Max Planck Odense Center on the Biodemography of Aging

University of Southern Denmark

Cox regression

Virginia Zarulli, Adam Lenart

MaxO - SDU

Introduction

- What is the Cox regression?
  - A semi-parametric proportional hazards regression.
    - Semi-parametric: does not assume any functional form for the baseline hazard.
    - proportional hazard: the covariate has the same proportional effect at all time points compared to the baseline hazard.
- How is it different from other techniques?
  - It concentrates (mainly) on the effect of the covariates by estimating the hazard function.
- What types of answers can it provide?
  - Hazard ratios for the covariates.
  - Rough estimate for the baseline hazard or survival function.

Cox regression

Recall that, the hazard, $h(t)$, in general,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

in the Cox model:

$$h(t) = f_0(t) e^{\beta_1 X_1 + \ldots + \beta_n X_n},$$

where

$f_0(t)$	unspecified baseline hazard
$\beta_1, \ldots, \beta_n$	coefficients of covariates $X_1, \ldots, X_n$
$X_1, \ldots, X_n$	covariates

Cox regression

Recall that, the hazard, $h(t)$, in general,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

in the Cox model:

$$h(t) = f_0(t) e^{\beta_1 X_1 + \ldots + \beta_n X_n},$$

where, for example, the coefficient β_1

- is the estimated influence of covariate X_1 while controlling for the effect of the other covariates.

Cox regression

Recall that, the hazard, $h(t)$, in general,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

in the Cox model:

$$h(t) = f_0(t) e^{\beta_1 X_1 + \ldots + \beta_n X_n},$$

where, for example, the coefficient β_1

- can be interpreted as the (logarithm of the) hazard ratio or relative risk of each increase in X_1 by 1.

Cox regression

Recall that, the hazard, $h(t)$, in general,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

in the Cox model:

$$h(t) = f_0(t) e^{\beta_1 X_1 + \ldots + \beta_n X_n},$$

where, for example, the coefficient β_1

- when $\beta_1 > 0$, the risk of dying increases and when $\beta_1 < 0$, the risk of death decreases.
- for $\exp(\beta_1)$, the hazard ratio, the risk of dying increases when $\exp(\beta_1) > 1$.

Cox regression (log hazard ratio)

Using the methadone dosage data for heroin addicts as in the Kaplan-Meier class

```
. stcox high_Dose, nohr

        failure _d:  Status == 1
  analysis time _t:  Time

Iteration 0:   log likelihood =  -705.6619
Iteration 1:   log likelihood = -696.67538
Iteration 2:   log likelihood = -696.52397
Iteration 3:   log likelihood = -696.52397
Refining estimates:
Iteration 0:   log likelihood = -696.52397

Cox regression -- Breslow method for ties

No. of subjects =          238                     Number of obs    =       238
No. of failures =          150
Time at risk    =        95812
                                                   LR chi2(1)       =     18.28
Log likelihood  =   -696.52397                     Prob > chi2      =    0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
high_Dose	-.7391731	.1690548	-4.37	0.000	-1.070515	-.4078317

A higher dosage of methadone has a -0.74 lower logarithm of risk of leaving a clinic than a lower dosage of methadone.

A higher dosage of methadone decreases the risk of leaving a clinic.

Cox regression (hazard ratio)

Using the methadone dosage data for heroin addicts as in the Kaplan-Meier class

```
. stcox high_Dose

        failure _d:  Status == 1
  analysis time _t:  Time

Iteration 0:   log likelihood =  -705.6619
Iteration 1:   log likelihood = -696.67538
Iteration 2:   log likelihood = -696.52397
Iteration 3:   log likelihood = -696.52397
Refining estimates:
Iteration 0:   log likelihood = -696.52397

Cox regression -- Breslow method for ties

No. of subjects =          238                     Number of obs    =       238
No. of failures =          150
Time at risk    =        95812
                                                   LR chi2(1)       =     18.28
Log likelihood  =   -696.52397                     Prob > chi2      =    0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
high_Dose	.4775086	.0807251	-4.37	0.000	.3428321	.6650908

The risk of leaving a clinic for patients on a high dose is 0.48*risk of leaving a clinic for patients on a low dose.

Patients on a high dose have 52% lower risk of leaving a clinic than patients on a low dose.

Cox regression - example 1

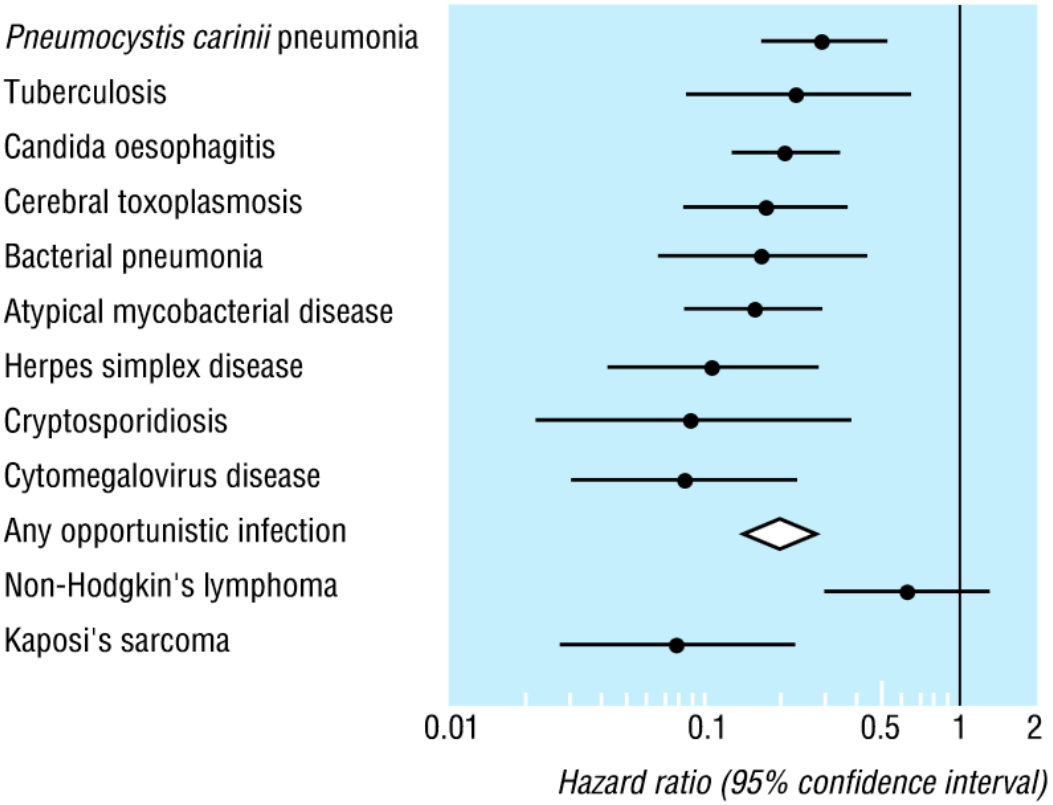
Table 2. Association of Sociodemographic Variables, Baseline Physical Health Status, and Baseline Caregiving Status With 4-Year Mortality*

Variables	Unadjusted Relative Risk (95% CI)	Adjusted Relative Risk (95% CI)
Sociodemographic factors		
Age	1.11 (1.07-1.15)†	1.10 (1.06-1.14)†
Sex	2.39 (1.58-3.62)†	1.88 (1.23-2.88)‡
Race	1.14 (0.61-2.13)	2.00 (1.03-3.89)§
Education, y	0.99 (0.95-1.03)	1.00 (0.96-1.05)
Stressful life events	0.93 (0.75-1.15)	0.83 (0.67-1.03)
Baseline physical health status		
Prevalent disease‖	4.55 (2.52-8.24)†	3.30 (1.79-6.08)†
Subclinical disease (no prevalent disease)‖	2.21 (1.20-4.08)§	1.84 (0.99-3.42)
Baseline caregiving status		
Not helping disabled spouse¶	1.84 (0.99-3.45)	1.37 (0.73-2.58)
Helping disabled spouse (no caregiving strain)¶	1.40 (0.81-2.42)	1.08 (0.61-1.90)
Helping disabled spouse (caregiving strain)¶	1.75 (1.10-2.80)§	1.63 (1.00-2.65)§

*Total number of subjects is 819. Total number of deaths is 103 (12.6%). CI indicates confidence interval.
†$P<.001$.
‡$P<.01$.
§$P<.05$.
‖Reference category is no subclinical or prevalent disease.
¶Reference category is no spouse disability (ie, control subjects).

Schulz et al. 1999: *Caregiving as a Risk Factor for Mortality*. JAMA 282(23), pp. 2215–2219.

Cox regression - example 2



Relative risk (hazard ratio) of AIDS defining opportunistic infections and malignancies, comparing 1992-4 (before introduction of potent antiretroviral combination therapy) with July 1997 to June 1998 (after introduction). Results from Cox regression models adjusted for transmission group, age, and CD4 cell count at baseline

Ledergerber et al. 1999: *Risk of HIV related Kaposi's sarcoma and non-Hodgkin's lymphoma with potent antiretroviral therapy: prospective cohort study*. BMJ 319(7201), pp. 23-24.

Cox regression with Stata

Running a Cox regression with Stata is extremely easy.

1. You need to declare the survival dataset with
   ```
   stset duration_variable,
   failure(failure_variable==code for failure)
   ```
2. Run Cox regression with
   ```
   stcox covariate1 covariate2 ..., options
   ```

Exercise 1

Analyzing the impact of methadone dosage on the risk of leaving a clinic.

1. Read `heroindata.dta` in.
2. Take a first look at Dose and Status.
   - `summarize` *varname*
   - `tabulate` *varname*`, summarize(`*varname*`)`
   - `hist` *varname*`, freq` to plot a histogram of the frequencies.

   What do you expect?
3. Create a dummy variable, 0 for low dosage ($<$ 60 mg), 1 for high dosage ($\geq$ 60 mg).
   - `gen` *varname* `= 1` creates a variable that is always 1.
   - `replace` *varname* `= 0 if` *varname* `<` *value* inputs a 0 wherever the if condition holds true.

Exercise 1 contd.

4. Set the survival model.
5. Check Kaplan-Meier by high dose. Interpret the results.
6. Run Cox regression. Interpret the results.
7. Compare Cox prediction with Kaplan-Meier.
   - `stcoxkm, by(`*varname*`)`

Cox regression - Stata output 1

By specifying high dosage as dose over 60 mg:

```
. stcox high_Dose, nohr

        failure _d:  Status == 1
  analysis time _t:  Time

Iteration 0:   log likelihood =  -705.6619
Iteration 1:   log likelihood = -696.67538
Iteration 2:   log likelihood = -696.52397
Iteration 3:   log likelihood = -696.52397
Refining estimates:
Iteration 0:   log likelihood = -696.52397

Cox regression -- Breslow method for ties

No. of subjects =          238                 Number of obs    =       238
No. of failures =          150
Time at risk    =        95812                 LR chi2(1)       =     18.28
Log likelihood  =   -696.52397                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   high_Dose |  -.7391731   .1690548    -4.37   0.000    -1.070515   -.4078317
------------------------------------------------------------------------------
```

We ask Stata to return the logarithm of the hazard rate (only for teaching purposes)

The regression model represents the data better than the null model (the model without covariates)

If we had 100 samples, 95 of them would have an estimated coef. between -1.07 and -.041

Wald-test score: Coef./Std. Err.

No effect if 0 is in the conf. interval

A higher dosage of methadone has a -0.74 lower logarithm of risk of leaving a clinic than a lower dosage of methadone.

A higher dosage of methadone decreases the risk of leaving a clinic.

high_Dose is statistically significant at a 0.000 level

Cox regression - Stata output 2

By specifying high dosage as dose over 60 mg:

```
. stcox high_Dose

        failure _d:  Status == 1
  analysis time _t:  Time

Iteration 0:   log likelihood =  -705.6619
Iteration 1:   log likelihood = -696.67538
Iteration 2:   log likelihood = -696.52397
Iteration 3:   log likelihood = -696.52397
Refining estimates:
Iteration 0:   log likelihood = -696.52397

Cox regression -- Breslow method for ties

No. of subjects =          238                 Number of obs   =       238
No. of failures =          150
Time at risk    =        95812

                                               LR chi2(1)      =     18.28
Log likelihood  =   -696.52397                 Prob > chi2     =    0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
high_Dose	.4775086	.0807251	-4.37	0.000	.3428321	.6650908

If we had 100 samples, the hazard ratio (relative risk) of 95 of them would be between 0.34 and 0.67

No effect if 1 is included in the conf. int.

The risk of leaving a clinic for patients on a high dose is 0.48*risk of leaving a clinic for patients on a low dose.

Patients on a high dose have 52% lower risk of leaving a clinic than patients on a low dose.

Exercise 1 contd. - baseline hazard

We can ask Stata to estimate and save the baseline hazard contribution of each observation with the `basehc(varname)` option.

8. Run Cox regression again with saving baseline hazard contribution in variable *h*.
   - `stcox ..., basehc(h)`
9. Plot baseline hazard contributions against time.
   - `line var1 var2, sort`
10. Draw a smoothed estimate of the hazard.
    - `stcurve, hazard at1(...) at2(...) kernel(...)`
    - Stata draws hazards at different covariate levels by setting `at1(varname=value) at2(varname=value)` etc.
    - Experiment with the `kernel(bw(value))` option which controls the smoothness of the curve(s).

Checking the proportionality assumption

The shape of the hazard,

$$h(t) = f_0(t)e^{\beta_1 X_1 + ... + \beta_n X_n},$$

stipulates that the covariates have a proportional effect on the baseline hazard.

- we need to check whether it is true.

Checking the proportionality assumption - Schoenfeld residuals

Schoenfeld residuals: difference between the covariate value for events and the weighted average (weighted by the estimated relative hazard of the Cox regression) of the covariate values of the remaining risk set when the event occurs.

- The (scaled) Schoenfeld residuals should not have any time trend.
  - **numerically:** `estat phtest, detail`
  - **graphically:** `estat phtest, plot(`*varname*`)`

Checking the proportionality assumption - Schoenfeld residuals

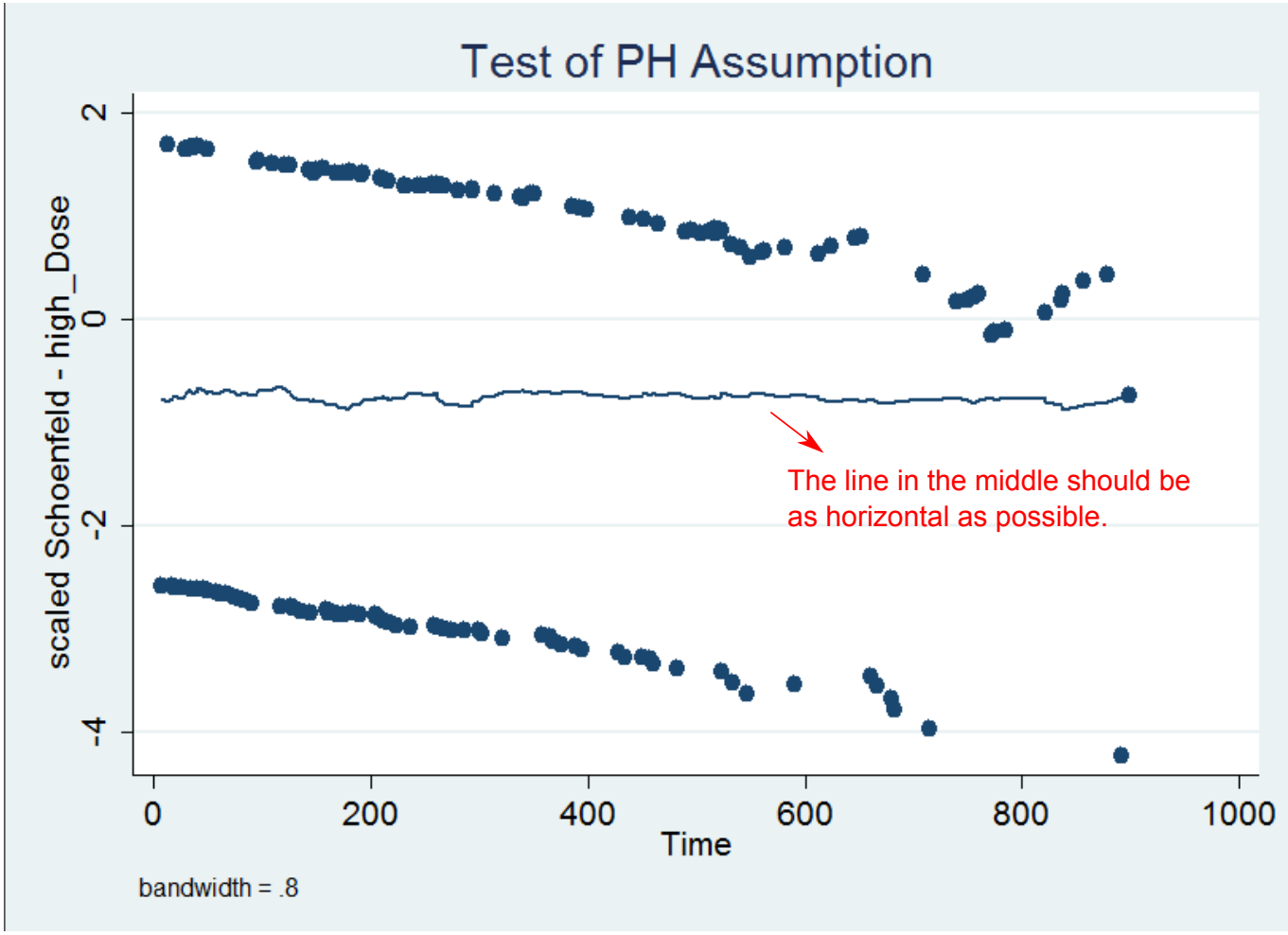
```
. estat phtest, detail

        Test of proportional-hazards assumption

        Time:  Time
```

	rho	chi2	df	Prob>chi2
high_Dose	**0.02174**	**0.07**	**1**	**0.7924**
global test		**0.07**	**1**	**0.7924**

Null hypothesis of the test is that there is no time trend. At p = 0.05, we do not reject the null hypothesis.

Global test yields the same result because we have only one variable in the model.

Checking the proportionality assumption - Schoenfeld residuals

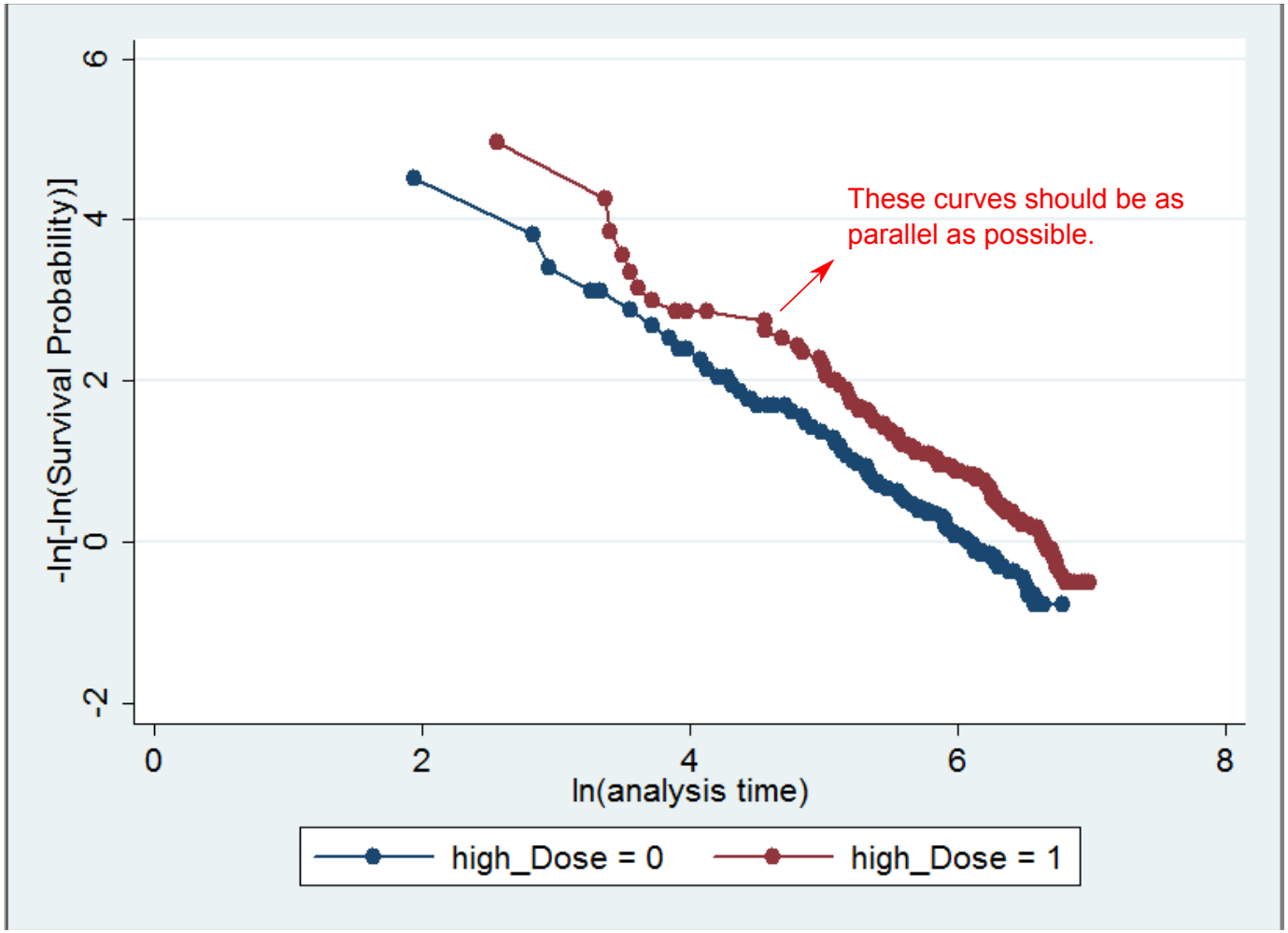


Checking the proportionality assumption - discrete covariates

For proportional hazards models, plotting - log of cumulative hazard against log of time should yield parallel curves.

- `stphplot, by(`*varname*`)`
  - If we had more covariates, we should include them with the `adjust(`*varnames*`)` option.
  - If we did a stratified regression, we should include `strata(`*varname*`)` option.

Checking the proportionality assumption - discrete covariates



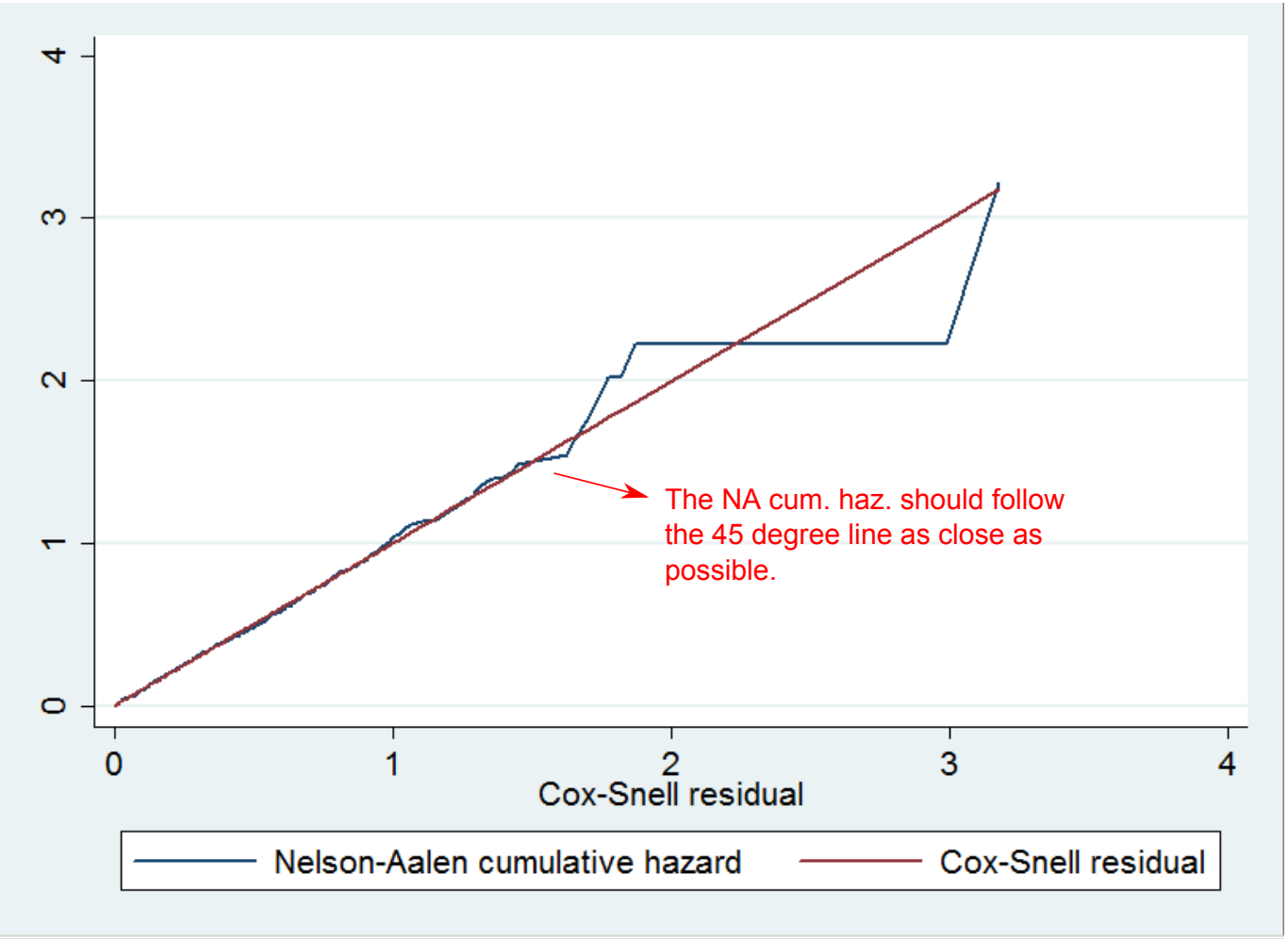
Goodness-of-fit

Martingale residuals: Observed events - Predicted events.
Cox-Snell residuals: Observed events - martingale residual.

For checking the goodness-of-fit of a Cox model, we use the Cox-Snell residuals. They should follow a right-censored exponential distribution with a mean 1.

- Plot histogram.
- Set the Cox-Snell residuals as the duration variable in a survival model with the same censoring variable as before, estimate its cumulative hazard function and plot it against the Cox-Snell residuals. The resulting curve should be a 45 degree line.

Checking the goodness-of-fit of the model



Exercise 1 - model diagnostics

11. Check whether higher dosage of methadone decreases the risk of leaving a clinic at all time points by the same proportion.
    - `estat phtest, detail`
    - `estat phtest, plot(`*varname*`)`
    - `stphplot, by(`*varname*`)`

Exercise 1 - model diagnostics contd.

12. Check how well the proposed simple model fits the data.
    - To get Cox-Snell residuals, first, we need to estimate the martingale residuals by running the Cox regression again and specifying the `mgale` option.
      - `stcox ..., mgale(`*`varname`*`)`
    - Then calculate the Cox-Snell residuals by `predict` *`varname`*`, csnell`.
    - Plot histogram of it.
      - `hist` *`varname`*
    - Generate the cumulative hazard function of the Cox-Snell residuals.
      - `sts gen` *`varname`* `= na`. The `na` stands for Nelson-Aalen estimator.
    - Plot Cox-Snell residuals against its cumulative hazard function.
      - `line` *`var_cumhaz var_cs var_cs`*`, sort xlab(0 1 to 4) ylab(0 1 to 4)`

Stratified Cox regression

Just like with the Kaplan-Meier, we can have a stratified Cox regression by adding the `strata(varname)` option to the regression.

$$h(t|X_j) = f_{0j}(t)e^{\beta_1 X_{1j}+...+\beta_n X_{nj}},$$

that is, we estimate a different baseline hazard for each category of the stratification variable.

Exercise 1 - stratification

The patients might be treated differently in the two clinics, the effect might not be proportional at all time points and we are not interested in it per se. However, we still assume that irrespective of the clinic, methadone dosage should have the same effect.

13. Run a stratified Cox regression by clinic.
    - First, you will have to reset the survival model because it was overwritten by the model with the Cox-Snell residuals.
    - `stcox ..., strata(`*varname*`)`

Exercise 1 - stratification contd.

14. Plot baseline cumulative hazard by each strata.
    - use `basech(varname)` option after Cox regression to save the baseline cumulative hazard.
        - `stcox ..., strata(varname) basech(varname)`
    - Save the two baseline hazards of each clinic in two new variables and plot them.
        - `gen varname = var_cumhaz if Clinic == 1`
        - `gen varname = var_cumhaz if Clinic == 2`
        - `line var1 var2 Time, connect(J J) sort`, with `connect(J J)` we ask Stata to plot step functions.

Continuous covariates

The interpretation of the coefficients of continuous covariates is the same as discrete covariates.

```
Cox regression -- Breslow method for ties

No. of subjects =          238                 Number of obs    =         238
No. of failures =          150
Time at risk    =        95812
                                               LR chi2(1)       =       36.89
Log likelihood  =   -687.21453                 Prob > chi2      =      0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Dose	.9647253	.0057667	-6.01	0.000	.9534886	.9760943

Each mg increase of methadone dosage decreases the risk of leaving a clinic by 3.5% (compared with the 1 mg lower level).

Functional form of covariates

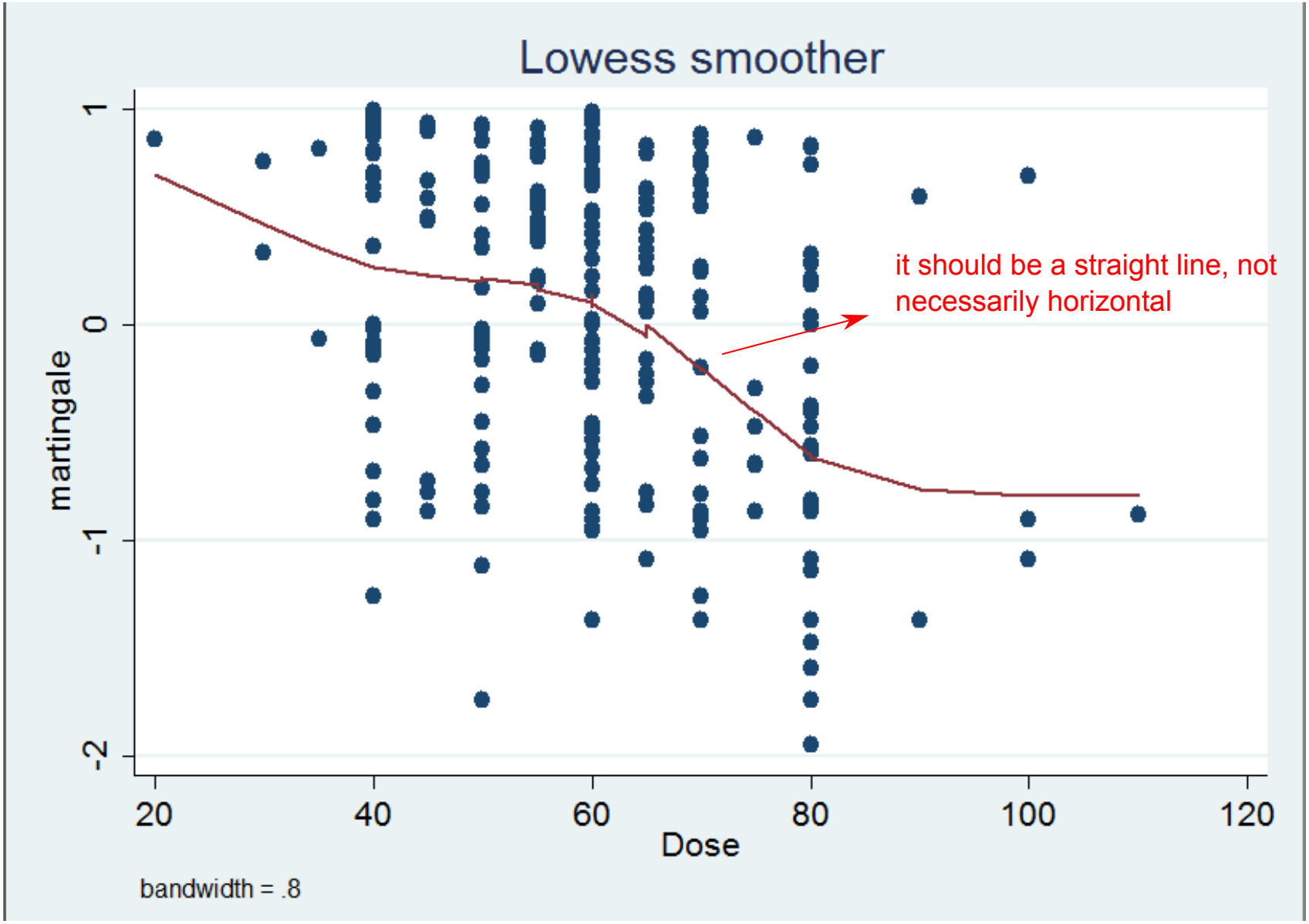
Difference between discrete and continuous covariates: functional transform

- martingale residuals of a Cox model without covariates might help to give an idea about a more suitable functional form.
  - essentially, the martingale residuals are censoring variable - cumulative hazard.

Often used functional transformation are:

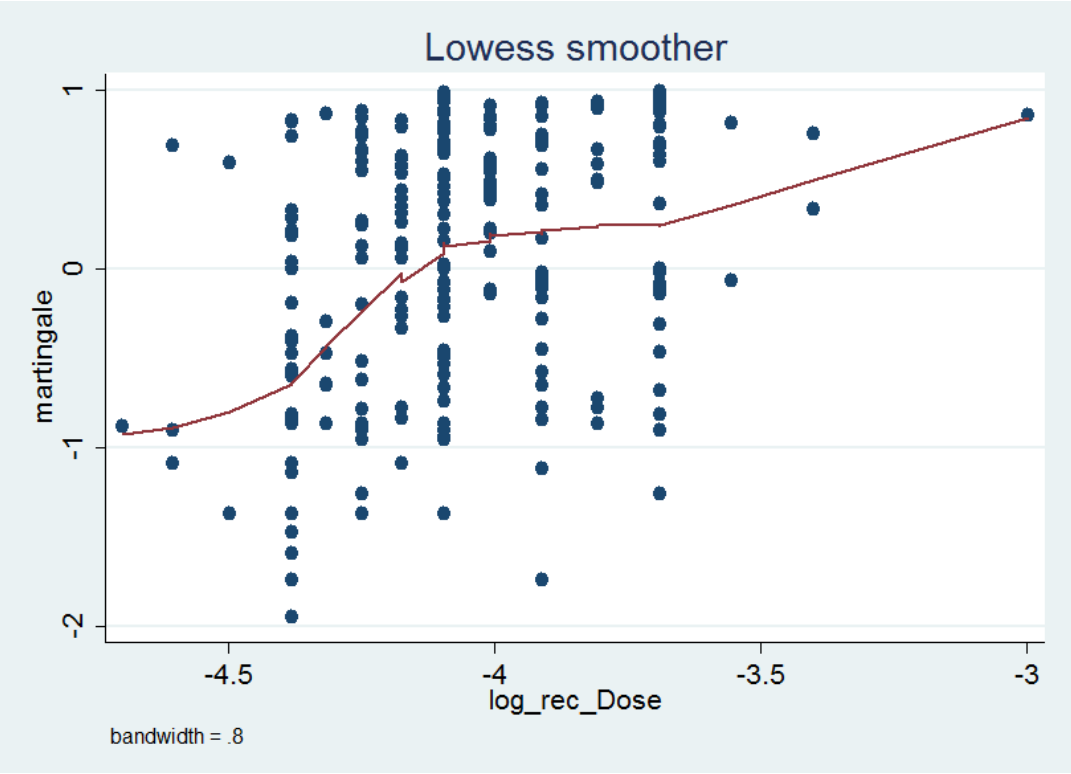
- logarithmic
- square root
- reciprocal
- and other power transforms (squared, cubic, etc.).

Functional form of covariates contd.



Functional form of covariates contd.

- Here I took the logarithm of the reciprocal of dosage.
- If we don't find a suitable functional form easily (as here), we can still turn the continuous variable into a categorical one.



Exercise 1 - find functional form

15. Run Cox regression without covariates (null model)
    - `stcox, mgale(`*varname*`) estimate`
16. Plot martingale residuals against methadone dosage.
    - `lowess` *var_martingale varname*
17. Generate transformed variable for methadone dosage and plot it against martingale residuals
    - *varname* `= log(1/`*Dose*`)`
    - `lowess` *var_martingale varname*

Time-varying covariates

Spell duration data: multiple observations for each subject, each covering a span of time.

Two types of time varying objects:

1. A covariate whose value changes over the spells (time-varying covariate)
2. A covariate whose effect changes over the spells (time-varying coefficient)

Time-varying covariates

1. **Time-varying covariate.** When during the spells the value of a specific covariate changes: the level of calcium in the patient during the time to bone fracture.

id	time0	time	fracture	calcium
1	0	3	1	11.77
2	0	2	0	11.82
2	2	5	1	11.34
3	0	4	0	11.98
3	4	7	0	11.69
3	7	9	1	11.08
...	...	...	...	...

Handled by Stata automatically, so just type:

```
stset time failure(fracture==1) time0(time0)
stcox calcium
```

Time-varying coefficients

2. **Time-varying coefficient (interaction between covariate and time)**. When the effect of a covariate changes over the time spells: the impact of mother's pre-pregnancy BMI on time to birth in different gestational weeks (for example: up to week 37, from week 37 on).

Time-varying coefficients

id	time0	time	status	BMI	BMI37w
1	0	35	1	26	0
2	0	37	0	19	0
2	37	39	1	19	19
3	0	37	0	23	0
3	37	41	1	23	23
4	0	37	1	25	0
...	...	...	...	...	...

```
stcox BMI BMI37w
```

effect BMI $= \beta_1 \cdot BMI + \beta_2 \cdot BMI37w$

Time-varying coefficients

id	time0	time	status	BMI
1	0	35	1	26
2	0	39	1	19
3	0	41	1	23
4	0	37	1	25
...	...	...	...	...

```
stcox BMI, tvc(BMI) texp(time>37)
```

effect BMI $= \beta_1 \cdot BMI + \beta_2 \cdot I(t > 37) \cdot BMI$

Limitation of this apprach: only one function of time can be specified.

Time-varying coefficients

We are interested in the impact of mother's pre-pregnancy BMI on time to birth in different gestational weeks: up to week 37, from week 37 to week 40, after week 40 (we partition the time into 3 intervals).

In this case we need to use the expanded version of the data, and we need to expand it according to our time intervals. `texp(...)` does not allow specifying more than one cut-point, so we need to use `stsplit(...)`.

Time-varying coefficients

`stsplit` *`newvar`*`, at(37 40)`

creates a new variable whose value is 0 for weeks before 37, 37 for weeks from 37 to 40 and 40 for weeks from 40 on.

`stsplit cuts, at(37 40)`

id	time0	time	status	BMI	cuts
1	0	35	1	26	0
2	0	37	0	19	0
2	37	39	1	19	37
3	0	37	0	23	0
3	37	40	0	23	37
3	40	41	1	23	40
4	0	37	1	25	0

Time-varying coefficients

`gen BMIw37=BMI*(cuts==37)`

`gen BMIw40=BMI*(cuts==40)`

generate *BMIw37* and *BMIW40* that are equal to *BMI* when *cuts* is respectively equal to 37 and 40.

id	time0	time	status	BMI	cuts	BMIw37	BMIw40
1	0	35	1	26	0	0	0
2	0	37	0	19	0	0	0
2	37	39	1	19	37	19	0
3	0	37	0	23	0	0	0
3	37	40	0	23	37	23	0
3	40	41	1	23	40	23	23
4	0	37	1	25	0	0	0

`stcox BMI BMI37w BMI40w`

effect BMI $= \beta_1 \cdot BMI + \beta_2 \cdot BMI37w + \beta_3 \cdot BMIw40$

Exercise 2 (optional)

1. Open the dataset reyes.dta.
2. Take a look at the data.
3. Prepare the data for survival analysis specifying also the id variable (we will need this for later).
4. Explore the data with KM (for the continuous variables, categorize them into 2 or 3 groups).
5. Run Cox model 1 with all the covariates.
6. Run Cox model 2 with the covariates that you think are necessary and the categorical age instead of the continuous age.

Exercise 2 (optional)

7. Check the proportionality assumption (`ph assump`). Which variables satisfy the ph assumption? Which ones do not?
8. Run Cox model 3 stratified by the variable that does not satisfy ph assump. and plot the baseline cumulative hazard for each stratum.
9. Interpret the results of the model 3 and the cumulative hazards plot. What is the effect of being treated vs non-treated? What percentage of increase/decrease in the risk of death is caused by one unit increase of age? and of *ammonia*? and of *sgot*?
10. Do you think any of the variables could be transformed for a better fit of the model? If yes, which variable would you transform? And how?

Exercise 2 (optional)

11. Run Cox model 4 with the transformed variable.
12. Does model 4 significantly increase the fit of the data? Compare the two models with AIC (`estat ic` after the regression; the lowest AIC shows the best model).
13. The KM plot by *`ftliver`* seems to show a time varying effect of this covariate, especially between 10 and 20 the curves seem to diverge and then, later, converge. To identify the cut point, list the KP curves by *`ftliver`* at times from 10 to 20. When do the two curves starting to converge?

Exercise 2 (optional)

14. Run a cox regression with data split at the cut point by using the `tvc() texp()` syntax.
15. Now run the same regression using the `stsplit` command.
16. Check goodness-of-fit.