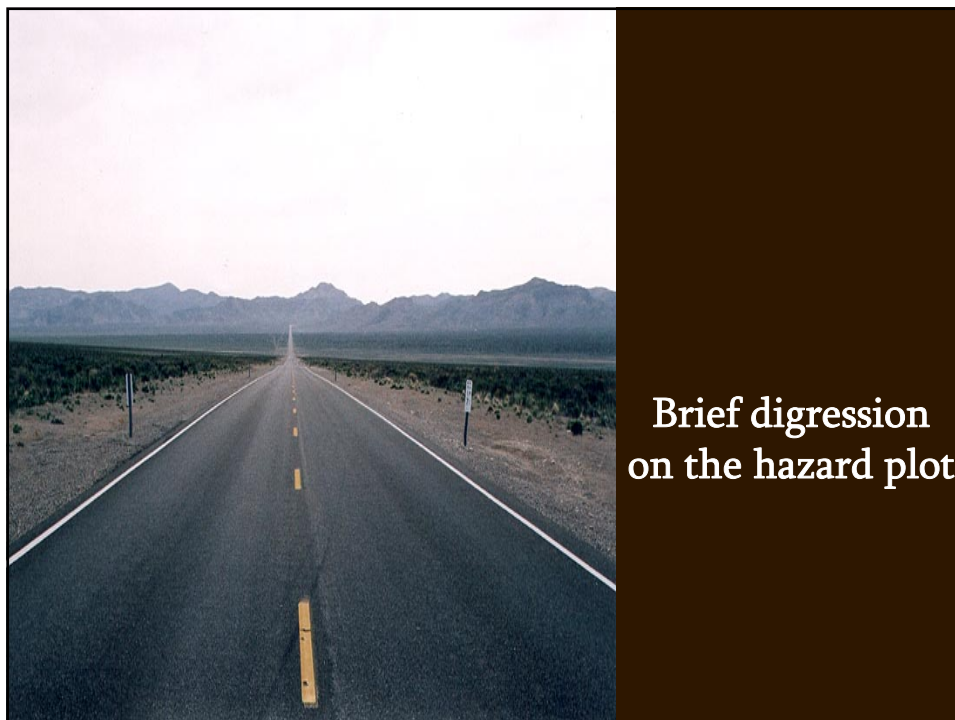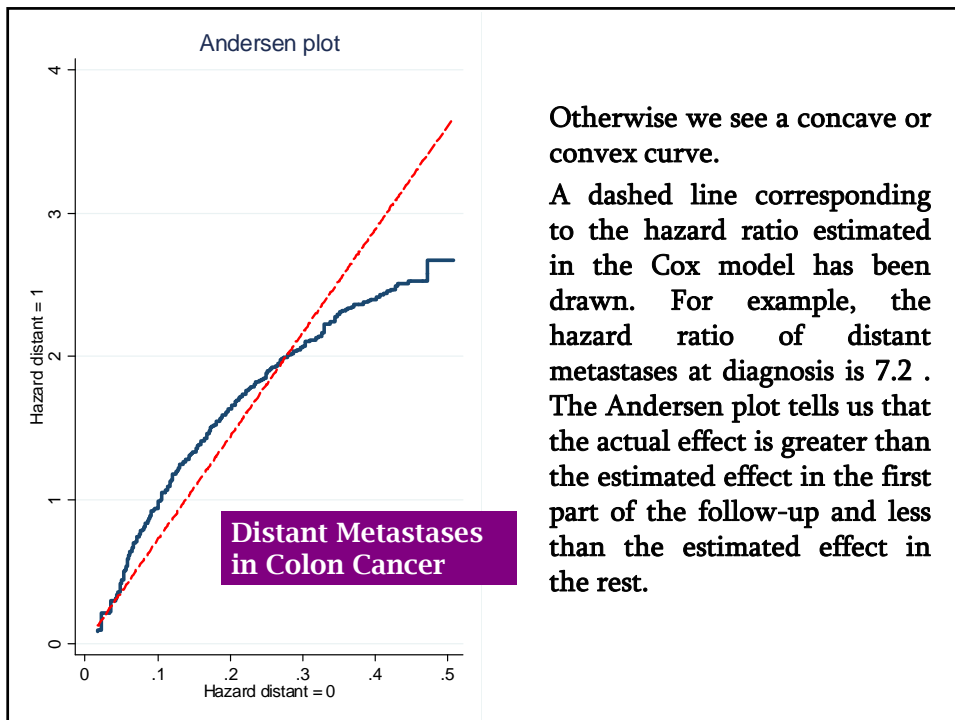
There is not a specific command to obtain this plot but the codes to be used are straightforward:

```
stcox , estimate strata(distant) basecha(Hdist)
g double Hdist1 = Hdist if distant==1
g double Hdist0 = Hdist if distant==0
g Hhat= Hdist0 * 7.225
sort _t
replace Hdist1 = Hdist1[_n−1] if Hdist1==.
replace Hdist0 = Hdist0[_n−1] if Hdist0==.
twoway line Hdist1 Hhat Hdist0 ···
```

Note that the variable we checked has been specified as **strata()** in -**stcox**-.

If the proportionality of hazards hold, we should see a straight line passing through the origin

Andersen plot

Invited to breast cancer screening test

5

Andersen plot

Distant Metastases in Colon Cancer

Otherwise we see a concave or convex curve.

A dashed line corresponding to the hazard ratio estimated in the Cox model has been drawn. For example, the hazard ratio of distant metastases at diagnosis is 7.2 . The Andersen plot tells us that the actual effect is greater than the estimated effect in the first part of the follow-up and less than the estimated effect in the rest.

Brief digression on the hazard plot

- Plotting the hazard function for each level of a categorical covariate is a direct approach to check the proportionality of hazards.

- To obtain this plot we have to use

  `-sts graph, hazard-`

  and not `-stcurve, hazard-`

  because we must compare non parametrical hazard estimates rather than model based hazard estimates.

- If the hazards are proportional, then the hazard curves are approximately parallel on a log scale.

- Hazard estimates are obtained by a kernel-smoothing technique applied to the cumulative hazard estimates.

- In Stata 10 smoothed hazard are bias-adjusted to take into account the restricted range of data available near the boundaries of the analysis time.

- Let me remember -stkerhaz-, a version 7 command available on SSC Archive, that computes the same bias-adjusted estimates by applying the so-called asymmetric kernel near the boundaries.

- Has -stkerhaz- still some utility since version 10 has been released?

In some case yes:

- -stkerhaz- allows non parametrical hazard estimates to be saved in a file (-stcurve- saves estimates, but -sts graph- does not). This could be useful for non standard graphs.

  For example, the bandwidth is a key parameter in the kernel smoothing method. This may be a sensible choice to employ short bandwidth at the start of the follow-up and larger at the end when the number of subjects still observed can be seriously reduced.

  By means of -stkerhaz- we can save smoothed hazard estimates obtained by using two (or more) bandwidths, we can combine the files and then produce a hazard plot with different degree of smoothing in different parts of the same curve.

- Excess risk model compares the mortality of a study population to the mortality of a reference population. By subtracting the cumulative expected hazard of the reference population (-stexpect-) to the cumulative hazard (-sts gen-) we can estimate the cumulative excess hazard.

  From the cumulative excess hazard, via -stkerhaz-, we can get and plot smoothed excess hazard estimates.
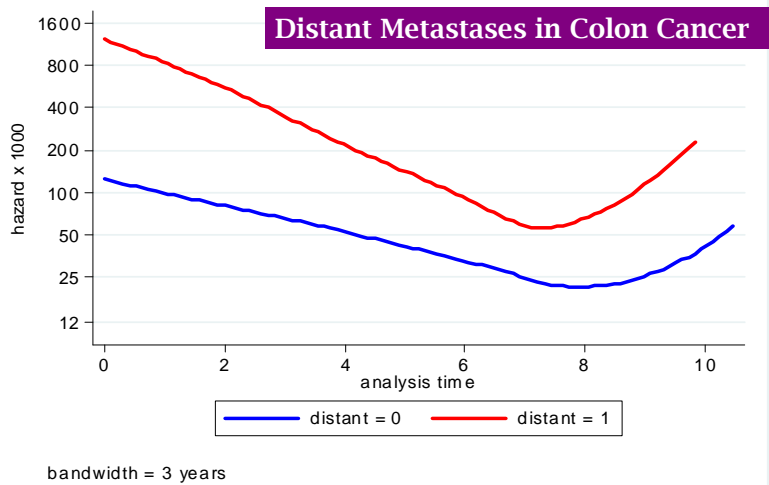
```
sts graph, hazard by(distant) width(3 3) k(epan2) ysca(log) ///
        yla(.012 .025 .05 0.1 0.2 0.4 0.8 1.6, angle(0))
```
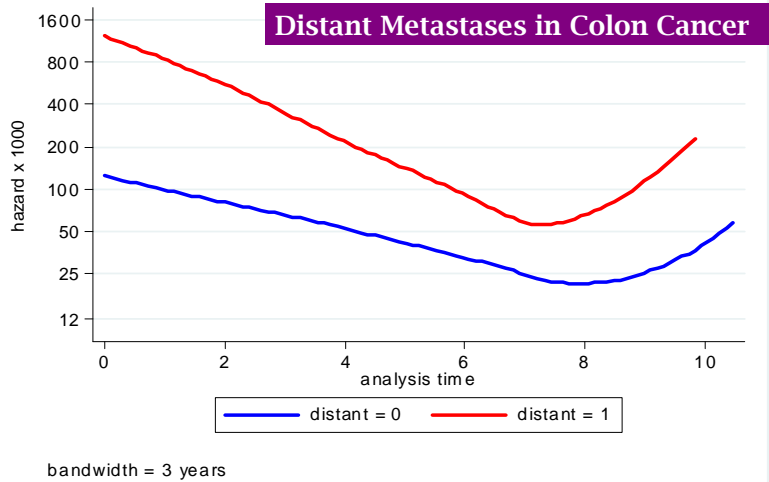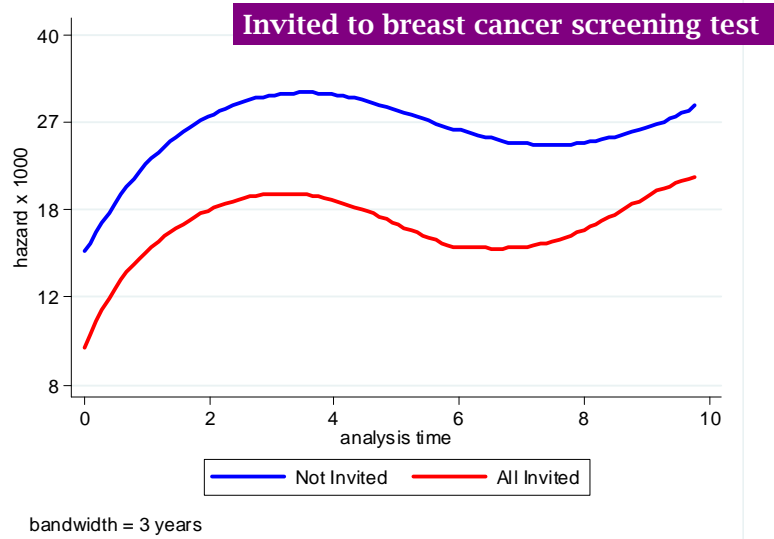

Here we see a decreasing distance of the hazard plots, i.e. a decreasing hazard ratio.

We should be cautious in evaluating the distance in the right tail of the curves since confidence intervals may be very large there.
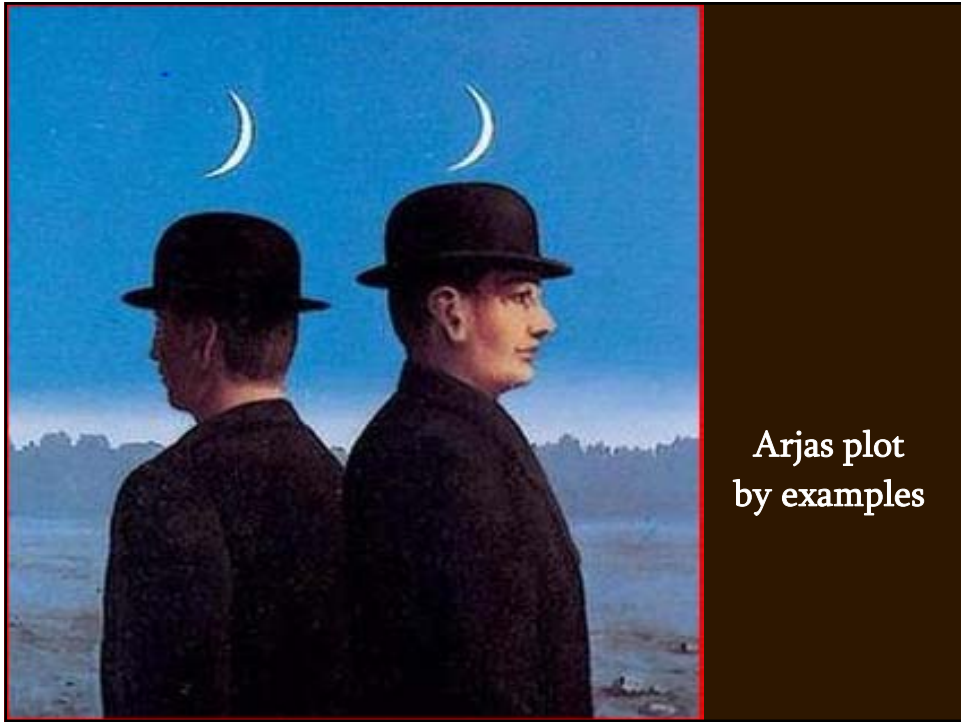
```
sts graph, hazard by(invited) width(3 3) k(epan2) ti("") ysca(log) ///
    yla(.008 .012 .018 .027 .04, angle(0))
```


Here we see a constant distance of the hazard plots, i.e. a constant hazard ratio.
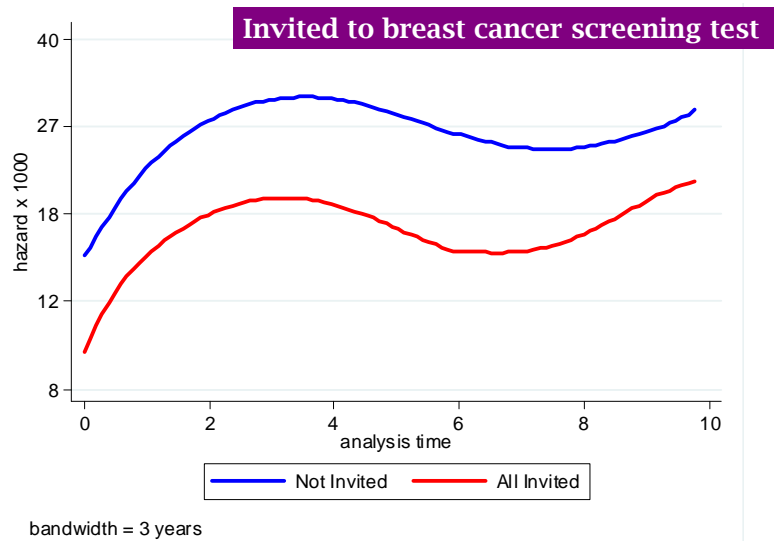
Arjas plot
by examples

- Arjas plot compares the total number of expected events to the total number of observed events occurring up to each event time.

- The cumulative hazard, calculated by the Nelson-Aalen estimator or by fitting a Cox model, is the base for working out the expected number of events.

- The Arjas plot allows to check:

  1. whether a covariate should be included in a proportional hazards model

  2. whether a covariate has proportional hazards effect.

1. Should a covariate be included in the model ?

· Assuming that a Cox model has been fitted with covariates x1 and x2, we wish to check whether a categorical covariate x3 should be included in the model.

· To this aim, we should first fit a Cox model with x1 and x2 covariates, ignoring x3.

· The resulting cumulative hazard estimate may be used to compute the total expected number of events at each event time and for each level of x3.

· If plotting the above estimate versus the total number of events for each level of x3 gives linear curves with slopes differing from 1, then x3 should be included in the model.

Let we consider the effect of method of detection of cases from the breast cancer screening study

. stcox invited, nolog nosh

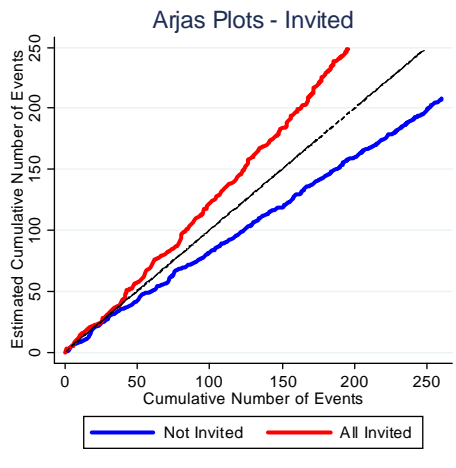Cox regression −− Breslow method for ties

No. of subjects =   2962
No. of failures =   456

         _t | Haz. Ratio
------------+-------------
    invited | .6283022

Let we consider the effect of method of detection of cases from the breast cancer screening study

starjas invited, … twoway options

In the plot we see a curve for each level of the invited covariate linearly differing from 45°

Arjas Plots - Invited

- In the breast cancer screening study we investigated whether the improved survival of cases invited to the screening test depended on the better tumour characteristics of cases diagnosed at the screening test or was biased by spurious factors (selection bias, lead time bias).

- To this aim we compared the survival experience of invited and not invited cases after adjusting for tumour characteristics: T , N and grading. If spurious factors play a role we should see some residual independent effect of the method of detection.

- In the Arjas plot we can investigate if a covariate should be included in the model after adjustment for other covariates.

- In the next slide Arjas plot for method of detection of cases is shown after adjusting for T , N and grading.

starjas invited, adjust（T2 T3 T4 T5 N2 N3 Gr2 Gr3 Gr9 yydx_m）

・The plot tell us that the method of detection has not residual effect after adjusting for T N and grading
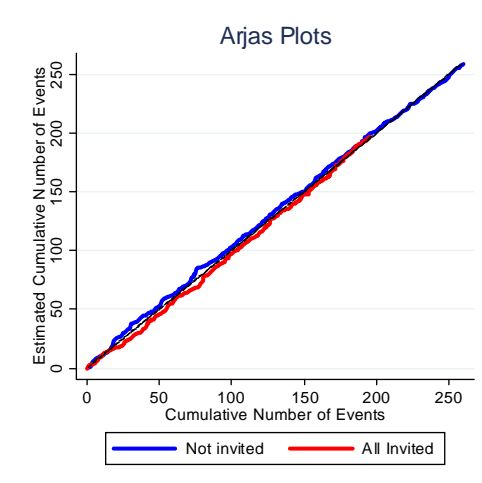
starjas invited, adjust（T2 T3 T4 T5 N2 N3 Gr2 Gr3 Gr9 yydx_m）

・The plot tell us that the method of detection has not residual effect after adjusting for T N and grading

. xi: stcox invited T2 T3 T4 T5 N2 N3 Gr2 Gr3 Gr9 yydx_m

Cox regression -- Breslow method for ties

No. of subjects =   2962   Number of obs =   2962
No. of failures =   456

```
------------------------------------------------------------------------------
      _t | Haz. Ratio   Std. Err.     z    P>|z|    [95% Conf. Interval]
---------+--------------------------------------------------------------------
 invited |  .9922653    .1020464   -0.08   0.940    .8111267    1.213855
    …omisis
------------------------------------------------------------------------------
```

・**Thus, using Arjas plot allows us to give a graphical expression to the effect of confounders on the main factor**

## 2. Does a covariate have proportional hazards effect?

Let we refer to the Bone Marrow Transplants data set (Klein and Moeschberger, p. 10 and p. 373) where the hazards of the autologous and allogeneic bone marrow transplant are clearly non-proportional

Smoothed hazard estimates
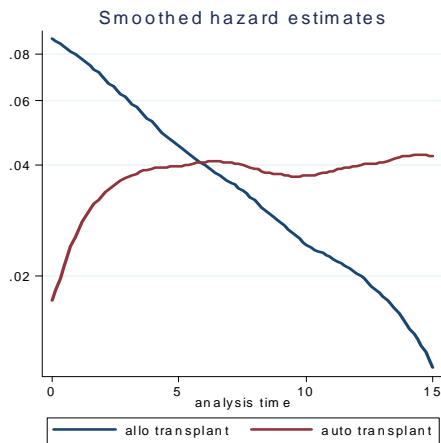
The hazard ratio of the "allotransplant" variable is greater than 1 at the beginning of the follow-up,

but it is sharply declining and becomes less than 1 after 6 months

---

starjas allotransplant

Arjas Plots - alloauto
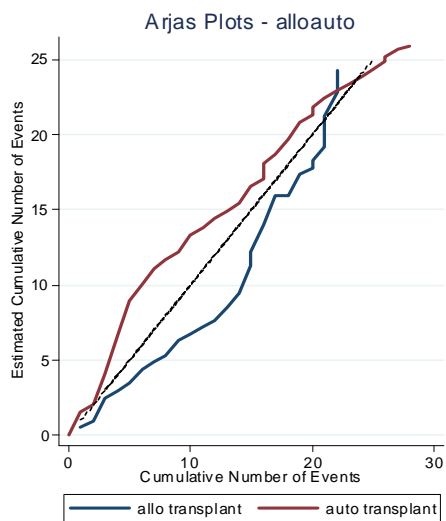
The curves

· differ nonlinearly from the 45° line

· show a decreasing divergence

· until they become parallel to the 45° line when the hazard ratio approaches the value 1

· and, finally, converge toward the 45° line when the hazard ratio turns < 1

starjas distant

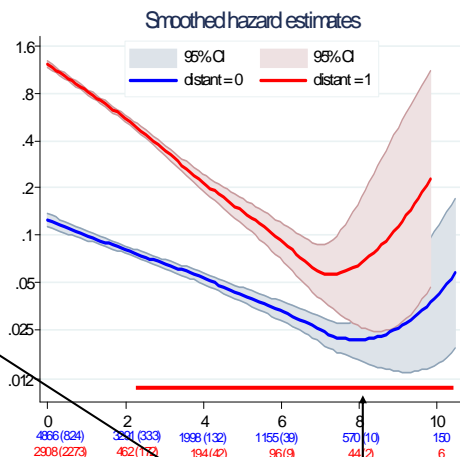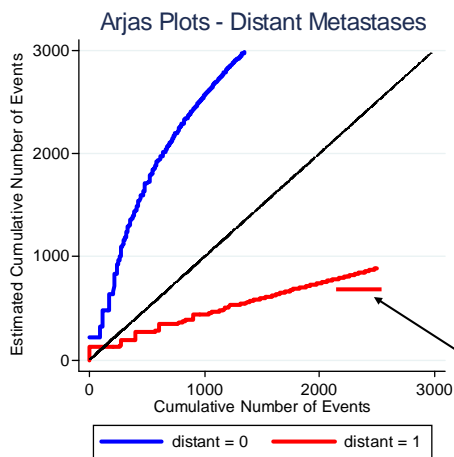Arjas Plots - Distant Metastases
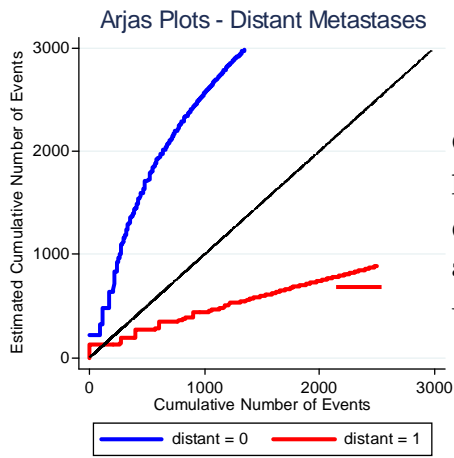
Cases without distant metastases produce a curve as expected

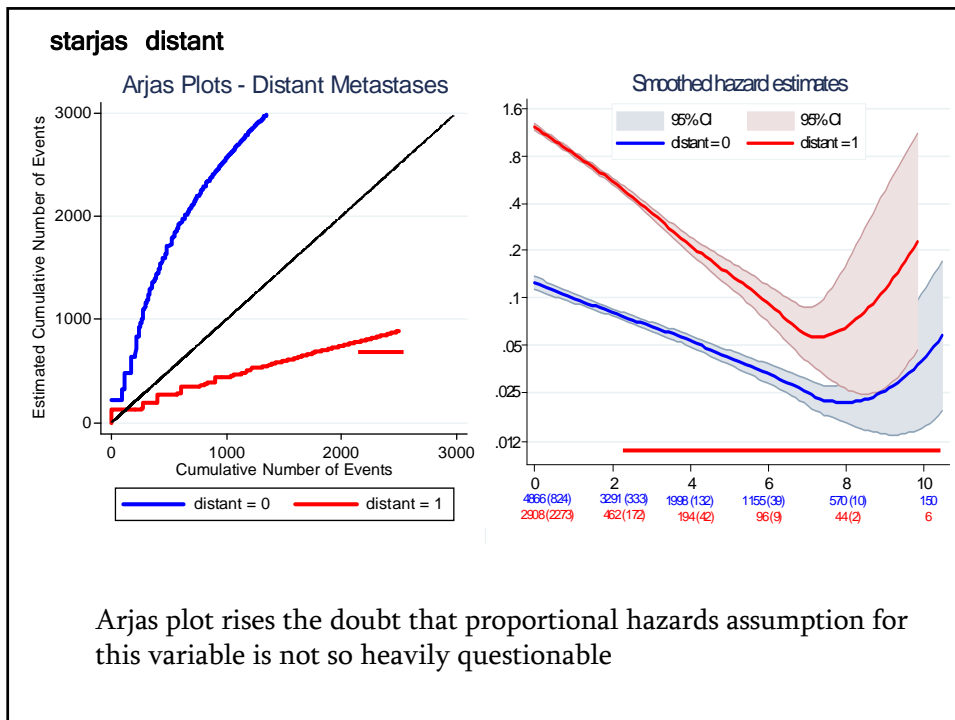Cases with distant metastases produce an approximately straight line.

Why this happens?

starjas distant

Arjas Plots - Distant Metastases

Smoothed hazard estimates

In the group with distant metastases only a few events (number in parentheses) occur after 2 years of follow-up.

Comparing observed and expected deaths after 2 years in the Arjas plot provides a very short line.

starjas   distant

Arjas plot rises the doubt that proportional hazards assumption for this variable is not so heavily questionable

starjas

```
help for  starjas
```

Title

    starjas —   Arjas plot to check proportional hazards assumption

Syntax

        starjas *varname* [**if** *exp*] [**in** *range*] [, **adjust**(*varlist*) **levelplot**(*#*
            [#] ..) atobs(*#*) **rrglance**(*#*) *twoway_options*]

options	description
Options	
adjust	specify the variables fitted in the Cox model before checking if *varname* is proportional
levelplot	specify the levels of *varname* to be displayed in the plot
atobs	restrict the plot to the first *#* events
rrglance	draw a line approximately corresponding in the case of binary covariate to a *#* relative risk
Y-Axis, X-Axis, Caption, Legend	
twoway_options	some of the options documented in **[G]** *twoway_options*
Add plot	
plot(*plot*)	add other plots to generated graph

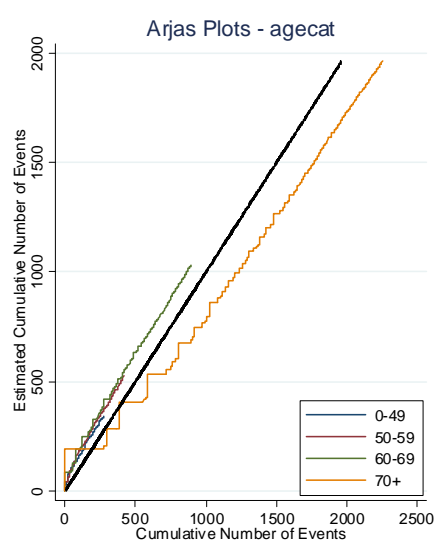   You must **stset** your data before using starjas; see **stset**.

---

levelplot (**#** [**#**] ...) can be useful when the variable we are checking has more than two levels. As an example, consider the effect of age at diagnosis in the Finnish colon cancer data set:

egen agecat, cut(0 50 60 70 100) icodes label

18
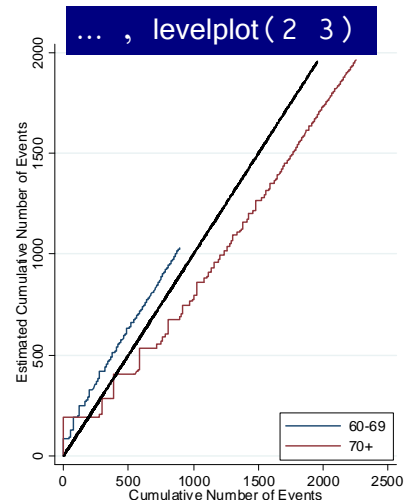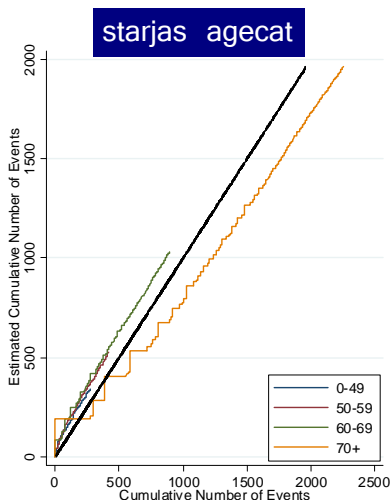
In interpreting plots for categorical covariates having more than two levels we must consider that the 45° line corresponds to the plot of expected and observed counts in the overall cohort of patients.

The plots for the age categories 0-49, 50-59 and 60-69 are above this line because the hazards of these groups are lower than the hazard of the overall cohort

and vice versa for cases in the age category ≥ 70.

Arjas Plots - agecat

Legend:
- 0-49
- 50-59
- 60-69
- 70+

(y-axis: Estimated Cumulative Number of Events; x-axis: Cumulative Number of Events)

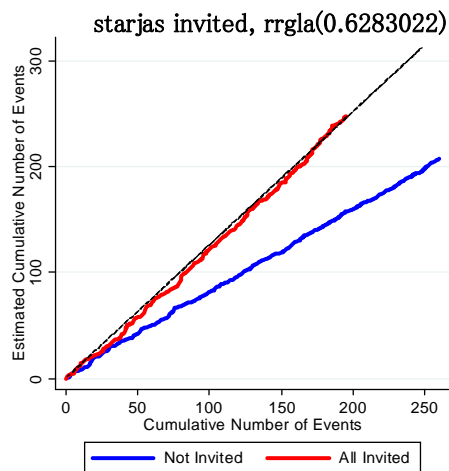In the case of a binary covariate, the dashed line in the Arjas plot indicates absence of effect, i.e. hazard ratio = 1.

rrglance(#) allows to move this line so that it corresponds to a # hazard ratio.

```
. stcox invited, nolog nosh

Cox regression -- Breslow
method for ties

No. of subjects =        2962

------------------------------
         _t |    Haz. Ratio
------------+-----------------
    invited |      .6283022
```

starjas invited, rrgla(0.6283022)

Legend:
- Not Invited
- All Invited

(y-axis: Estimated Cumulative Number of Events; x-axis: Cumulative Number of Events)

Conclusions

- Arjas plot is useful to address two critical questions in modeling survival data by a Cox regression:

  - Does the variable have to be included in the model ?

  - Is its effect proportional ?

- By this approach, we can try to answer to both questions using the same graph.

- -starjas- allows us to easily obtain Arjas plot and may be a complement to other "Cox diagnostic plots" available in the official package.

- It is freely downloadable from the SSC Archive by typing from within Stata:

<div align="center">

ssc install starjas

</div>

and hope the users enjoy it.

Thanks