BIOST/EPI 537
Survival data analysis for epidemiology

Chapter 3: Regression models in survival analysis: the proportional hazards model

Marco Carone
Department of Biostatistics
School of Public Health, University of Washington

Winter 2020

LAST UPDATED ON DEC 22ND 2019

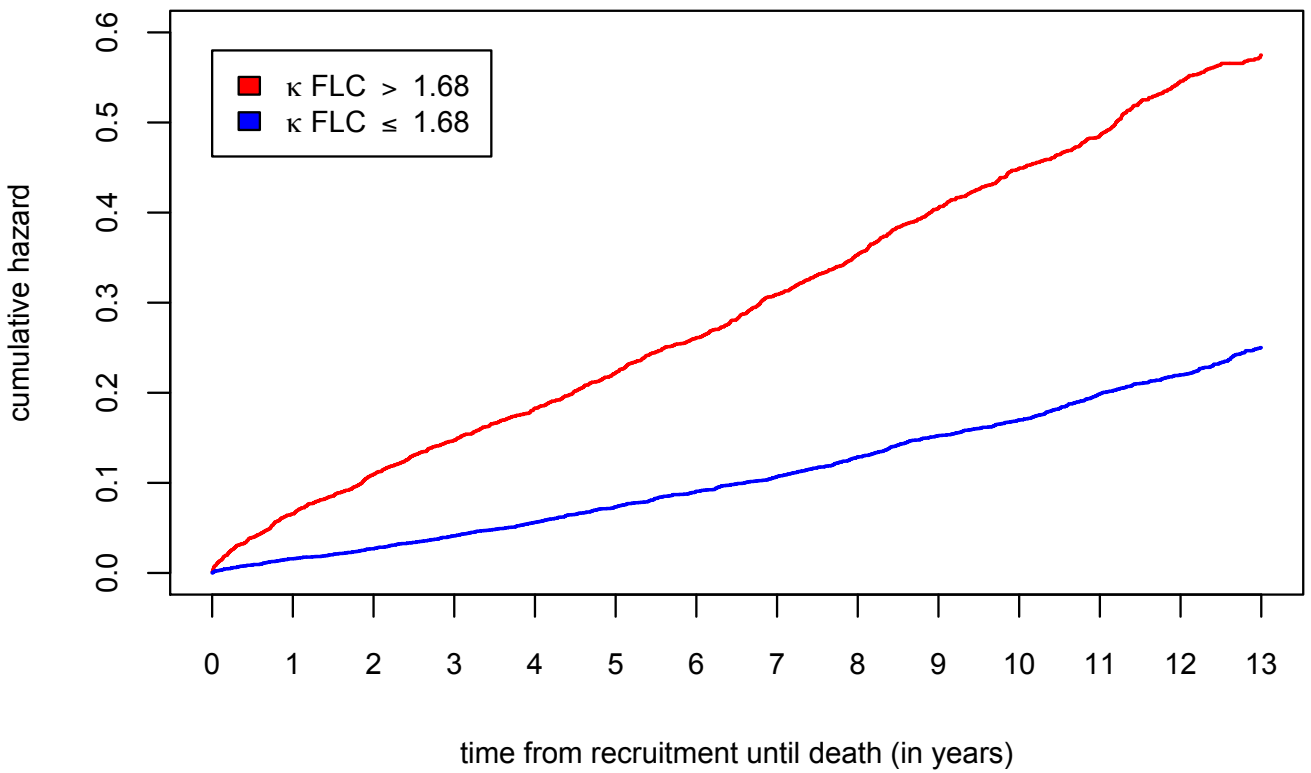
Contents of this chapter

- Formulation, properties and parameter interpretation
- Parametric proportional hazards model
- Semiparametric proportional hazards model
- Stratified proportional hazards model
- Diagnostic tools

Formulation, properties and parameter interpretation

We begin with the simple problem of **describing the survivorship of two subgroups**.

A nonparametric approach would consist of splitting the dataset according to these subgroups and performing a separate nonparametric analysis in each.



Formulation, properties and parameter interpretation

This plot was based on data from a study of the association between serum free light chains (FLCs) and mortality in residents of Olmsted County of age 50 years or more.

It appears that the (cumulative) hazard function in patients with higher levels of κ FLCs is proportional to the (cumulative) hazard function in other patients.

Inspired by this observation, rather than estimating the survival distribution separately in each group, we may have considered the model

$$h_1(t) = c \cdot h_0(t) \ ,$$

where h_0 and h_1 are the subgroup-specific hazard functions.

This is the simplest example of a **proportional hazards model**, which stipulates that...

...the hazard functions corresponding to every subgroup considered are proportional to one another.

In our simple two-sample version of this model, we can re-express the model as

$$\frac{h_1(t)}{h_0(t)} = c \quad \text{for every time } t$$

and so, **the hazard ratio comparing the two subgroups is constant over time**.

The hazard ratio at time t tells us...

> how much more likely it is that an event will happen in subgroup 1 at time t versus in subgroup 0 *given that it has not yet happened in either*.

$$\frac{h_1(t)}{h_0(t)} = \frac{h_1(t)\Delta t}{h_0(t)\Delta t} \approx \frac{P\left(t \leqslant T < t + \Delta t \mid T \geqslant t, \text{ subgroup } 1\right)}{P\left(t \leqslant T < t + \Delta t \mid T \geqslant t, \text{ subgroup } 0\right)}$$

Formulation, properties and parameter interpretation

$c < 1$ = **decreased hazard** in subgroup 1 relative to subgroup 0

The event tends to occur ..... in subgroup 1 versus subgroup 0 since *at any time the hazard of an event in subgroup 1 is ..... than in subgroup 0 by* $100 \times (1 - c)\%$.

$c > 1$ = **increased hazard** in subgroup 1 relative to subgroup 0

The event tends to occur ..... in subgroup 1 versus subgroup 0 since *at any time the hazard of an event in subgroup 1 is ..... than in subgroup 0 by* $100 \times (c - 1)\%$.

time-to-event	$c < 1$	$c > 1$
transplant → organ rejection onset of disease → death	+	−
recruitment → tenure treatment → remission	−	+

What does this say about the survival function in each group?

$$S_1(t) = \exp\{-H_1(t)\} = \exp\{-cH_0(t)\} = [\exp\{-H_0(t)\}]^c = \{S_0(t)\}^c$$

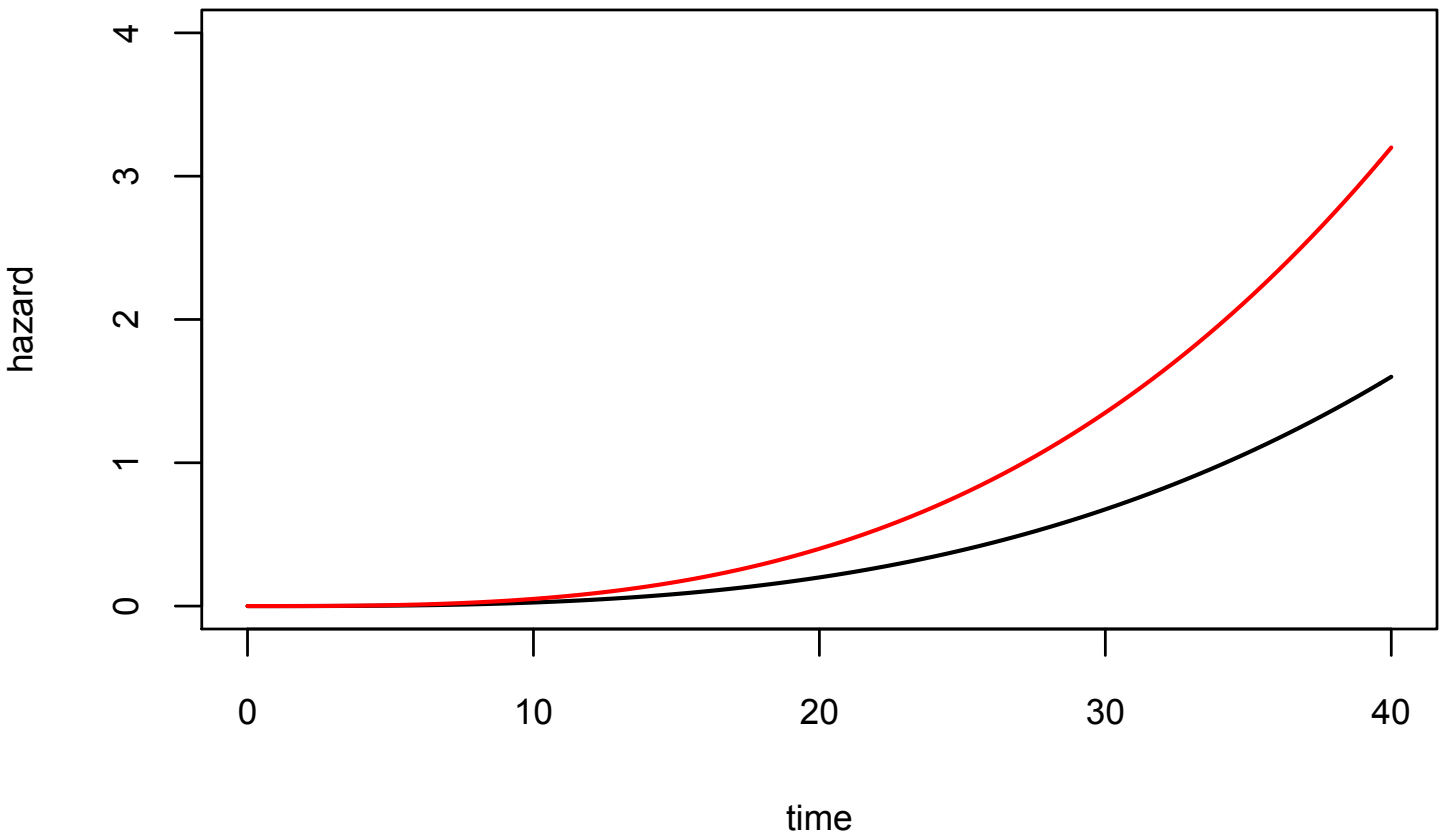
Raising $S_0(t)$ to the power c will make it **bigger** if $c < 1$ and **smaller** if $c > 1$, as expected in view of the previous slide.

Mathematically, why is that?

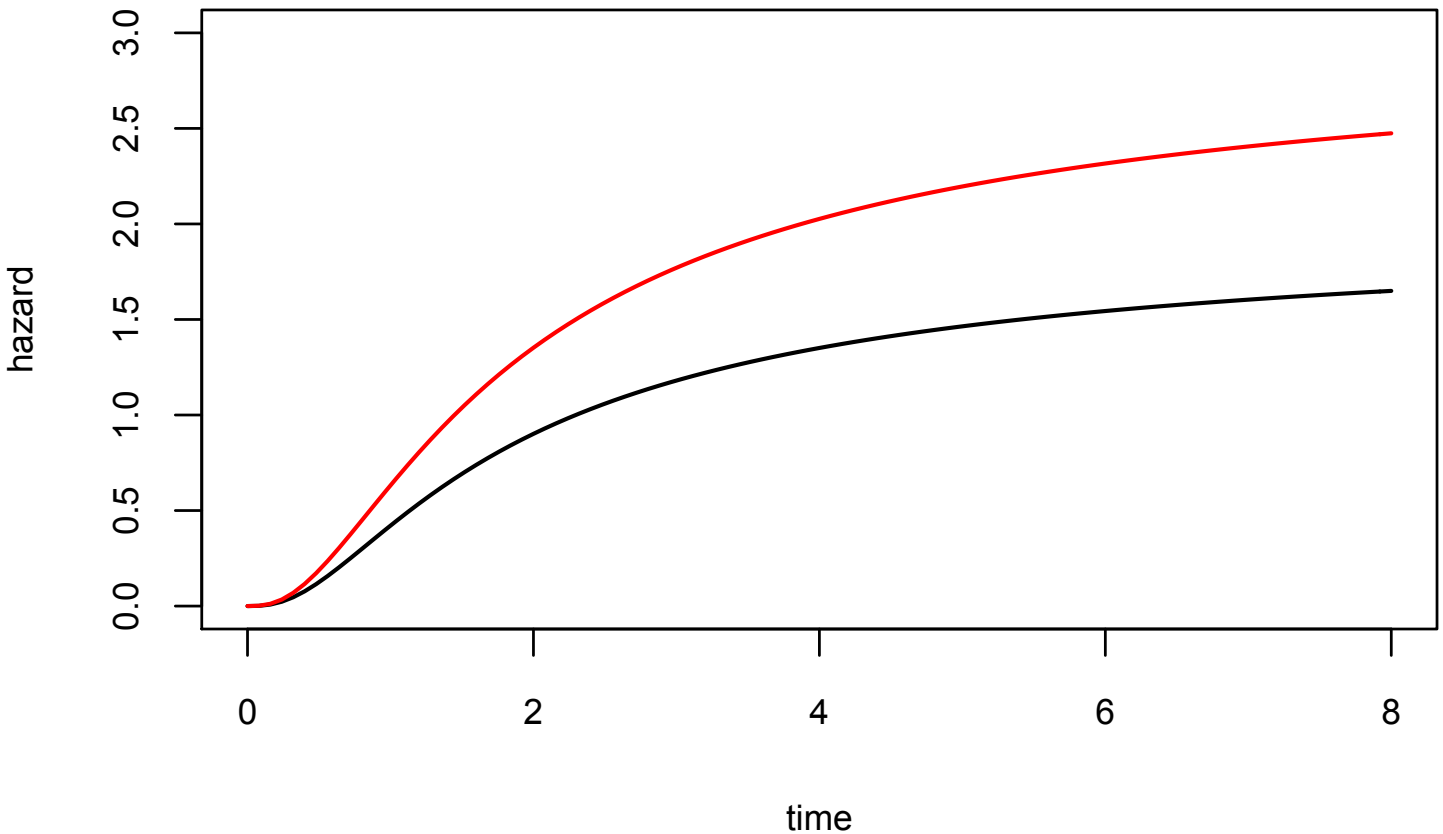
.....

In particular, this implies that if **two survival curves cross each other**, they cannot possibly have proportional hazard functions.

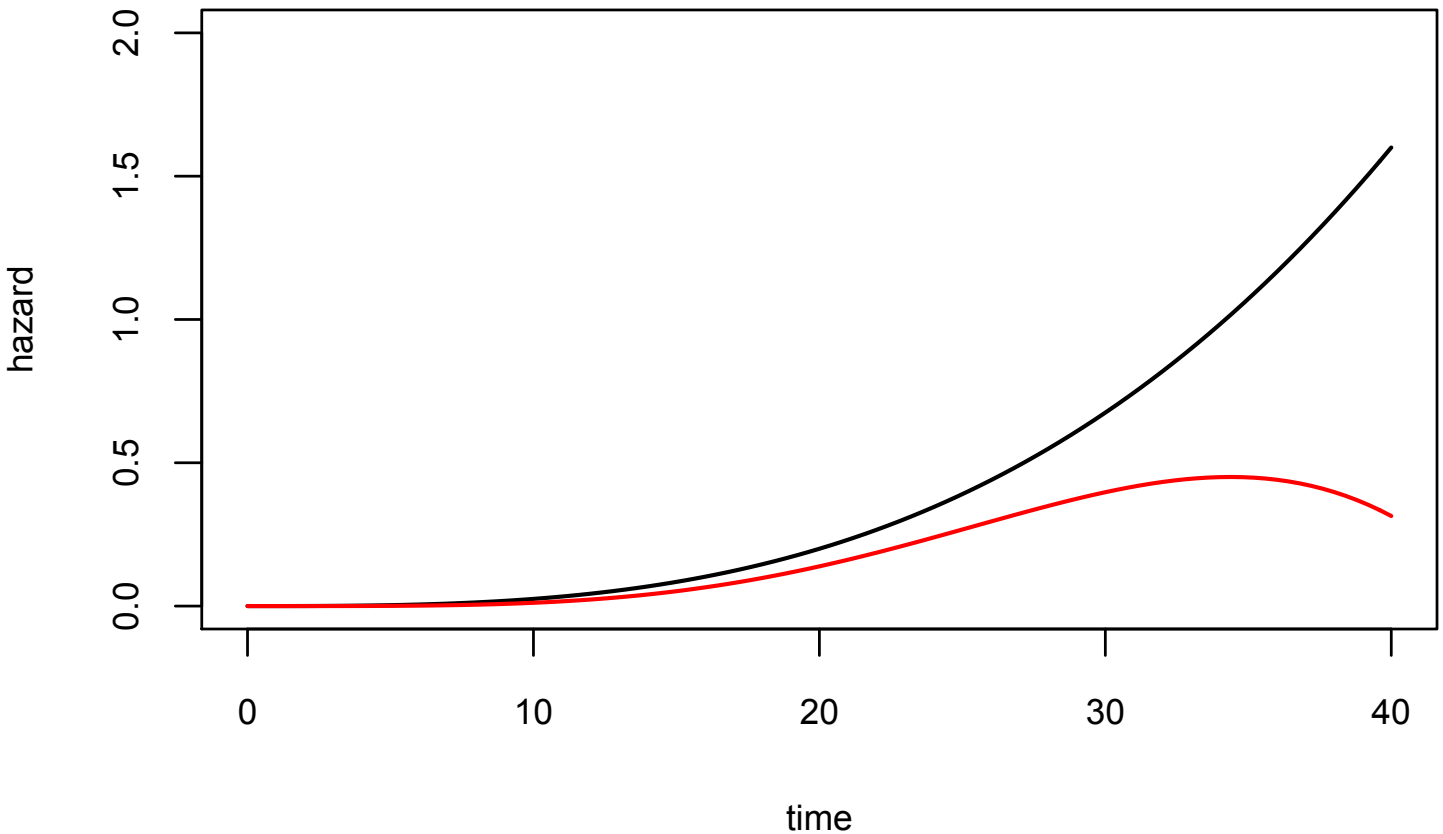
Can this pair of curves satisfy the proportional hazards model?



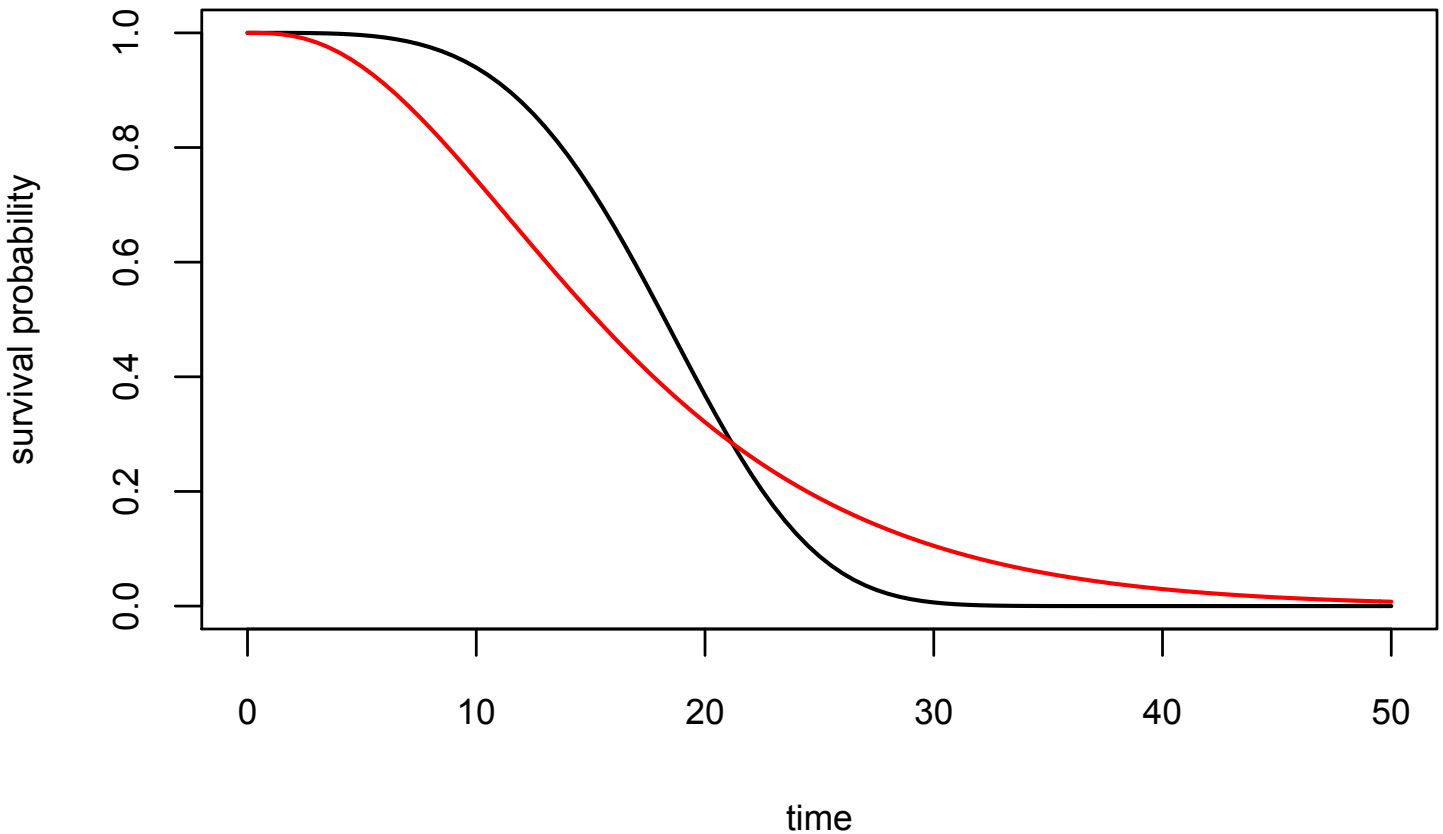
Can this pair of curves satisfy the proportional hazards model?



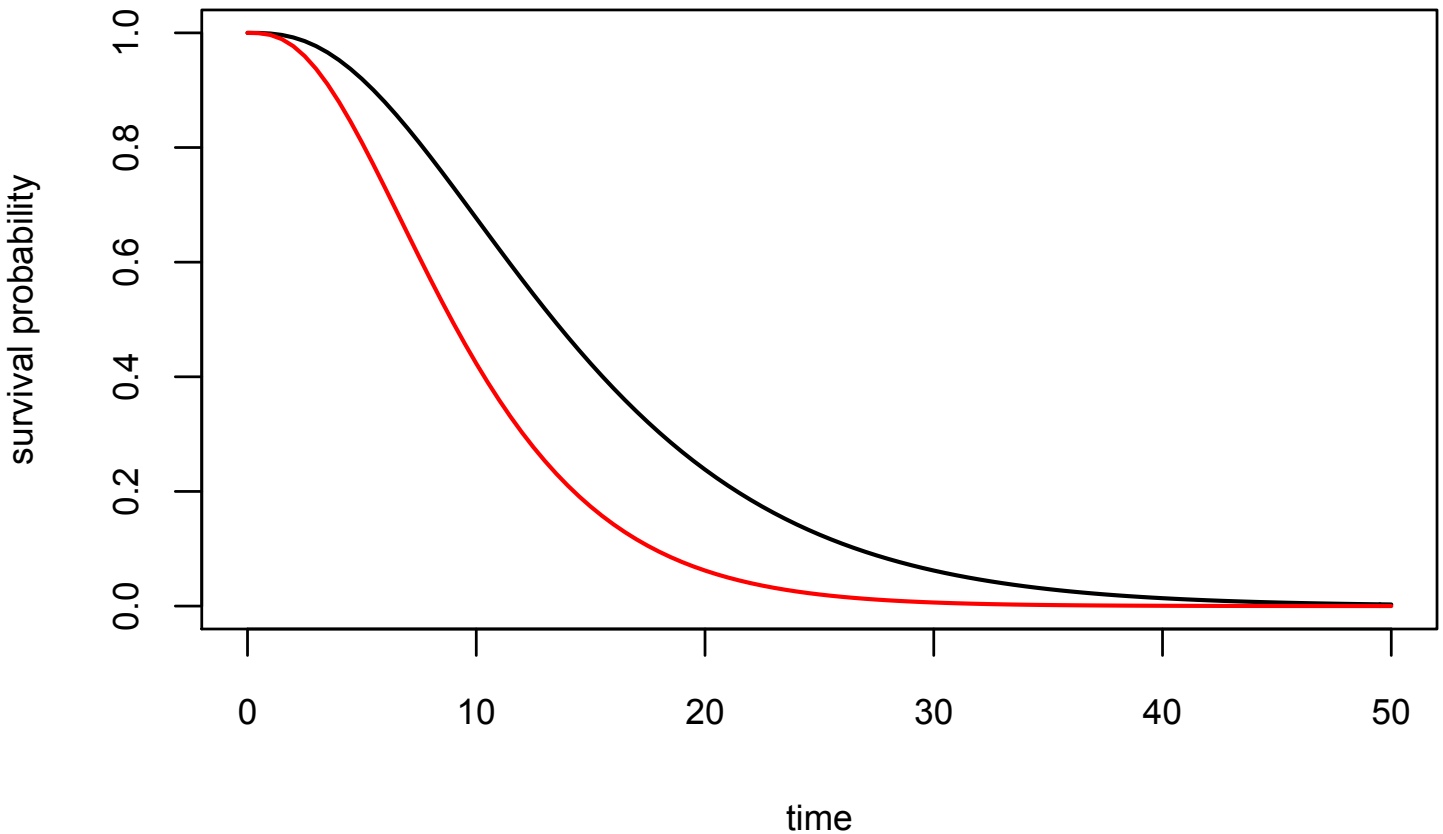
Can this pair of curves satisfy the proportional hazards model?



Can this pair of curves satisfy the proportional hazards model?



Can this pair of curves satisfy the proportional hazards model?



Formulation, properties and parameter interpretation

How is this useful? In practice, we have more than two subgroups!

Suppose that $Z = (Z_1, Z_2, \ldots, Z_q)$ is the covariate vector of interest.

Denoting the hazard function of T given $Z_1 = z_1, Z_2 = z_2, \ldots, Z_q = z_q$ at time t by $h(t \mid z_1, z_2, \ldots, z_q)$, the basic **proportional hazards model** states that

$$h(t \mid z_1, z_2, \ldots, z_q) = h_0(t) e^{\beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_q z_q}$$

for each t, where h_0 is an unspecified hazard function.

Since a distribution is completely specified by its hazard function, this model boils down the distribution of T given Z to

1. the vector of coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_q)$ – what is of interest;
2. a hazard function h_0 – often not of interest (a nuisance, in statistical parlance).

Without further assumptions, this is therefore a so-called **semiparametric model**.

Formulation, properties and parameter interpretation

As in its simpler two-sample form, the premise of this model is that

the hazard ratio comparing any two subgroups is constant over time.

Take two covariate profiles, say $(z_1, z_2, \ldots, z_q)$ and $(\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_q)$, then we have that

$$\begin{aligned} \frac{h(t \mid \tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_q)}{h(t \mid z_1, z_2, \ldots, z_q)} &= \frac{h_0(t) e^{\beta_1 \tilde{z}_1 + \beta_2 \tilde{z}_2 + \ldots + \beta_q \tilde{z}_q}}{h_0(t) e^{\beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_q z_q}} \\ &= e^{\beta_1(\tilde{z}_1 - z_1) + \beta_2(\tilde{z}_2 - z_2) + \ldots + \beta_q(\tilde{z}_q - z_q)} \neq \text{ function of time } t \ . \end{aligned}$$

If the covariate profiles $\tilde{z}$ and z differ only in one component – say the first – then

$$HR_{\tilde{z},z}(t) := \frac{h(t \mid \tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_q)}{h(t \mid z_1, z_2, \ldots, z_q)} = e^{\beta_1(\tilde{z}_1 - z_1)} \ .$$

Formulation, properties and parameter interpretation

Suppose that Z_1 is a binary variable (e.g., $Z_1 = 1$ for treated and $Z_1 = 0$ for controls).

$$\begin{aligned} HR(t) &:= \frac{h(t \mid 1, z_2, \ldots, z_p)}{h(t \mid 0, z_2, \ldots, z_p)} = e^{\beta_1 (1-0)} = e^{\beta_1} \\ \log HR(t) &= \beta_1 \end{aligned}$$

e^{β_1} = hazard ratio comparing subgroups of treated and control patients that otherwise have the same characteristics as determined by all other covariates

= hazard ratio comparing treated patients to controls *adjusting for all other covariates*

We note that $HR < 1$ if $\beta_1 < 0$ and instead $HR > 1$ if $\beta_1 > 0$.

Formulation, properties and parameter interpretation

Suppose that Z_1 is a continuous variable (e.g., Z_1 = systolic blood pressure).

$$HR(t) := \frac{h(t \mid z_1 + 1, z_2, \ldots, z_p)}{h(t \mid z_1, z_2, \ldots, z_p)} = e^{\beta_1 (z_1 + 1 - z_1)} = e^{\beta_1}$$

$$\log HR(t) = \beta_1$$

e^{β_1} = hazard ratio comparing subgroups of patients whose SBP differs by one unit but otherwise have the same characteristics as determined by all other covariates

= hazard ratio comparing patients with a certain SBP to other patients with SBP lower by one unit *adjusting for all other covariates*

Again, we note that $HR < 1$ if $\beta_1 < 0$ and instead $HR > 1$ if $\beta_1 > 0$.

A few examples of simple models involving categorical groupings...

Example 1: (comparing two subgroups)

Z = binary group indicator (i.e., $Z = 1$ if in subgroup 1 and $= 0$ otherwise)

$$h(t \mid z) = h_0(t)e^{\beta z}$$

$$\begin{aligned} e^{\beta} &= \frac{h(t \mid 1)}{h(t \mid 0)} \\ &= \text{HR comparing patients in subgroup 1 to patients in subgroup 0} \\ h_0(t) &= \text{hazard rate at time } t \text{ for patients in subgroup 0} \end{aligned}$$

A few examples of simple models involving categorical groupings...

Example 2: (comparing $r + 1 > 2$ subgroups)

Z = categorical group indicator (i.e., $Z = j$ if in subgroup j, $j = 0, 1, \ldots, r$)

$$h(t \mid z) = h_0(t)e^{\beta z}$$

$$\begin{aligned} e^{\beta} &= \frac{h(t \mid j+1)}{h(t \mid j)} \\ &= \text{HR comparing patients in subgroup } j+1 \text{ to patients in subgroup } j \\ h_0(t) &= \text{hazard rate at time } t \text{ for patients in subgroup 0} \end{aligned}$$

CAUTION: Can the hazards be expected to vary between subgroups like this?!!!

A few examples of simple models involving categorical groupings...

Example 3: (comparing $r + 1 > 2$ subgroups)

Z_j = group indicator (i.e., $Z_j = 1$ if in subgroup j and $= 0$ otherwise), $j = 1, 2, \ldots, r$

$$h(t \mid z_1, z_2, \ldots, z_r) = h_0(t) e^{\beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_r z_r}$$

$$e^{\beta_1} = \frac{h(t \mid 1, 0, 0, \ldots, 0)}{h(t \mid 0, 0, 0, \ldots, 0)} = \text{HR comparing patients in subgroup 1 versus subgroup 0}$$

$$e^{\beta_2} = \frac{h(t \mid 0, 1, 0, \ldots, 0)}{h(t \mid 0, 0, 0, \ldots, 0)} = \text{HR comparing patients in subgroup 2 versus subgroup 0}$$

$\ldots$

$h_0(t)$ = hazard rate at time t for patients in subgroup 0

What is the trade-off relative to Example 2? .....

.....

Formulation, properties and parameter interpretation

As it turns out, the **logrank test** can be framed as a **score test of the null hypothesis of no association** in these simple models!

Example 1: binary group indicator Z

model: $h(t \mid z) = h_0(t)e^{\beta z}$
score test of $\mathcal{H}_0 : \beta = 0 \rightarrow$ logrank test

Example 2: categorical group indicator Z coded as single variable

model: $h(t \mid z) = h_0(t)e^{\beta z}$
score test of $\mathcal{H}_0 : \beta = 0 \rightarrow$ Tarone's trend test

Example 3: categorical group indicator Z coded as binary indicators $Z_1, Z_2, \ldots, Z_r$

model: $h(t \mid z_1, z_2, \ldots, z_r) = h_0(t)e^{\beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_r z_r}$
score test of $\mathcal{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_r = 0 \rightarrow$ multigroup logrank test

Why should this not be surprising? When are the logrank tests most powerful?

What about interactions?

Suppose that $Z = (Z_1, Z_2)$ and that both Z_1 and Z_2 are binary variables.

Define the following Z_2-specific hazard ratios comparing levels of Z_1:

$$HR_{Z_1:Z_2=1}(t) := \frac{h(t \mid 1,1)}{h(t \mid 0,1)} \quad \text{and} \quad HR_{Z_1:Z_2=0}(t) := \frac{h(t \mid 1,0)}{h(t \mid 0,0)} \ .$$

Model w/o interaction: $h(t \mid z_1, z_2) = h_0(t)e^{\beta_1 z_1 + \beta_2 z_2}$

The hazard ratio comparing levels of Z_1 is the same across levels of Z_2, since

$$HR_{Z_1:Z_2=1}(t) = e^{\beta_1} = HR_{Z_1:Z_2=0}(t) \ .$$

Model w/ interaction: $h(t \mid z_1, z_2) = h_0(t)e^{\beta_1 z_1 + \beta_2 z_2 + \gamma z_1 z_2}$

The hazard ratio comparing levels of Z_1 differs across levels of Z_2, and in fact,

$$HR_{Z_1:Z_2=1}(t) = \frac{h(t \mid 1,1)}{h(t \mid 0,1)} = e^{\beta_1+\gamma} \neq e^{\beta_1} = \frac{h(t \mid 1,0)}{h(t \mid 0,0)} = HR_{Z_1:Z_2=0}(t) \ .$$

How hazard ratios for Z_1 vary across levels of Z_2 is captured by the parameter γ:

$$\frac{HR_{Z_1:Z_2=1}(t)}{HR_{Z_1:Z_2=0}(t)} = \frac{e^{\beta_1+\gamma}}{e^{\beta_1}} = e^{\gamma} \ .$$

If, for example, in the above model we have that $e^{\gamma} = 2$, we conclude that...

...the hazard ratio comparing individuals in subgroup $Z_1 = 1$ to those in subgroup $Z_1 = 0$ is twice as large in the subpopulation of individuals in subgroup $Z_2 = 1$ compared to individuals in subgroup $Z_2 = 0$.

Formulation, properties and parameter interpretation

We have focused on the regression coefficients so far – rightfully so since they tell us about the association we are studying.

What about the so-called **baseline hazard function** h_0?

$$\begin{aligned} h_0(t) &= h(t \mid 0, 0, \ldots, 0) \\ &= \text{hazard function at time } t \text{ in the subgroup of individuals with} \\ &\quad \text{covariate values } Z_1 = Z_2 = \ldots = Z_r = 0 \end{aligned}$$

By recentering covariates, we may render the baseline hazard's interpretation more meaningful.

The baseline hazard function provides an **absolute** baseline risk level for the population but no information about how covariates are associated to the hazard.

The regression coefficients provide a **relative** association measure capturing how covariates relate to the hazard but no information about absolute risk.

Parametric proportional hazards model

Once we select a model of interest, we need to estimate its parameters. Why not find the parameter values that **maximize the likelihood function**?

- contribution of uncensored individuals

$$= f(y_i \mid z_i) = h(y_i \mid z_i) S(y_i \mid z_i)$$
$$= h(y_i \mid z_i) e^{-H(y_i|z_i)} = h_0(y_i) \exp\{\beta z_i - e^{\beta z_i} H_0(y_i)\}$$

- contribution of censored individuals

$$= S(y_i \mid z_i) = e^{-H(y_i|z_i)} = \exp\{-e^{\beta z_i} H_0(y_i)\}$$

So, the likelihood function is given by

$$\begin{aligned} L_n(\beta, h_0) &:= \prod_{i=1}^{n} \left[f(y_i \mid z_i)^{\delta_i} S(y_i \mid z_i)^{1-\delta_i} \right] \\ &= \prod_{i=1}^{n} \left[h_0(y_i)^{\delta_i} \exp\left\{ \delta_i \left[\beta z_i - e^{\beta z_i} H_0(y_i) \right] - (1-\delta_i) e^{\beta z_i} H_0(y_i) \right\} \right] \\ &= \prod_{i=1}^{n} \left[h_0(y_i)^{\delta_i} \exp\left\{ \delta_i \beta z_i - e^{\beta z_i} H_0(y_i) \right\} \right] . \end{aligned}$$

Parametric proportional hazards model

Here, we need to optimize the likelihood over β values and h_0 values (i.e., whole functions) – not a straightforward problem at all!

Why not assume a **parametric model for the baseline hazard** h_0?

For example, we may consider a variety of models for the true baseline hazard function:

exponential model:	$h_0(t) = \lambda$, with λ unknown
Weibull model:	$h_0(t) = p\lambda^{p-1}t^p$, with λ, p unknown
gamma model:	no closed form
generalized gamma model:	no closed form
piecewise exponential model:	$h_0(t) = \sum_{k=1}^{m} \lambda_k I(u_{k-1} < t \leqslant u_k)$, with $\lambda_1, \lambda_2, \ldots, \lambda_k$ unknown and $u_0 := 0 < u_1 < u_2 < \ldots < u_m := +\infty$ specified

The likelihood function then only depends on a relatively small number of unknowns. It can therefore be maximized using standard optimization software.

Confidence intervals can be constructed in the usual way.

Semiparametric proportional hazards model

There are pros and cons to using parametric proportional hazards models:

- \+ they are simple to use in practice;
- \+ the MLE that we compute is asymptotically efficient;
- − the risk of model misspecification is greater because of the need to also specify a parametric model for the baseline distribution.

Surprisingly, we can easily do without this extra modelling assumption!

The likelihood involves both β and h_0, and it is difficult to use without parametric assumptions on h_0. However, the so-called partial likelihood function depends only on β and can serve as a likelihood function as well!

Semiparametric proportional hazards model

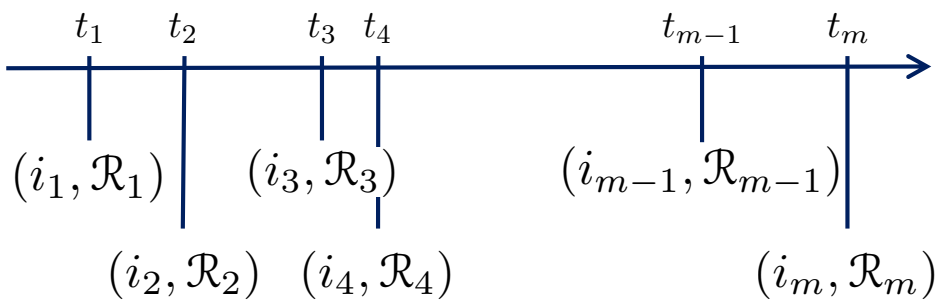
The data at our disposal consist of the triplets (y_i, δ_i, z_i) for each study participant. This gives rise to the usual likelihood.

This information allows us to write down, for each observed event time:

- which participants were at risk of experiencing the event (i.e., the risk set);
- which participant experienced the event.

Suppose that m events occur, say at times $t_1 < t_2 < \ldots < t_m$.

Denote by i_k the index of the participant who experienced an event at time t_k, and by $\mathcal{R}_k$ the risk set at that time.



Semiparametric proportional hazards model

Suppose the data, written as (observation time, covariate), are

$$\{(5, z_1), (10, z_2), (2+, z_3), (6+, z_4), (7, z_5), (15, z_6), (11+, z_7), (3, z_8)\} .$$

We can sort the data on the observed times and tabulate them as

patient index	3	8	1	4	5	2	7	6
observed time	2	3	5	6	7	10	11	15
event indicator	0	1	1	0	1	1	0	1
covariate	z_3	z_8	z_1	z_4	z_5	z_2	z_7	z_6

This helps us to fill the following table.

event time	3	5	7	10	15
event index					
risk set					

Semiparametric proportional hazards model

We observe that, at time $t = 3$,

- each of patients 1, 2, 4, 5, 6, 7 and 8 could have had an event;
- in reality, only patient 8 did.

What is the probability that patient 8 had an event given that only one out of patients 1, 2, 4, 5, 6, 7 and 8 had an event?

$P(\text{ patient 8 had an event} \mid \text{one out of patients 1,2,4,5,6,7,8 had an event})$

$$= \frac{h(3 \mid z_8)}{h(3 \mid z_1) + h(3 \mid z_2) + h(3 \mid z_4) + h(3 \mid z_5) + h(3 \mid z_6) + h(3 \mid z_7) + h(3 \mid z_8)}$$

$$= \frac{h_0(3)e^{\beta z_8}}{h_0(3)e^{\beta z_1} + h_0(3)e^{\beta z_2} + h_0(3)e^{\beta z_4} + h_0(3)e^{\beta z_5} + h_0(3)e^{\beta z_6} + h_0(3)e^{\beta z_7} + h_0(3)e^{\beta z_8}}$$

$$= \frac{e^{\beta z_8}}{e^{\beta z_1} + e^{\beta z_2} + e^{\beta z_4} + e^{\beta z_5} + e^{\beta z_6} + e^{\beta z_7} + e^{\beta z_8}}$$

Semiparametric proportional hazards model

We observe that, at time $t = 10$,

- each of patients 2, 6 and 7 could have had an event;
- in reality, only patient 2 did.

What is the probability that patient 2 had an event given that only one out of patients 2, 6 and 7 had an event?

$$P(\text{ patient 2 had an event} \mid \text{one out of patients 2,6,7 had an event }) = \frac{h(10 \mid z_2)}{h(10 \mid z_2) + h(10 \mid z_6) + h(10 \mid z_7)} = \frac{h_0(10)e^{\beta z_2}}{h_0(10)e^{\beta z_2} + h_0(10)e^{\beta z_6} + h_0(10)e^{\beta z_7}} = \frac{e^{\beta z_2}}{e^{\beta z_2} + e^{\beta z_6} + e^{\beta z_7}}$$

Semiparametric proportional hazards model

We can do this at each observed event time (i.e., $t = 3, 5, 7, 10, 15$) and then multiply the resulting collection of probabilities.

$$
\begin{aligned}
PL_n(\beta) \;:=\; & \left(\frac{e^{\beta z_8}}{e^{\beta z_1} + e^{\beta z_2} + e^{\beta z_4} + e^{\beta z_5} + e^{\beta z_6} + e^{\beta z_7} + e^{\beta z_8}} \right) && t = 3 \\
& \times \left(\frac{e^{\beta z_1}}{e^{\beta z_1} + e^{\beta z_2} + e^{\beta z_4} + e^{\beta z_5} + e^{\beta z_6} + e^{\beta z_7}} \right) && t = 5 \\
& \times \left(\frac{e^{\beta z_5}}{e^{\beta z_2} + e^{\beta z_5} + e^{\beta z_6} + e^{\beta z_7}} \right) && t = 7 \\
& \times \left(\frac{e^{\beta z_2}}{e^{\beta z_2} + e^{\beta z_6} + e^{\beta z_7}} \right) && t = 10 \\
& \times \left(\frac{e^{\beta z_6}}{e^{\beta z_6}} \right) && t = 15
\end{aligned}
$$

This is the **partial likelihood** introduced by Sir David Cox in 1972.

Formulation, properties and parameter interpretation

Regression Models and Life-Tables

By D. R. Cox

Imperial College, London

[Read before the Royal Statistical Society, at a meeting organized by the Research Section, on Wednesday, March 8th, 1972, Mr M. J. R. Healy in the Chair]

Summary

The analysis of censored failure times is considered. It is assumed that on each individual are available values of one or more explanatory variables. The hazard function (age-specific failure rate) is taken to be a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time. A conditional likelihood is obtained, leading to inferences about the unknown regression coefficients. Some generalizations are outlined.

Keywords: LIFE TABLE; HAZARD FUNCTION; AGE-SPECIFIC FAILURE RATE; PRODUCT LIMIT ESTIMATE; REGRESSION; CONDITIONAL INFERENCE; ASYMPTOTIC THEORY; CENSORED DATA; TWO-SAMPLE RANK TESTS; MEDICAL APPLICATIONS; RELIABILITY THEORY; ACCELERATED LIFE TESTS.

Sir David Cox
at a symposium for his 90th birthday

Sir David Cox
at a symposium for his 90th birthday

Semiparametric proportional hazards model

In general, for a proportional hazards model, the partial likelihood is given by

$$PL_n(\beta) := \prod_k \left[\frac{HR_{z_{i_k}}(\beta)}{\sum_{j \in \mathcal{R}_k} HR_{z_{i_j}}(\beta)} \right],$$

where $HR_z(\beta) =$ hazard ratio comparing patients with covariate z to baseline group.

Some observations about the partial likelihood:

- unlike the likelihood, the partial likelihood only involves what is of interest, β;
- in the absence of censoring, the partial likelihood is the marginal likelihood based on the observation ranks conditional on covariates;
- surprisingly, there is only very little loss of information from ignoring the actual values of the observed times beyond their ranks;
- the partial likelihood can be treated just like a regular likelihood:
  - the maximizer of the partial likelihood is an estimator of the true value of β;
  - the inverse of the observed information matrix can be used to gauge uncertainty;
  - we can perform Wald, score and likelihood ratio tests as usual.

Semiparametric proportional hazards model

A marginal Cox analysis of the herpes data...

```
herpes = read.csv("herpes.csv")
s.herpes = with(herpes, Surv(timetorec, event))

herpes.cox.1 = coxph(s.herpes ~ male, data=herpes)
summary(herpes.cox.1)
```

```
## Call:
## coxph(formula = s.herpes ~ male, data = herpes)
##
##   n= 456, number of events= 372
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## male 0.1410    1.1515   0.1093 1.291    0.197
##
##      exp(coef) exp(-coef) lower .95 upper .95
## male     1.151     0.8684    0.9295     1.427
##
## Concordance= 0.51  (se = 0.014 )
## Rsquare= 0.004   (max possible= 1 )
## Likelihood ratio test= 1.64  on 1 df,   p=0.2002
## Wald test            = 1.67  on 1 df,   p=0.1968
## Score (logrank) test = 1.67  on 1 df,   p=0.1965
```

Semiparametric proportional hazards model

```
herpes.cox.2 = coxph(s.herpes ~ type, data=herpes)
summary(herpes.cox.2)

## Call:
## coxph(formula = s.herpes ~ type, data = herpes)
##
##   n= 456, number of events= 372
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## type 0.70757   2.02906  0.09318 7.594  3.1e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## type     2.029     0.4928      1.69     2.436
##
## Concordance= 0.599  (se = 0.014 )
## Rsquare= 0.118   (max possible= 1 )
## Likelihood ratio test= 57.37  on 1 df,   p=3.608e-14
## Wald test            = 57.67  on 1 df,   p=3.098e-14
## Score (logrank) test = 57.54  on 1 df,   p=3.308e-14
```

Semiparametric proportional hazards model

```
herpes$type. = as.factor(herpes$type)
herpes.cox.3 = coxph(s.herpes ~ type., data=herpes)
summary(herpes.cox.3)
```

```
## Call:
## coxph(formula = s.herpes ~ type., data = herpes)
##
##   n= 456, number of events= 372
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## type.2 1.1511    3.1618   0.1786 6.446 1.15e-10 ***
## type.3 1.5238    4.5896   0.2168 7.029 2.08e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## type.2     3.162     0.3163     2.228     4.487
## type.3     4.590     0.2179     3.001     7.019
##
## Concordance= 0.599  (se = 0.014 )
## Rsquare= 0.139   (max possible= 1 )
## Likelihood ratio test= 68.07  on 2 df,   p=1.665e-15
## Wald test            = 52.93  on 2 df,   p=3.206e-12
## Score (logrank) test = 59.09  on 2 df,   p=1.475e-13
```

Semiparametric proportional hazards model

```
herpes.cox.4 = coxph(s.herpes ~ treat, data=herpes)
summary(herpes.cox.4)

## Call:
## coxph(formula = s.herpes ~ treat, data = herpes)
##
##   n= 456, number of events= 372
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## treat 0.005313  1.005327 0.052546 0.101    0.919
##
##       exp(coef) exp(-coef) lower .95 upper .95
## treat     1.005     0.9947    0.9069     1.114
##
## Concordance= 0.504  (se = 0.015 )
## Rsquare= 0   (max possible= 1 )
## Likelihood ratio test= 0.01  on 1 df,   p=0.9195
## Wald test            = 0.01  on 1 df,   p=0.9195
## Score (logrank) test = 0.01  on 1 df,   p=0.9195
```

Semiparametric proportional hazards model

```
herpes$treat. = as.factor(herpes$treat)
herpes.cox.5 = coxph(s.herpes ~ treat., data=herpes)
summary(herpes.cox.5)
```

```
## Call:
## coxph(formula = s.herpes ~ treat., data = herpes)
##
##   n= 456, number of events= 372
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## treat.1 -0.2467    0.7814   0.1947 -1.267    0.205
## treat.2  0.1473    1.1587   0.1164  1.266    0.206
## treat.3 -0.5101    0.6005   0.3234 -1.577    0.115
##
##         exp(coef) exp(-coef) lower .95 upper .95
## treat.1    0.7814      1.280    0.5335     1.145
## treat.2    1.1587      0.863    0.9223     1.456
## treat.3    0.6005      1.665    0.3186     1.132
##
## Concordance= 0.538  (se = 0.015 )
## Rsquare= 0.016   (max possible= 1 )
## Likelihood ratio test= 7.37  on 3 df,   p=0.0611
## Wald test            = 6.78  on 3 df,   p=0.07928
## Score (logrank) test = 6.89  on 3 df,   p=0.07563
```

Semiparametric proportional hazards model

```
herpes$treat.any = as.numeric(herpes$treat>0)
herpes.cox.6 = coxph(s.herpes ~ treat.any, data=herpes)
summary(herpes.cox.6)
```

```
## Call:
## coxph(formula = s.herpes ~ treat.any, data = herpes)
##
##   n= 456, number of events= 372
##
##               coef exp(coef) se(coef)    z Pr(>|z|)
## treat.any 0.001049  1.001049 0.105607 0.01    0.992
##
##           exp(coef) exp(-coef) lower .95 upper .95
## treat.any     1.001      0.999    0.8139     1.231
##
## Concordance= 0.503  (se = 0.014 )
## Rsquare= 0   (max possible= 1 )
## Likelihood ratio test= 0  on 1 df,   p=0.9921
## Wald test            = 0  on 1 df,   p=0.9921
## Score (logrank) test = 0  on 1 df,   p=0.9921
```

Semiparametric proportional hazards model

```
herpes.cox.7 = coxph(s.herpes ~ duration, data=herpes)
summary(herpes.cox.7)

## Call:
## coxph(formula = s.herpes ~ duration, data = herpes)
##
##   n= 456, number of events= 372
##
##              coef exp(coef) se(coef)     z Pr(>|z|)
## duration 0.022159  1.022407 0.006706 3.304 0.000952 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## duration     1.022     0.9781     1.009     1.036
##
## Concordance= 0.552  (se = 0.017 )
## Rsquare= 0.022   (max possible= 1 )
## Likelihood ratio test= 10.37  on 1 df,   p=0.001283
## Wald test            = 10.92  on 1 df,   p=0.0009515
## Score (logrank) test = 10.92  on 1 df,   p=0.0009494
```

Semiparametric proportional hazards model

And now a joint Cox analysis of the herpes data...

```
herpes.cox.joint = coxph(s.herpes ~ male + type. + treat. + duration, data=herpes)
summary(herpes.cox.joint)$coef
```

```
##                 coef exp(coef)    se(coef)          z     Pr(>|z|)
## male       0.22852298 1.2567424 0.111383094  2.0516846 4.020033e-02
## type.2     1.23614875 3.4423306 0.182607471  6.7694314 1.292899e-11
## type.3     1.57593310 4.8352513 0.218351309  7.2174200 5.297984e-13
## treat.1   -0.03503918 0.9655676 0.196958056 -0.1779017 8.588002e-01
## treat.2    0.38977598 1.4766500 0.124006792  3.1431825 1.671216e-03
## treat.3   -0.26651358 0.7660456 0.326019709 -0.8174769 4.136559e-01
## duration   0.02593113 1.0262703 0.006896099  3.7602614 1.697359e-04
```

Semiparametric proportional hazards model

```
herpes.cox.joint.int = coxph(s.herpes ~ male + type.*treat. + duration, data=herpes)

## Warning in coxph(s.herpes ~ male + type.  * treat.  + duration, data = herpes):  X matrix
deemed to be singular; variable 13

summary(herpes.cox.joint.int)$coef

##                       coef exp(coef)    se(coef)           z     Pr(>|z|)
## male             0.21523104 1.2401484 0.112424796  1.91444460 5.556337e-02
## type.2           1.24559382 3.4749977 0.279560605  4.45554130 8.368178e-06
## type.3           1.65306887 5.2229839 0.327189703  5.05232547 4.364631e-07
## treat.1          0.01006587 1.0101167 0.567348976  0.01774194 9.858447e-01
## treat.2          0.42197861 1.5249759 0.361077178  1.16866596 2.425382e-01
## treat.3         -0.39411156 0.6742788 1.035814767 -0.38048459 7.035857e-01
## duration         0.02491140 1.0252243 0.006991283  3.56320910 3.663486e-04
## type.2:treat.1  -0.12572931 0.8818535 0.614326604 -0.20466200 8.378362e-01
## type.3:treat.1   0.21353392 1.2380455 0.723646506  0.29508043 7.679324e-01
## type.2:treat.2   0.01323508 1.0133231 0.388561993  0.03406171 9.728279e-01
## type.3:treat.2  -0.25265523 0.7767356 0.470192615 -0.53734411 5.910299e-01
## type.2:treat.3   0.14592257 1.1571066 1.090844360  0.13377029 8.935842e-01
## type.3:treat.3          NA        NA 0.000000000          NA           NA
```

Semiparametric proportional hazards model

Why is there no hazard ratio estimate for the type=3/treat=3 subgroup?

```
table(herpes$treat, herpes$type, dnn = c("treat", "type"))
```

```
##      type
## treat   1   2   3
##     0  32 206  32
##     1   9  25   6
##     2  26  83  21
##     3   3  13   0
```

Semiparametric proportional hazards model

How do more complex models compare to simpler ones?

```
anova(herpes.cox.joint.int, herpes.cox.joint)
```

```
## Analysis of Deviance Table
##  Cox model: response is  s.herpes
##  Model 1: ~ male + type. * treat. + duration
##  Model 2: ~ male + type. + treat. + duration
##    loglik  Chisq Df P(>|Chi|)
## 1 -1969.1
## 2 -1969.9 1.4623  5    0.9174
```

```
anova(herpes.cox.joint, herpes.cox.5)
```

```
## Analysis of Deviance Table
##  Cox model: response is  s.herpes
##  Model 1: ~ male + type. + treat. + duration
##  Model 2: ~ treat.
##    loglik  Chisq Df P(>|Chi|)
## 1 -1969.9
## 2 -2013.0 86.271  4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Semiparametric proportional hazards model

A marginal Cox analysis of the myocardial infarction data...

```
trace = read.csv("trace.csv")
s.trace = with(trace, Surv(time, (status==9)))
trace.cox.1 = coxph(s.trace ~ wmi, data=trace)
summary(trace.cox.1)
```

```
## Call:
## coxph(formula = s.trace ~ wmi, data = trace)
##
##   n= 500, number of events= 259
##
##        coef exp(coef) se(coef)      z Pr(>|z|)
## wmi -1.2585    0.2841   0.1573 -7.999 1.22e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## wmi    0.2841       3.52    0.2087    0.3867
##
## Concordance= 0.645  (se = 0.018 )
## Rsquare= 0.119   (max possible= 0.998 )
## Likelihood ratio test= 63.26  on 1 df,   p=1.776e-15
## Wald test            = 63.98  on 1 df,   p=1.221e-15
## Score (logrank) test = 66.08  on 1 df,   p=4.441e-16
```

Semiparametric proportional hazards model

```
trace.cox.2 = coxph(s.trace ~ chf, data=trace)
summary(trace.cox.2)

## Call:
## coxph(formula = s.trace ~ chf, data = trace)
##
##   n= 500, number of events= 259
##
##       coef exp(coef) se(coef)    z Pr(>|z|)
## chf 1.0774    2.9370   0.1352 7.97 1.55e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## chf     2.937     0.3405     2.253     3.828
##
## Concordance= 0.634  (se = 0.016 )
## Rsquare= 0.13   (max possible= 0.998 )
## Likelihood ratio test= 69.71  on 1 df,   p=1.11e-16
## Wald test            = 63.52  on 1 df,   p=1.554e-15
## Score (logrank) test = 69.69  on 1 df,   p=1.11e-16
```

Semiparametric proportional hazards model

```
trace.cox.3 = coxph(s.trace ~ age, data=trace)
summary(trace.cox.3)

## Call:
## coxph(formula = s.trace ~ age, data = trace)
##
##   n= 500, number of events= 259
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## age 0.066692  1.068966 0.006584 10.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## age     1.069     0.9355     1.055     1.083
##
## Concordance= 0.693  (se = 0.019 )
## Rsquare= 0.209   (max possible= 0.998 )
## Likelihood ratio test= 117.4  on 1 df,   p=0
## Wald test            = 102.6  on 1 df,   p=0
## Score (logrank) test = 105.2  on 1 df,   p=0
```

Semiparametric proportional hazards model

```
trace.cox.4 = coxph(s.trace ~ sex, data=trace)
summary(trace.cox.4)

## Call:
## coxph(formula = s.trace ~ sex, data = trace)
##
##   n= 500, number of events= 259
##
##        coef exp(coef) se(coef)      z Pr(>|z|)
## sex -0.1937    0.8239   0.1296 -1.495    0.135
##
##     exp(coef) exp(-coef) lower .95 upper .95
## sex    0.8239      1.214    0.6392     1.062
##
## Concordance= 0.519  (se = 0.015 )
## Rsquare= 0.004   (max possible= 0.998 )
## Likelihood ratio test= 2.19  on 1 df,   p=0.1388
## Wald test            = 2.23  on 1 df,   p=0.135
## Score (logrank) test = 2.24  on 1 df,   p=0.1344
```

Semiparametric proportional hazards model

```
trace.cox.5 = coxph(s.trace ~ diabetes, data=trace)
summary(trace.cox.5)

## Call:
## coxph(formula = s.trace ~ diabetes, data = trace)
##
##   n= 500, number of events= 259
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## diabetes 0.4344    1.5441   0.1916 2.268   0.0234 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## diabetes     1.544     0.6476     1.061     2.248
##
## Concordance= 0.522  (se = 0.009 )
## Rsquare= 0.009   (max possible= 0.998 )
## Likelihood ratio test= 4.61  on 1 df,   p=0.03183
## Wald test            = 5.14  on 1 df,   p=0.02336
## Score (logrank) test = 5.22  on 1 df,   p=0.02229
```

Semiparametric proportional hazards model

```
trace.cox.6 = coxph(s.trace ~ vf, data=trace)
summary(trace.cox.6)

## Call:
## coxph(formula = s.trace ~ vf, data = trace)
##
##   n= 500, number of events= 259
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## vf 0.5050    1.6571   0.2445 2.066   0.0389 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##    exp(coef) exp(-coef) lower .95 upper .95
## vf     1.657     0.6035     1.026     2.676
##
## Concordance= 0.519  (se = 0.007 )
## Rsquare= 0.007   (max possible= 0.998 )
## Likelihood ratio test= 3.7  on 1 df,   p=0.05432
## Wald test            = 4.27  on 1 df,   p=0.03887
## Score (logrank) test = 4.36  on 1 df,   p=0.03684
```

Semiparametric proportional hazards model

A look at pairwise correlations between the various covariates of interest may be helpful in interpreting marginal associations.

```
corr.trace = with(trace, cor(cbind(wmi, chf, age, sex, diabetes, vf)))
corr.trace[upper.tri(corr.trace)] = NA
corr.trace
```

```
##                  wmi         chf          age          sex    diabetes vf
## wmi       1.00000000          NA           NA           NA          NA NA
## chf      -0.37464791  1.00000000           NA           NA          NA NA
## age      -0.19288208  0.34197944  1.00000000           NA          NA NA
## sex      -0.02354281 -0.13205952 -0.27356308  1.00000000          NA NA
## diabetes -0.13860246  0.08871325  0.04685823 -0.05523671  1.00000000 NA
## vf       -0.08966805  0.13467109 -0.03983927 -0.04760863 -0.05060615  1
```

Semiparametric proportional hazards model

```
trace.cox.joint.1 = coxph(s.trace ~ wmi + chf + age + sex + diabetes + vf, data=trace)
summary(trace.cox.joint.1)$coef
```

```
##                 coef exp(coef)    se(coef)         z     Pr(>|z|)
## wmi      -0.83153960 0.4353785 0.172622714 -4.817093 1.456651e-06
## chf       0.51742531 1.6777025 0.148890903  3.475198 5.104771e-04
## age       0.06633974 1.0685897 0.007240442  9.162389 0.000000e+00
## sex       0.25110266 1.2854420 0.136825912  1.835198 6.647633e-02
## diabetes  0.19942396 1.2206994 0.198349340  1.005418 3.146957e-01
## vf        0.85029120 2.3403282 0.252933771  3.361715 7.746009e-04
```

Semiparametric proportional hazards model

```
trace.cox.joint.2 = coxph(s.trace ~ wmi + age + sex + diabetes + chf*vf, data=trace)
summary(trace.cox.joint.2)$coef
```

```
##                  coef exp(coef)   se(coef)            z     Pr(>|z|)
## wmi       -0.836125885 0.4333863 0.17220731 -4.855345029 1.201773e-06
## age        0.067609123 1.0699470 0.00728456  9.281153716 0.000000e+00
## sex        0.274288883 1.3155948 0.13760516  1.993303734 4.622820e-02
## diabetes   0.206409989 1.2292571 0.19835759  1.040595376 2.980634e-01
## chf        0.465327418 1.5925355 0.15220713  3.057198634 2.234162e-03
## vf        -0.005717004 0.9942993 0.71830694 -0.007958998 9.936497e-01
## chf:vf     1.060843562 2.8888068 0.76915886  1.379225570 1.678252e-01
```

Semiparametric proportional hazards model

```
trace$chf_vf = trace$chf*trace$vf
trace$nochf_vf = (1-trace$chf)*trace$vf
trace.cox.joint.3 = coxph(s.trace ~ wmi + age + sex + diabetes +  chf_vf + nochf_vf + chf, data=trace)
summary(trace.cox.joint.3)$coef
```

```
##                  coef exp(coef)   se(coef)            z     Pr(>|z|)
## wmi      -0.836125885 0.4333863 0.17220731 -4.855345029 1.201773e-06
## age       0.067609123 1.0699470 0.00728456  9.281153716 0.000000e+00
## sex       0.274288883 1.3155948 0.13760516  1.993303734 4.622820e-02
## diabetes  0.206409989 1.2292571 0.19835759  1.040595376 2.980634e-01
## chf_vf    1.055126558 2.8723386 0.27549446  3.829937451 1.281758e-04
## nochf_vf -0.005717004 0.9942993 0.71830694 -0.007958998 9.936497e-01
## chf       0.465327418 1.5925355 0.15220713  3.057198634 2.234162e-03
```

Semiparametric proportional hazards model

We could have also used the `deltamethod` command to estimate the HR comparing patients in congestive heart failure with and without ventricular fibrillation (adjusting for other potential confounders).

```
loghr.est = coef(trace.cox.joint.3)[5]
loghr.se = sqrt(vcov(trace.cox.joint.3)[5,5])
exp(loghr.est+c(-1,0,1)*1.96*loghr.se)
```

```
## [1] 1.673902 2.872339 4.928801
```

```
loghr.est = sum(coef(trace.cox.joint.2)[c(6,7)])
loghr.se = deltamethod(g=~(x6+x7),mean=coef(trace.cox.joint.2),cov=vcov(trace.cox.joint.2))
exp(loghr.est+c(-1,0,1)*1.96*loghr.se)
```

```
## [1] 1.673902 2.872339 4.928801
```

Semiparametric proportional hazards model

Note that the numerator of the partial likelihood contributions hinges on a single event occurring at the event times.

In practice, ties occur!

Recall data from the 6-MP leukemia trial:

week	control	6-MP
1	× ×	
2	× ×	
3	×	
4	× ×	
5	× ×	
6		× × × ○
7		×
8	× × × ×	
9		○
10		× ○
11	× ×	○
12	× ×	
13		×
14		
15	×	
16		×
17	×	○
18		

week	control	6-MP
19		○
20		○
21		
22	×	×
23	×	×
24		
25		○
26		
27		
28		
29		
30		
31		
32		○ ○
33		
34		○
35		○

Semiparametric proportional hazards model

Why do we observe tied event times in practice?

(A) the event time random variable may truly have a discrete distribution;

(B) the coarseness of the measurement of event times leads to ties.

To be sensible, the proportional hazards model requires continuity of the underlying time-to-event distribution. Why?

$$h(t \mid z) = h_0(t)e^{\beta z}$$

In case (A), the proportional hazards model should be avoided. Instead, regression models for discrete survival data should be used (e.g., proportional odds model).

In case (B), the proportional hazards model is still appropriate, but not the usual partial likelihood, as it requires a single event at each observed time.

Several remedies have been proposed in the literature!

Semiparametric proportional hazards model

If the event time distribution is truly continuous, the underlying data – which we observed a coarsened version of – must not have had ties.

The latent ordering is unknown. We may **consider all possible scenarios** consistent with the observed data.

Suppose that at event time t two individuals, with covariates z_1 and z_2, experienced an event, and one more individual, with covariate z_3, was also at risk.

SCENARIO 1: (individual 1 first, then individual 2)

$$\text{partial likelihood contribution} = \left(\frac{e^{\beta z_1}}{e^{\beta z_1} + e^{\beta z_2} + e^{\beta z_3}} \right) \left(\frac{e^{\beta z_2}}{e^{\beta z_1} + e^{\beta z_2} + e^{\beta z_3} - e^{\beta z_1}} \right)$$

SCENARIO 2: (individual 2 first, then individual 1)

$$\text{partial likelihood contribution} = \left(\frac{e^{\beta z_2}}{e^{\beta z_1} + e^{\beta z_2} + e^{\beta z_3}} \right) \left(\frac{e^{\beta z_1}}{e^{\beta z_1} + e^{\beta z_2} + e^{\beta z_3} - e^{\beta z_2}} \right)$$

Semiparametric proportional hazards model

At event time t, the total contribution to the partial likelihood is the average of each contribution from a possible ordering:

$$\frac{1}{2}\left(\frac{e^{\beta z_1}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}}\right)\left(\frac{e^{\beta z_2}}{e^{\beta z_2}+e^{\beta z_3}}\right)+\frac{1}{2}\left(\frac{e^{\beta z_2}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}}\right)\left(\frac{e^{\beta z_1}}{e^{\beta z_1}+e^{\beta z_3}}\right).$$

At each event time with ties, this replaces the usual partial likelihood contribution.

This is the **exact marginal analysis method** (Kalbfleisch & Prentice, 1979).

- \+ it is a principled and effective extension of the partial likelihood to tied data;
- − it can be computationally very intensive:
  - if d individuals out of n individuals at risk had an event, there are $d!$ orderings to consider, each of which involve computing $d[n-(d-1)/2]$ hazard ratios;
  - in the 6MP example, at time $t=8$, $d=4$ and $n=28$: this yields 24 different orderings to consider, each of which involves 106 hazard ratios!!!

Semiparametric proportional hazards model

Several approximations have been proposed to reduce the computational burden of the exact method and these are still widely used.

Breslow (1975) method: perform exact method without depleting the risk set

$$\begin{aligned}
&\frac{1}{2}\left(\frac{e^{\beta z_1}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}}\right)\left(\frac{e^{\beta z_2}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}-e^{\beta z_1}}\right) \\
&\qquad\qquad + \frac{1}{2}\left(\frac{e^{\beta z_2}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}}\right)\left(\frac{e^{\beta z_1}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}-e^{\beta z_2}}\right) \\
\approx\ &\frac{1}{2}\left(\frac{e^{\beta z_1}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}}\right)\left(\frac{e^{\beta z_2}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}}\right)+\frac{1}{2}\left(\frac{e^{\beta z_2}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}}\right)\left(\frac{e^{\beta z_1}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}}\right) \\
=\ &\frac{e^{\beta(z_1+z_2)}}{(e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3})^2}
\end{aligned}$$

- if at each time d is small compared to the size n of the risk set, the excess hazard ratios included in the denominator will only produce a slight distortion;
- beware if d/n is not very small as this approach can lead to substantial bias.

Semiparametric proportional hazards model

Several approximations have been proposed to reduce the computational burden of the exact method and these are still widely used.

Efron (1977) method: use denominator averaged over all orderings

$$\frac{1}{2}\left(\frac{e^{\beta z_1}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}}\right)\left(\frac{e^{\beta z_2}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}-e^{\beta z_1}}\right)$$
$$+\frac{1}{2}\left(\frac{e^{\beta z_2}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}}\right)\left(\frac{e^{\beta z_1}}{e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}-e^{\beta z_2}}\right)$$
$$\approx \frac{e^{\beta(z_1+z_2)}}{(e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3})[e^{\beta z_1}+e^{\beta z_2}+e^{\beta z_3}-\frac{1}{2}(e^{\beta z_1}+e^{\beta z_2})]}$$

- this scheme is slightly more difficult to explain;
- however, it tends to perform quite well and it is relatively easy to implement.

Semiparametric proportional hazards model

How do these various approaches differ in the 6-MP dataset?

```
mp = read.csv("6mp.csv")
mp$tx = ifelse(mp$treat=="6-MP", 1, 0)
s.mp = with(mp, Surv(time, cens))
coxph(s.mp ~ tx, data=mp, ties="breslow")

## Call:
## coxph(formula = s.mp ~ tx, data = mp, ties = "breslow")
##
##       coef exp(coef) se(coef)     z       p
## tx -1.509     0.221    0.410 -3.68 0.00023
##
## Likelihood ratio test=15.2  on 1 df, p=9.61e-05
## n= 42, number of events= 30
```

Semiparametric proportional hazards model

How do these various approaches differ in the 6-MP dataset?

```
coxph(s.mp ~ tx, data=mp, ties="efron")

## Call:
## coxph(formula = s.mp ~ tx, data = mp, ties = "efron")
##
##       coef exp(coef) se(coef)     z       p
## tx -1.572     0.208    0.412 -3.81 0.00014
##
## Likelihood ratio test=16.4  on 1 df, p=5.26e-05
## n= 42, number of events= 30
```

Semiparametric proportional hazards model

How do these various approaches differ in the 6-MP dataset?

```
coxph(s.mp ~ tx, data=mp, ties="exact")

## Call:
## coxph(formula = s.mp ~ tx, data = mp, ties = "exact")
##
##       coef exp(coef) se(coef)     z       p
## tx -1.628     0.196    0.433 -3.76 0.00017
##
## Likelihood ratio test=16.2  on 1 df, p=5.54e-05
## n= 42, number of events= 30
```

Semiparametric proportional hazards model

So far, we have focused uniquely on estimation of the log-hazard ratio β, ignoring completely the baseline hazard function h_0.

When may estimating h_0 also be useful to us?

- to characterize in absolute terms the subgroup-specific risk of an event;
- to describe more interpretable summaries of subgroup-specific survivorship (e.g., survival probability, median survival time);
- to perform model checking.

As in the one-sample case, estimating the hazard function is difficult. The cumulative hazard function is an easier quantity to estimate. So, we will focus one estimating

$$H_0(t) := \int_0^t h_0(u)du \ .$$

Semiparametric proportional hazards model

Recall that, in the one-sample case, we can estimate the cumulative hazard function $H(t)$ by the Nelson-Aalen estimator

$$\widehat{H}(t) := \sum_{i:t_i \leqslant t} \frac{d_i}{n_i} \ ,$$

where d_i is the number of events occurring at t_i and n_i is the size of the risk set at t_i.

In the regression setting, we wish to estimate the **baseline cumulative hazard function**.

- If we had a large number of study participants in the baseline subgroup, we could use the Nelson-Aalen estimator applied on that subset.
- In most cases, we will have very few (if any) such partipants.
  Nevertheless, through the model structure, all individuals can provide information about the baseline subgroup.

Semiparametric proportional hazards model

The Breslow estimator of the baseline cumulative hazard function is

$$\widehat{H}_0(t) := \sum_{i:t_i \leqslant t} \frac{d_i}{n_i(\hat{\beta})} ,$$

where $n_i(\beta) := \sum_{j \in \mathcal{R}_i} e^{\beta z_j}$ is the size of the **adjusted risk set** at time t_i and $\hat{\beta}$ is the maximum partial likelihood estimator of β.

If all individuals have the same risk profile (i.e., $\beta = 0$), we have $n_i(\beta) = n_i(0) = n_i$.

Why do we need to adjust the risk set?

- Available individuals are from various subgroups, each of which may have a different survivorship compared to the baseline subgroup.
- We need to find the baseline subgroup equivalent of each member of the risk set.
  - $HR_j = e^{\beta z_j} > 1$: one high-risk individual = several baseline individuals;
  - $HR_j = e^{\beta z_j} < 1$: one low-risk individual = a fraction of a baseline individual.

Semiparametric proportional hazards model

Suppose that Z is a binary covariate (e.g., treatment vs control). Here, the baseline group is therefore the control group.

If n_{0i} and n_{1i} are, respectively, the number of controls and treated individuals in the risk set at time t_i, we can write

$$n_i(\beta) = n_{0i} + n_{1i} e^{\beta} \ .$$

- **Case I:** $\beta < 0$

  If the treated have longer durations, a sample of baseline individuals alone would have yielded a risk set with less than $n_{0i} + n_{1i}$;

- **Case II:** $\beta > 0$

  If the treated have shorter durations, a sample of baseline individuals alone would have yielded a risk set with more than $n_{0i} + n_{1i}$.

Semiparametric proportional hazards model

Once we have constructed an estimator of the baseline cumulative hazard function, we get an estimator of the subgroup-specific survival function for free.

Using the fact that

$$\begin{aligned} S(t \mid z_1, z_2, \ldots, z_q) &= \exp\{-H(t \mid z_1, z_2, \ldots, z_q)\} \\ &= \exp\{-\exp(\beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_q z_q) \cdot H_0(t)\} \ , \end{aligned}$$

we notice that

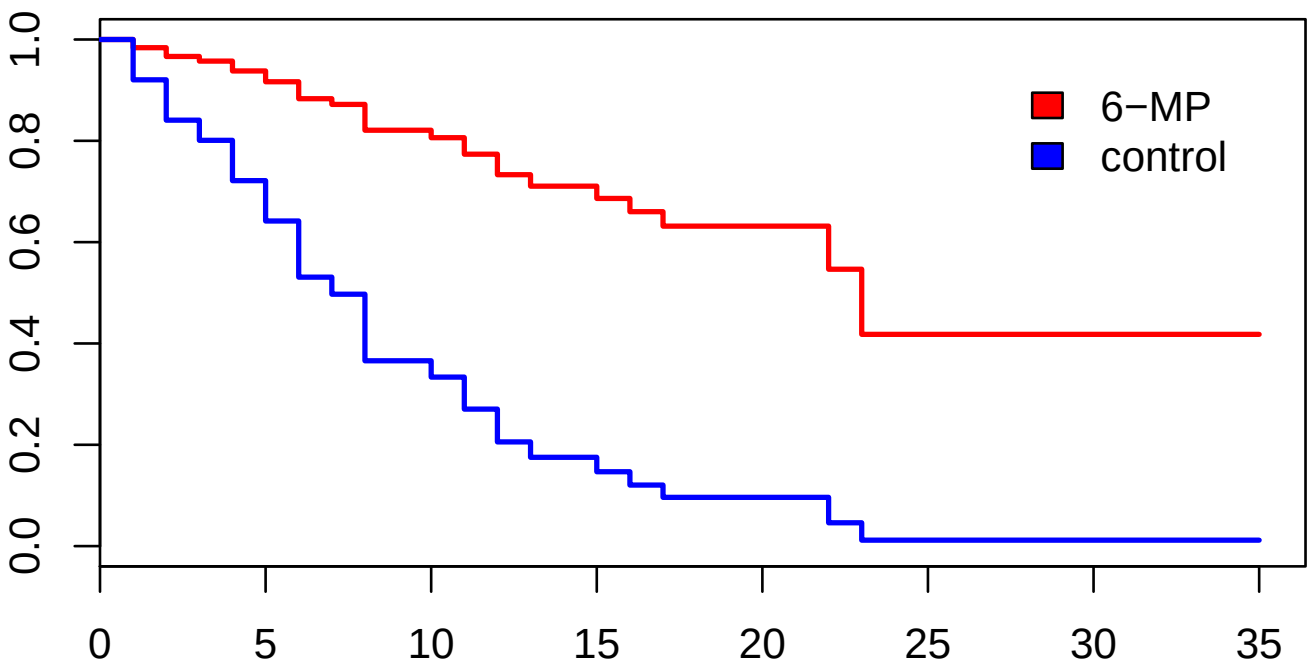
$$\widehat{S}(t \mid z_1, z_2, \ldots, z_q) = \exp\{-\exp(\hat{\beta}_1 z_1 + \hat{\beta}_2 z_2 + \ldots + \hat{\beta}_q z_q) \cdot \widehat{H}_0(t)\}$$

is an estimator of the subgroup-specific survival probability $S(t \mid z_{1,2}, \ldots, z_q)$.

Semiparametric proportional hazards model

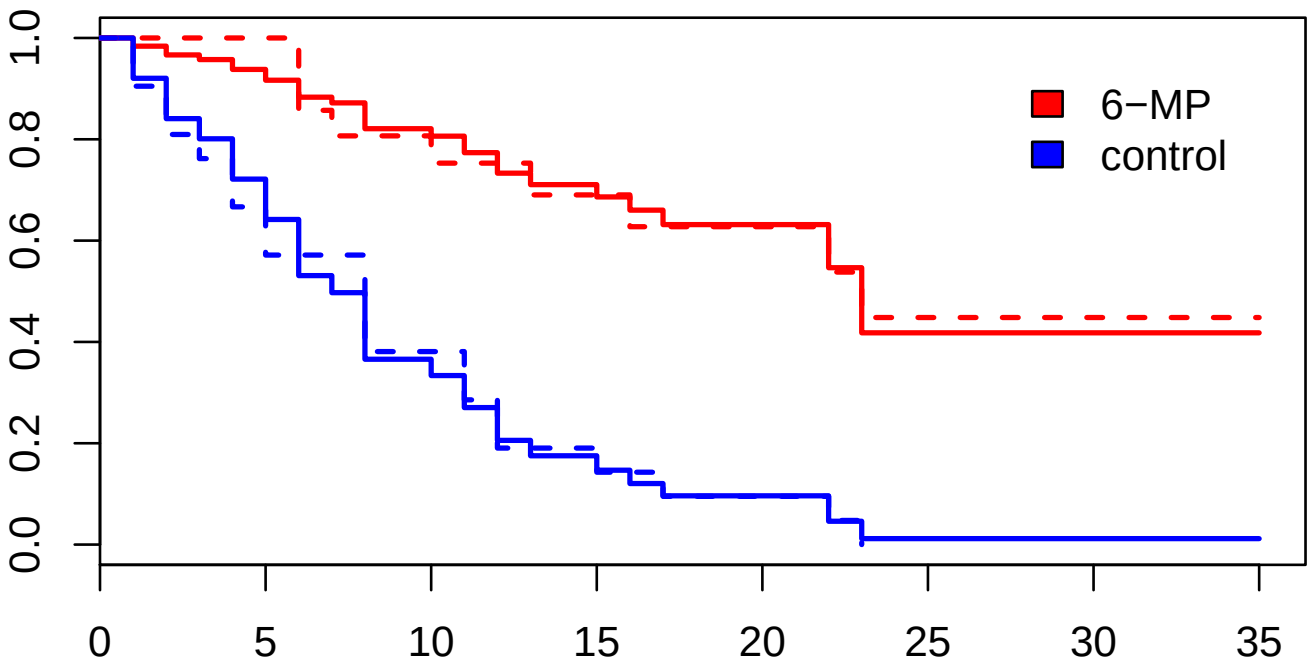
What are the estimated treatment group-specific survival functions of time until relapse in the 6MP study based on the proportional hazards model?

```
cox.mp = coxph(s.mp~tx, data=mp, ties="exact")
plot(survfit(cox.mp,newdata=data.frame(tx=1),conf.type="none"),col=2,lwd=2)
lines(survfit(cox.mp,newdata=data.frame(tx=0),conf.type="none"),col=4,lwd=2)
legend(27,0.97,fill=c(2,4),legend=c("6-MP","control"),bty='n')
```



Semiparametric proportional hazards model

```
cox.mp = coxph(s.mp~tx, data=mp, ties="exact")
plot(survfit(cox.mp,newdata=data.frame(tx=1),conf.type="none"),col=2,lwd=2)
lines(survfit(cox.mp,newdata=data.frame(tx=0),conf.type="none"),col=4,lwd=2)
lines(survfit(s.mp~1,subset=(mp$tx==1),conf.type="none"),col=2,lwd=2,lty=2)
lines(survfit(s.mp~1,subset=(mp$tx==0),conf.type="none"),col=4,lwd=2,lty=2)
legend(27,0.97,fill=c(2,4),legend=c("6-MP","control"),bty='n')
```

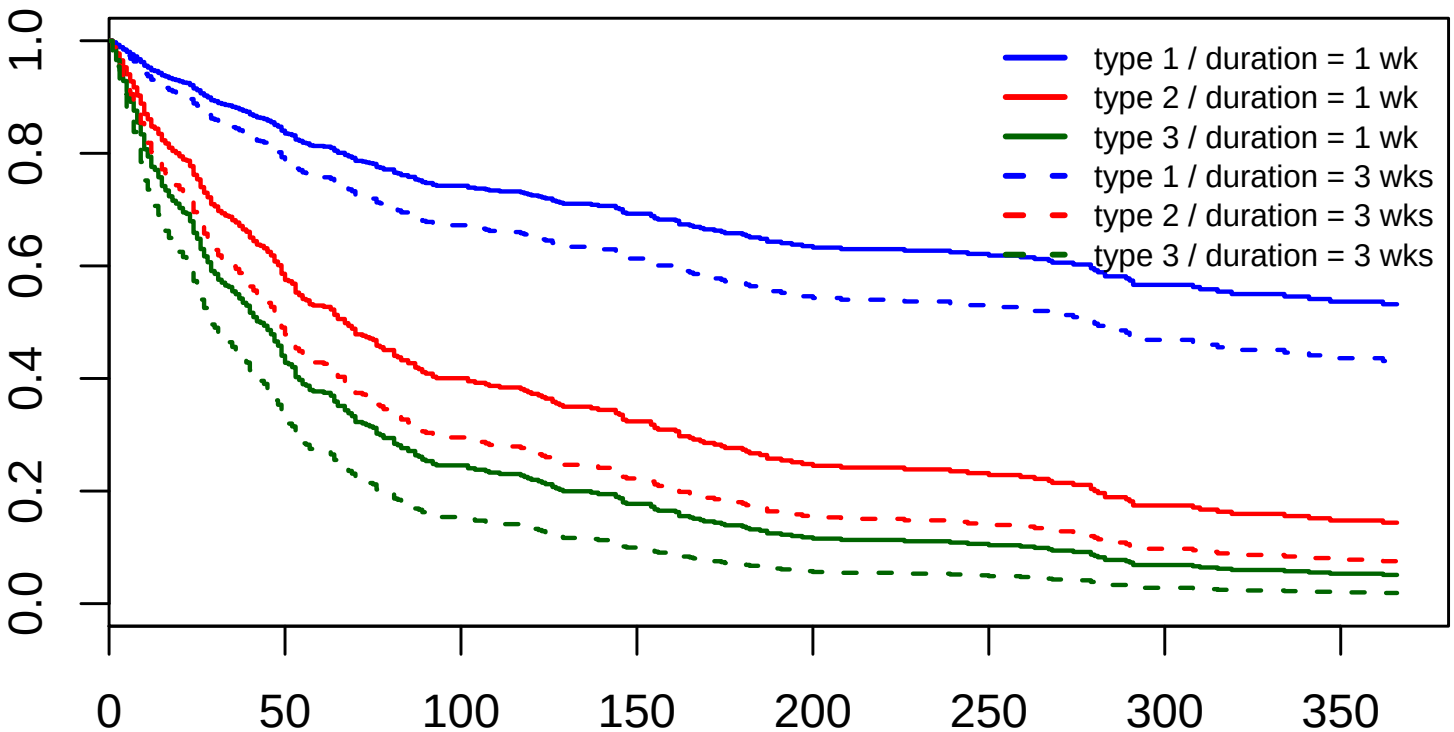


Semiparametric proportional hazards model

In the herpes example, suppose we fit a proportional hazards model with type as a dummy variable and duration as a linear term. We can then look at subgroup-specific estimates of the survival function.

```
herpes$type2 = as.numeric(herpes$type==2); herpes$type3 = as.numeric(herpes$type==3)
coxfit = coxph(s.herpes~type2+type3+duration,data=herpes)

profile1 = data.frame(type2=0,type3=0,duration=7)
profile2 = data.frame(type2=1,type3=0,duration=7)
profile3 = data.frame(type2=0,type3=1,duration=7)
profile4 = data.frame(type2=0,type3=0,duration=21)
profile5 = data.frame(type2=1,type3=0,duration=21)
profile6 = data.frame(type2=0,type3=1,duration=21)

plot(survfit(coxfit,newdata=profile1,conf.type="none"),col=4,lwd=1.5)
lines(survfit(coxfit,newdata=profile2,conf.type="none"),col=2,lwd=1.5)
lines(survfit(coxfit,newdata=profile3,conf.type="none"),col="darkgreen",lwd=1.5)
lines(survfit(coxfit,newdata=profile4,conf.type="none"),col=4,lwd=1.5,lty=2)
lines(survfit(coxfit,newdata=profile5,conf.type="none"),col=2,lwd=1.5,lty=2)
lines(survfit(coxfit,newdata=profile6,conf.type="none"),col="darkgreen",lwd=1.5,lty=2)

legend("topright",
       legend=c("type 1 / duration = 1 wk","type 2 / duration = 1 wk",
                "type 3 / duration = 1 wk","type 1 / duration = 3 wks",
                "type 2 / duration = 3 wks","type 3 / duration = 3 wks"),
       col=c(4,2,"darkgreen",4,2,"darkgreen"),lwd=rep(2,6),lty=c(rep(1, 3),rep(2,3)),cex=.7,bty="n")
```

Semiparametric proportional hazards model

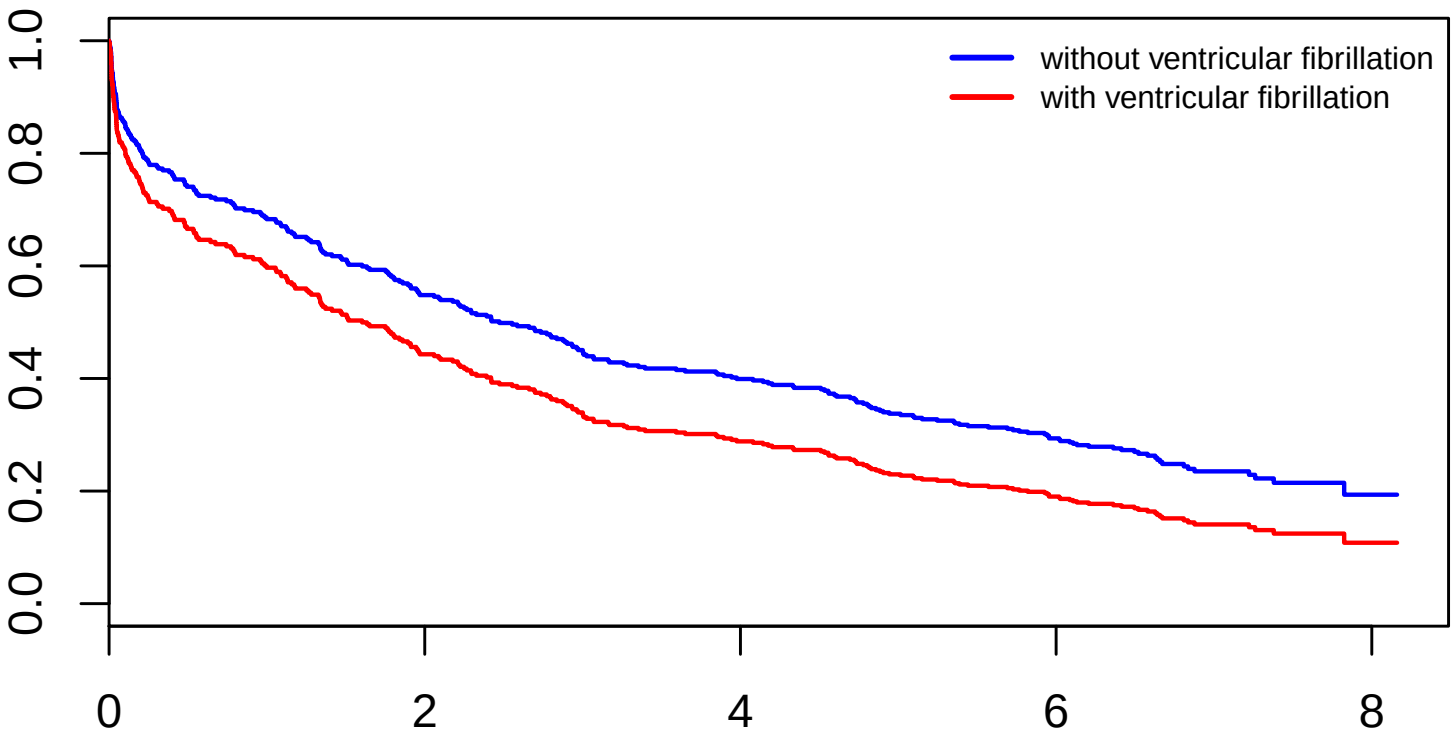


Semiparametric proportional hazards model

Based on a Cox model, what is the estimated survival function of a female patient in ventricular fibrillation with a WMI of 1, congestive heart failure and no diabetes? What about a similar patient not in ventricular fibrillation?

```
coxfit = coxph(s.trace~wmi+chf+sex+diabetes+vf,data=trace)

profile1 = data.frame(wmi=1,chf=1,sex=1,diabetes=0,vf=0)
profile2 = data.frame(wmi=1,chf=1,sex=1,diabetes=0,vf=1)

plot(survfit(coxfit,newdata=profile1,conf.type="none"),col=4,lwd=1.5)
lines(survfit(coxfit,newdata=profile2,conf.type="none"),col=2,lwd=1.5)

legend("topright",
       legend=c("without ventricular fibrillation",
                "with ventricular fibrillation"),
       col=c(4,2),lty=rep(1,2),lwd=rep(2,2),cex=.7,bty="n")
```

Semiparametric proportional hazards model



Stratified proportional hazards model

Suppose that Z_2 is a confounding factor, say with three levels (e.g., 0, 1 and 2).

We have seen that we can flexibly adjust for Z_2 by incorporating the dummy variables

$$Z_{2,1} := \begin{cases} 1 & : \ Z_2 = 1 \\ 0 & : \ Z_2 \neq 1 \end{cases} \quad \text{and} \quad Z_{2,2} := \begin{cases} 1 & : \ Z_2 = 2 \\ 0 & : \ Z_2 \neq 2 \end{cases}$$

as main terms, as in the proportional hazards model

$$h(t \mid z_1, z_2) = h_0(t) \exp(\beta_1 z_1 + \beta_{2,1} z_{2,1} + \beta_{2,2} z_{2,2}) \ .$$

But is this flexible enough?

- this model requires proportional hazards between levels of Z_2;
- if this assumption fails, there may be residual confounding from Z_2 since its full confounding effects may not be properly accounting for;
- if we are focused on adjusting for Z_2, do we really need proportionality of hazards?

Stratified proportional hazards model

In this simple approach, all subgroup-specific hazard functions must be proportional to one another!

To relax this constraint, we may instead use the **stratified proportional hazards model**

$$h(t \mid z_1, z_2) = h_0(t \mid z_2) e^{\beta_1 z_1} \ ,$$

where $h_0(t \mid z_2)$ is an unspecified baseline hazard function (at time t) corresponding to the subgroup $Z_2 = z_2$.

In contrast to the proportional hazards model, each subgroup defined by levels of a categorical variable Z_2 has its own baseline hazard function – there is **no requirement of proportionality across levels of the stratification variable** Z_2.

Stratified proportional hazards model

How do we interpret parameters in this stratified model?

$$
\begin{aligned}
e^{\beta_1} &= \frac{h_0(t \mid z_2)e^{\beta_1}}{h_0(t \mid z_2)} = \frac{h(t \mid 1, z_2)}{h(t \mid 0, z_2)} \\
&= \text{hazard ratio comparing individuals with } Z_1 = 1 \text{ to other individuals with } Z_1 = 0 \text{ but the same level of } Z_2 \\
h_0(t \mid z_2) &= h(t \mid 0, z_2) \\
&= \text{hazard function for individuals with } Z_1 = 0 \text{ and } Z_2 = z_2
\end{aligned}
$$

A few important remarks:

- this model provides no concise measure of association between hazard and Z_2;
- for this reason, stratification should only be performed with respect to variables that we need to adjust for but are otherwise not of interest;
- here, the association between hazard and Z_1 is assumed fixed across strata of Z_2;
- this model is much more flexible but can lead to less precise estimates, particularly for the subgroup-specific survival functions.

Stratified proportional hazards model

In a stratification model, can we include a main term for the stratification variable?

.....

$$\begin{aligned} h(t \mid z_1, z_2) &= h_0(t \mid z_2) e^{\beta_1 z_1 + \beta_2 z_2} \\ &= \underbrace{h_0(t \mid z_2) e^{\beta_2 z_2}}_{h_0^*(t|z_2)} e^{\beta_1 z_1} = h_0^*(t \mid z_2) e^{\beta_1 z_1} \end{aligned}$$

An interaction term involving the stratification variable Z_2 allows the association between hazard and Z_1 to vary in different strata of Z_2, as in the model

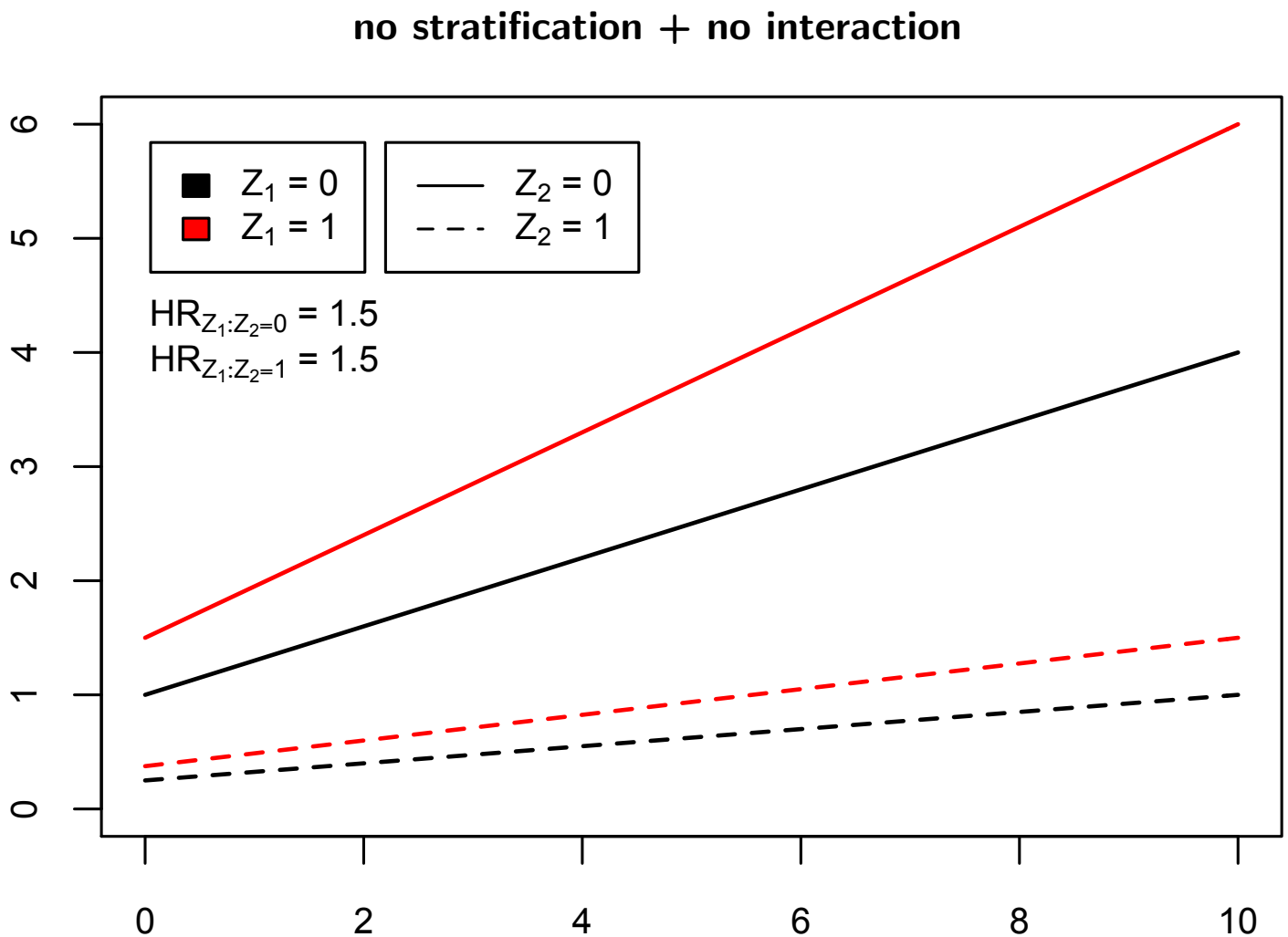
$$h(t \mid z_1, z_2) = h_0(t \mid z_2) e^{\beta_1 z_1 + \gamma z_1 z_2}$$

$HR_{Z_1:Z_2=0}$ = ...

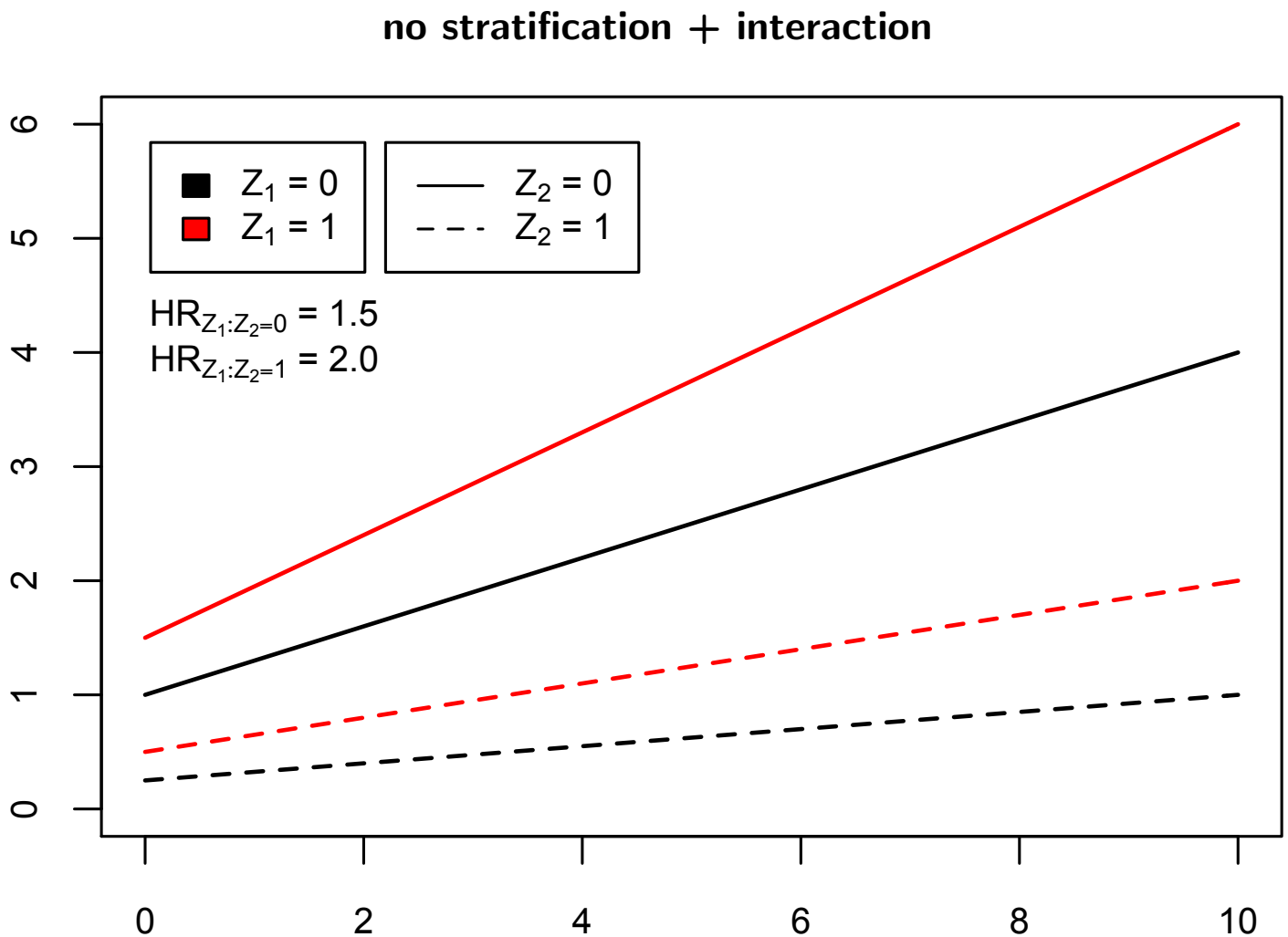
$HR_{Z_1:Z_2=1}$ = ...

$\frac{HR_{Z_1:Z_2=1}}{HR_{Z_1:Z_2=0}}$ = ...

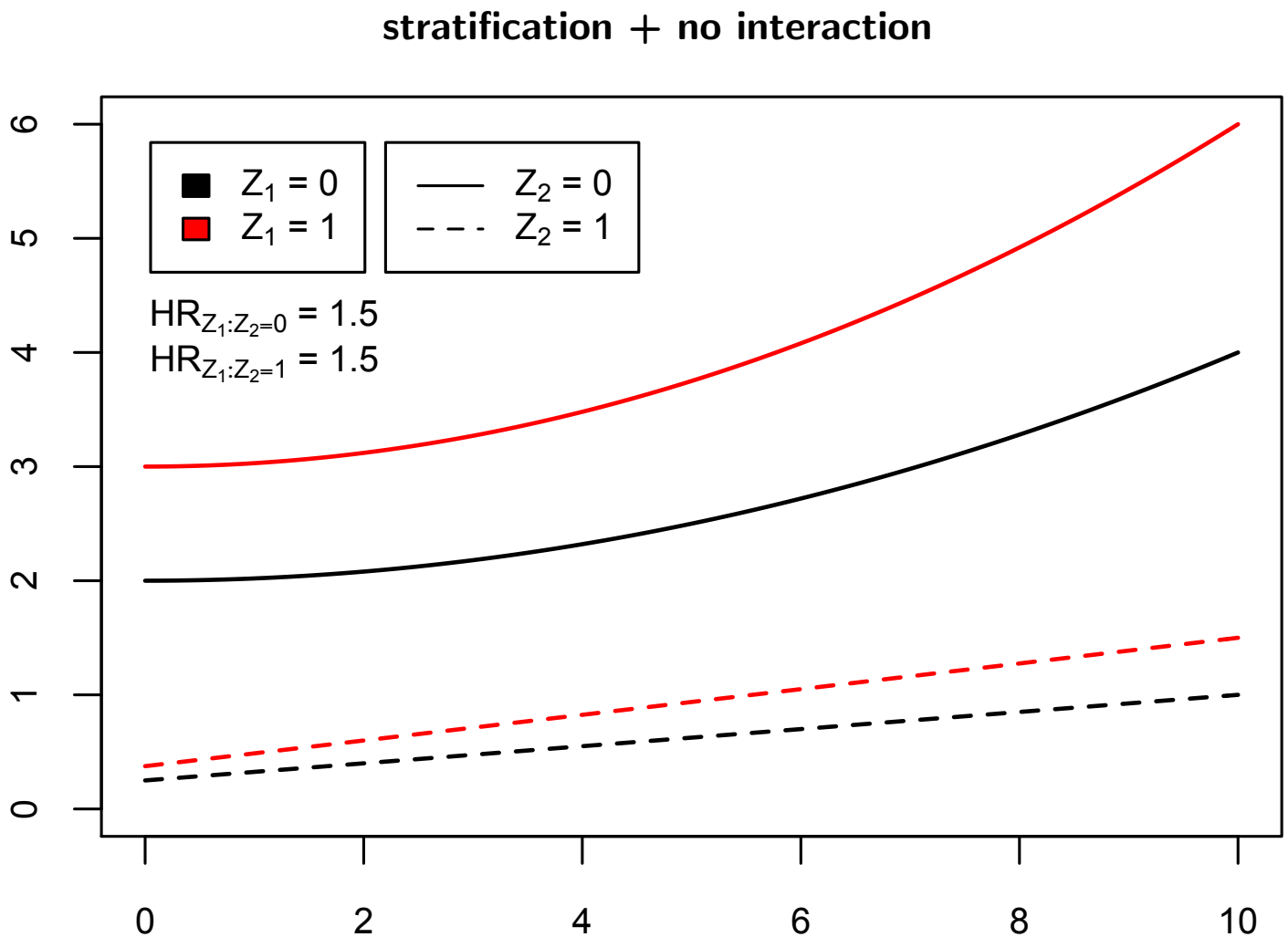
Stratified proportional hazards model



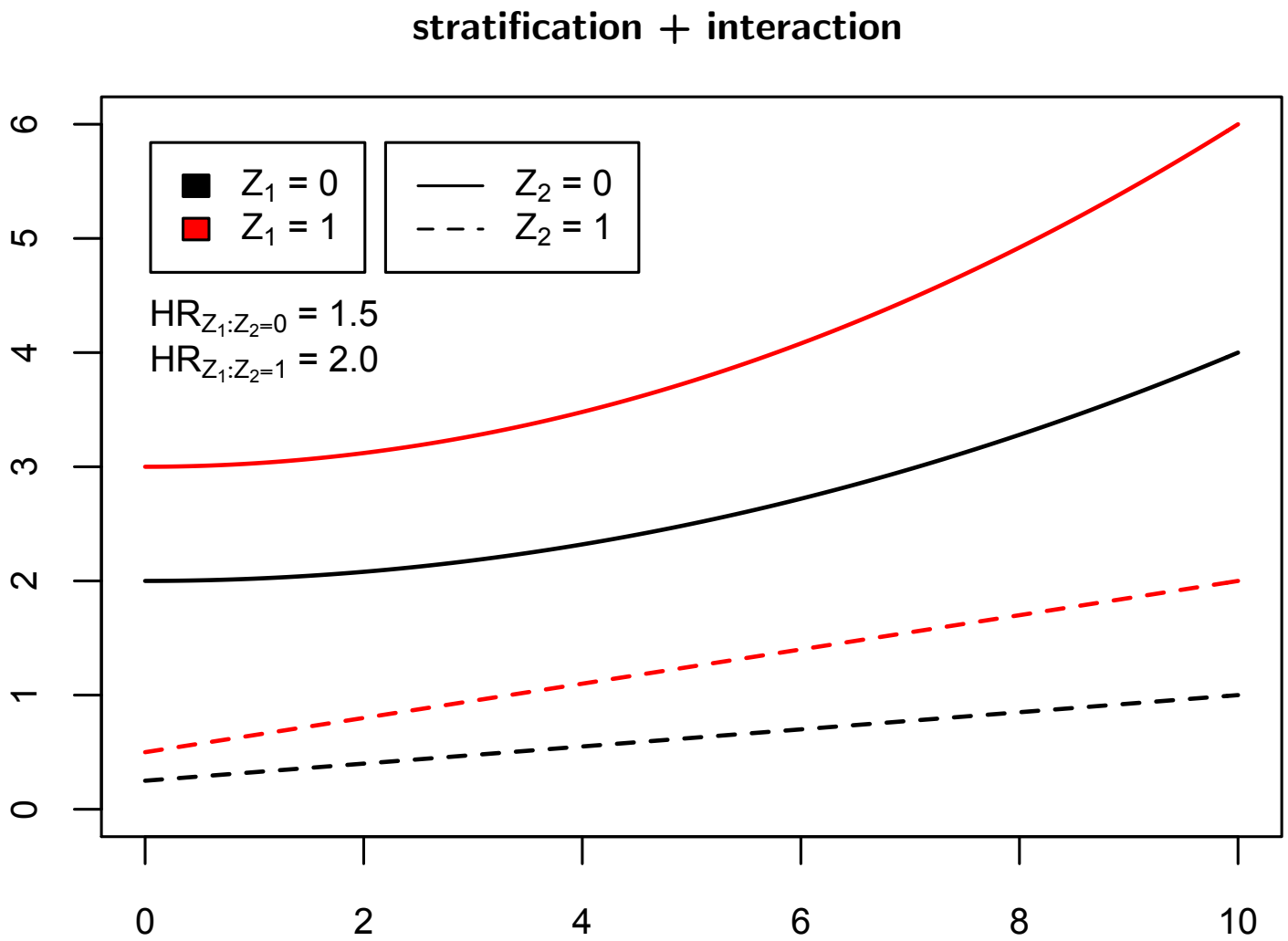
Stratified proportional hazards model



Stratified proportional hazards model



Stratified proportional hazards model



Stratified proportional hazards model

We wish to determine the association between treatment with acyclovir and time until recurrence of genital herpes infection, adjusting for HSV type, sex and duration of primary lesion.

```
quantile(herpes$duration,seq(0.2,0.8,by=0.2))

## 20% 40% 60% 80%
##  13  17  20  25

herpes$durgrp = 0
herpes$durgrp[herpes$duration>13 & herpes$duration<=17] = 1
herpes$durgrp[herpes$duration>17 & herpes$duration<=20] = 2
herpes$durgrp[herpes$duration>20 & herpes$duration<=25] = 3
herpes$durgrp[herpes$duration>25] = 4
herpes$durgrp.=as.factor(herpes$durgrp)
```

Stratified proportional hazards model

```
summary(coxph(s.herpes~treat.+type.+durgrp.+male,data=herpes))$coef
```

```
##                 coef exp(coef)  se(coef)          z     Pr(>|z|)
## treat.1  -0.05671657 0.9448618 0.1965206 -0.2886037 7.728846e-01
## treat.2   0.39255783 1.4807635 0.1253668  3.1312744 1.740495e-03
## treat.3  -0.24459529 0.7830214 0.3272378 -0.7474543 4.547894e-01
## type.2    1.24373084 3.4685299 0.1822810  6.8231517 8.906431e-12
## type.3    1.61809053 5.0434508 0.2191625  7.3830616 1.546541e-13
## durgrp.1  0.34268283 1.4087219 0.1571789  2.1802092 2.924196e-02
## durgrp.2  0.37164403 1.4501167 0.1703293  2.1819143 2.911586e-02
## durgrp.3  0.67693059 1.9678284 0.1797636  3.7656709 1.661025e-04
## durgrp.4  0.46202118 1.5872789 0.1781871  2.5928986 9.517084e-03
## male      0.18662789 1.2051787 0.1122327  1.6628659 9.633929e-02
```

```
summary(coxph(s.herpes~treat.+strata(type.)+strata(durgrp.)+strata(male),data=herpes))$coef
```

```
##               coef exp(coef)  se(coef)          z    Pr(>|z|)
## treat.1 -0.1509985 0.8598490 0.2174795 -0.6943114 0.487486965
## treat.2  0.3620175 1.4362241 0.1341846  2.6979065 0.006977704
## treat.3 -0.3384116 0.7129018 0.3315843 -1.0205898 0.307448827
```

Diagnostic tools

The regression models we use are a simplification of reality – we do not expect them to precisely mimic the relationships of interest. We do hope they can be useful.

It may be useful to determine how faithfully the models used represent reality.

Several **model fit diagnostic tools** are used for this purpose, including:

1. residual scatterplots and other visualization techniques;
2. extended models (e.g., stratification, time-varying coefficients) and formal hypothesis testing.

In practice, it may also be useful to determine if certain observations inordinately affect parameter estimates – **influence diagnostic tools** may be used for this.

Diagnostic tools

Residuals play a critical role in the evaluation of fit for linear mean models of the type

$$E(Y \mid Z = z) = \beta z$$

since then $R := Y - \beta Z$ is known to have mean zero if the model is correct.

In survival analysis, the outcome data are generally censored (and possibly truncated). Moreover, in the proportional hazards model, we do not model the mean outcome (or any transformation of the outcome). What then can be used as a residual?

Here, we will briefly discuss how two different residual proposals can be useful:

1. the martingale residuals (and related deviance residuals);
2. the Schoenfeld residuals.

Diagnostic tools

One of the objects most reminiscent of a residual from linear models is the so-called **martingale residual**, defined as

$$M_i := \Delta_i - \widehat{H}(Y_i \mid Z_i) \ ,$$

where $\widehat{H}(t \mid z)$ is an estimator of $H(t \mid z)$.

These residuals are essentially of the form "observed count – expected count".

If a correct model is used, these residuals are uncorrelated and have mean nearly zero. Additionally, in any given sample, these residuals have sample mean zero.

These residuals provide an assessment of overall model fit and are useful to determine the functional form through which covariates should be included in the model.

Diagnostic tools

Suppose that the covariate vector $Z = (Z_1, Z_2)$, where Z_2 is scalar but Z_1 may be a vector, and that the hazard rate of T given Z is given by

$$h(t \mid z_1, z_2) = h_0(t) \exp\{\beta z_1 + f_0(z_2)\} \ .$$

To determine the form of f_0, the following approach can be used:

1. fit the model $h(t \mid z_1) = h_{0,*}(t) \exp(\beta_* z_1)$ omitting covariate Z_2;
2. compute the martingale residuals M_i, $i = 1, 2, \ldots, n$, based on this model;
3. use a smoother (e.g., lowess) to estimate $E(M \mid Z_2 = z_2)$ as a function of z_2.

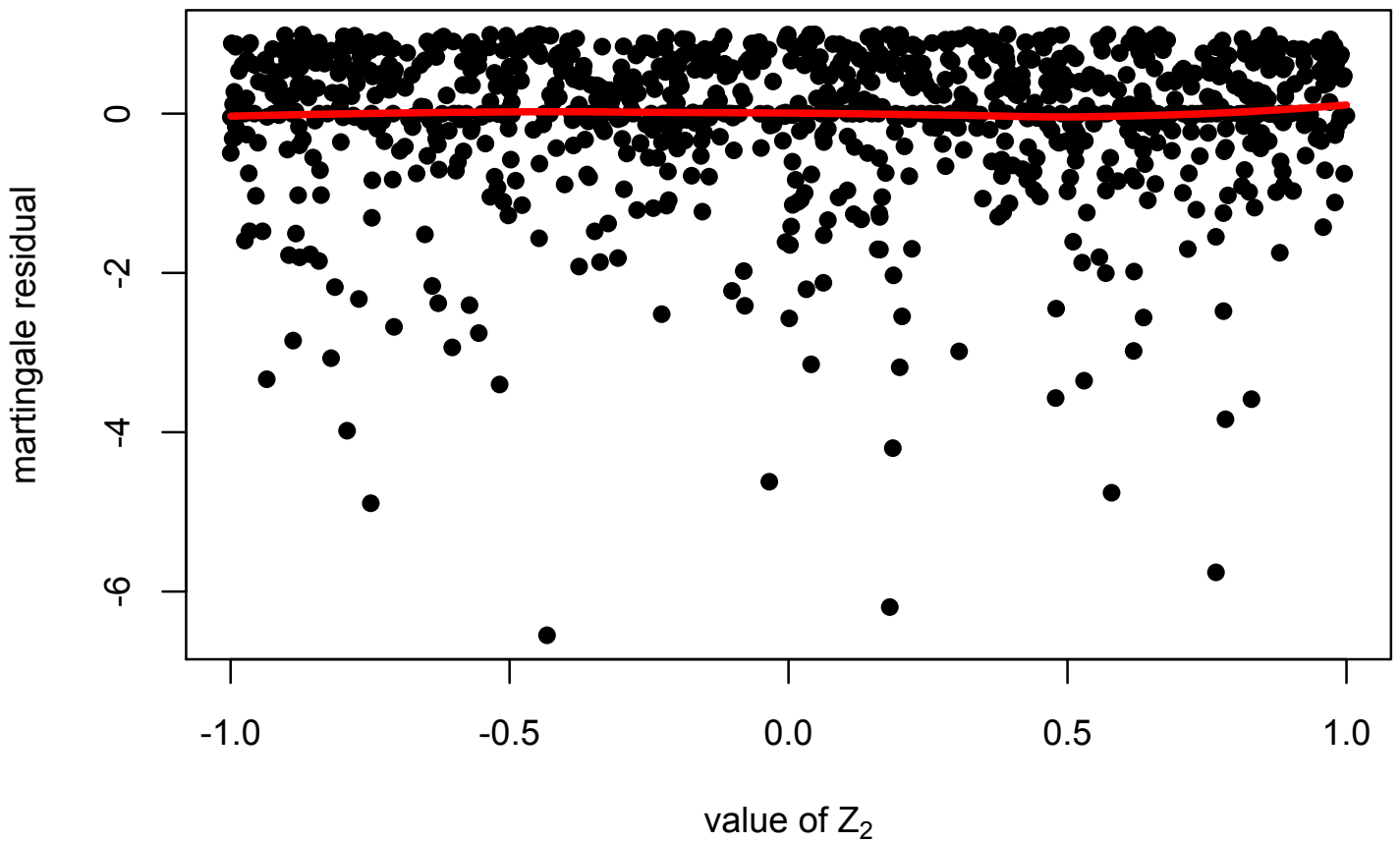
$E(M \mid Z_2 = z_2)$ is approximately proportional to $f_0(z_2)$ as a function of z_2.

This can be useful when deciding whether Z_2 should be included in the model, and if so, what functional form it should take (e.g., threshold indicator, linear, quadratic).

Diagnostic tools

model fitted: $h(t \mid z_1, z_2) = h_0(t) \exp(\beta_1 z_1 + \beta_2 z_2)$

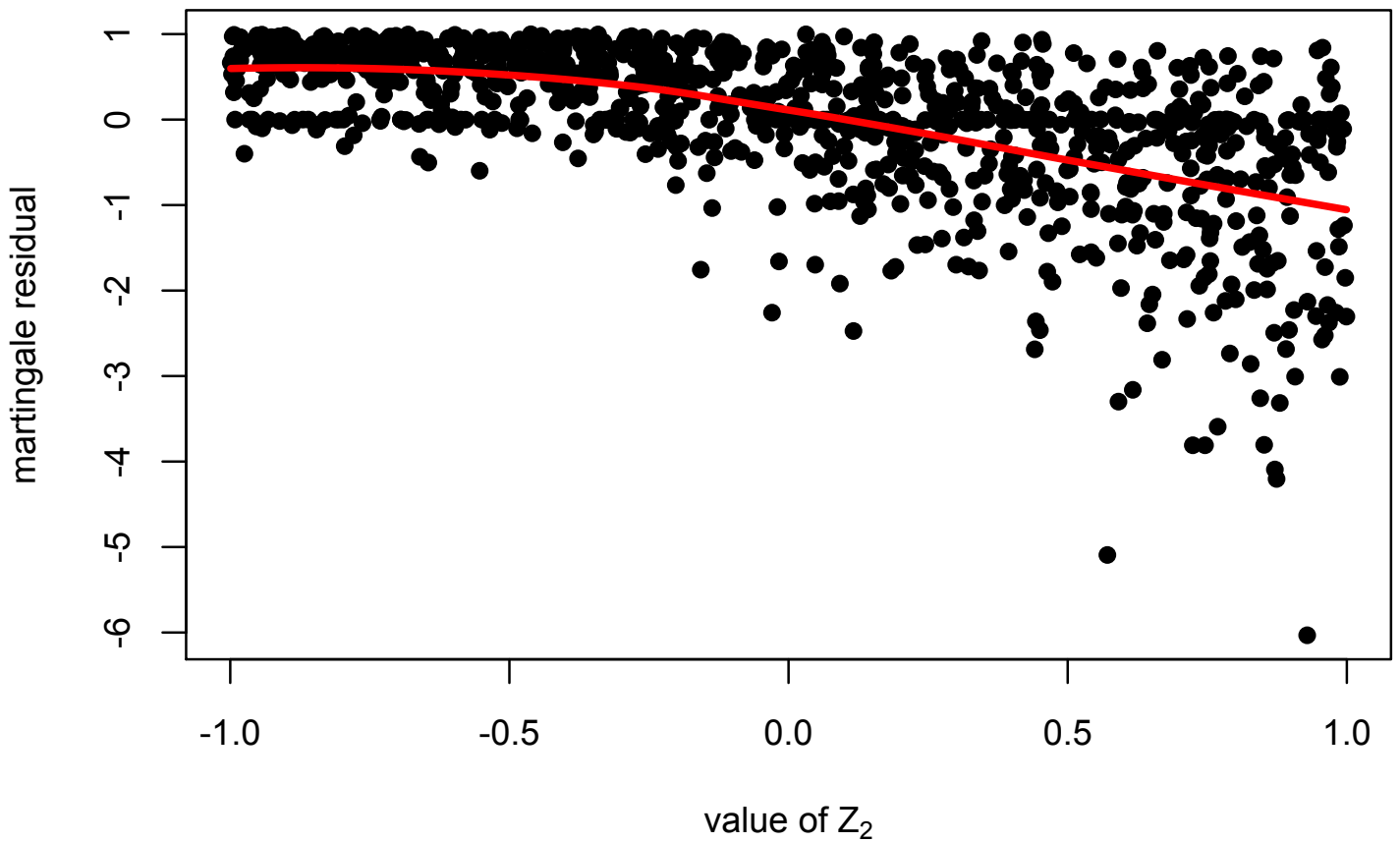
actual model: $h(t \mid z_1, z_2) = h_0(t) \exp(2z_1 - 2z_2)$



Diagnostic tools

model fitted: $h(t \mid z_1, z_2) = h_0(t) \exp(\beta_1 z_1)$

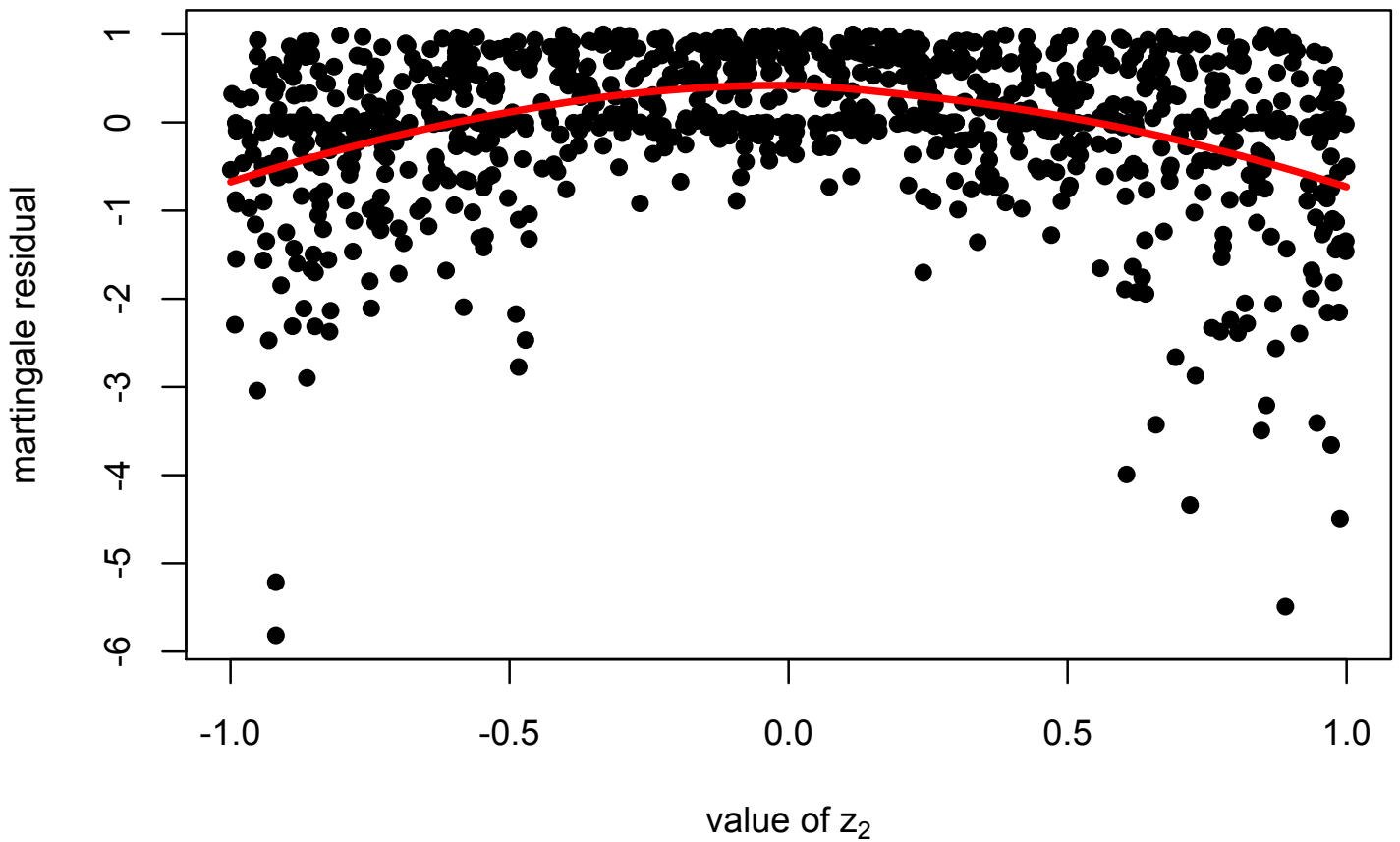
actual model: $h(t \mid z_1, z_2) = h_0(t) \exp(2z_1 - 2z_2)$



Diagnostic tools

model fitted: $h(t \mid z_1, z_2) = h_0(t) \exp(\beta_1 z_1)$

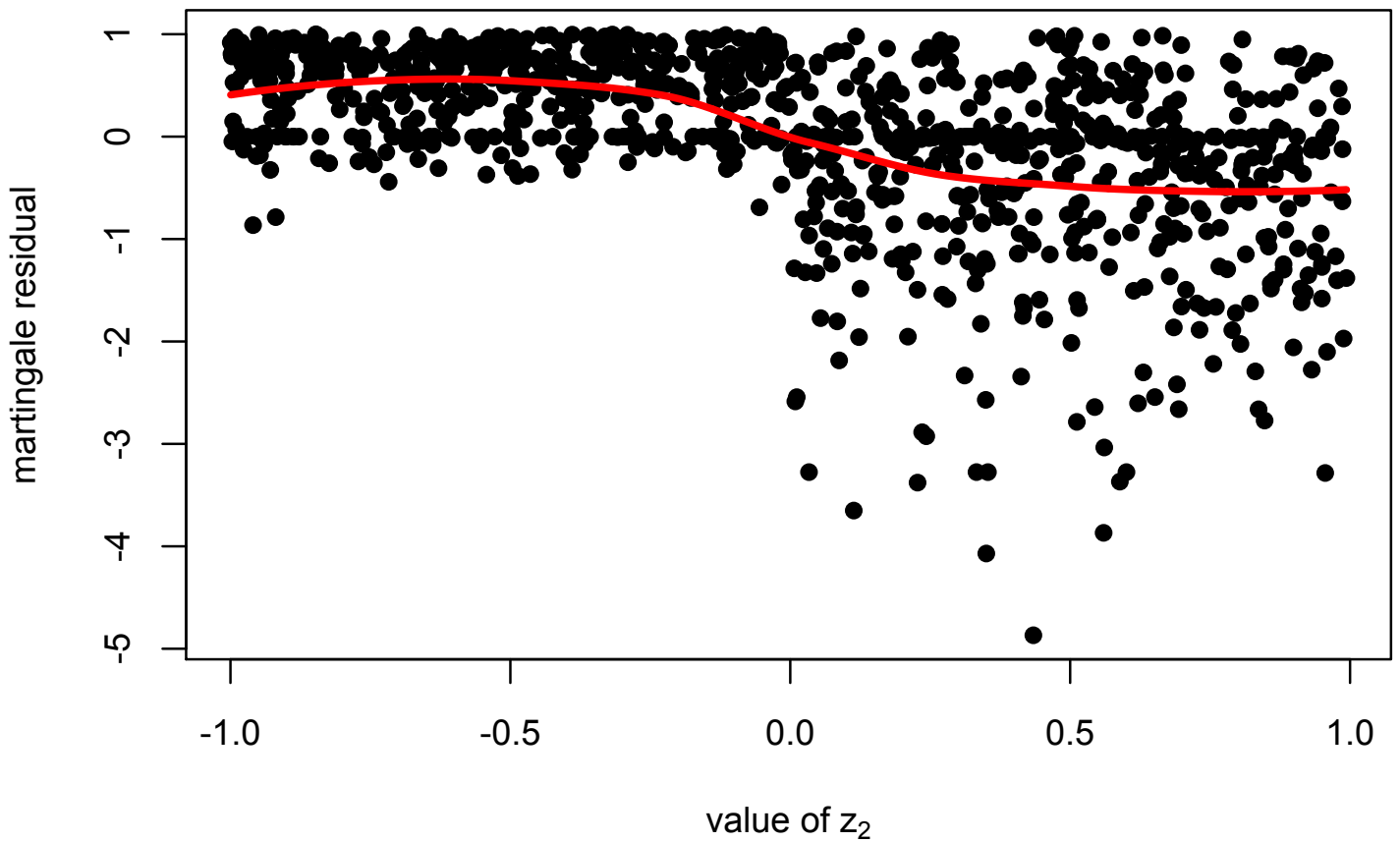
actual model: $h(t \mid z_1, z_2) = h_0(t) \exp(2z_1 - 2|z_2|)$



Diagnostic tools

model fitted: $h(t \mid z_1, z_2) = h_0(t) \exp(\beta_1 z_1)$

actual model: $h(t \mid z_1, z_2) = h_0(t) \exp(2z_1 - 2I(z_2 > 0))$



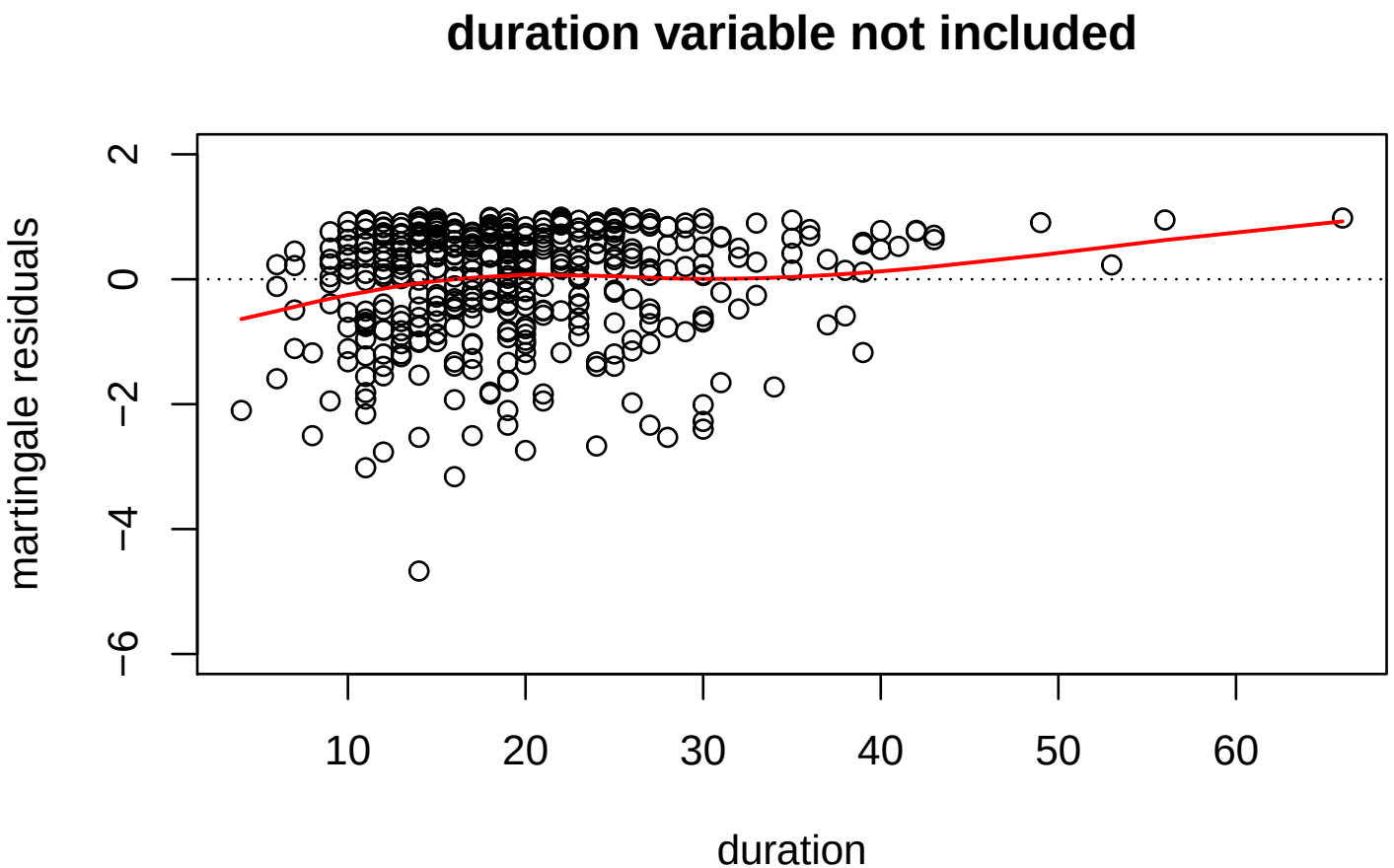
Diagnostic tools

We can investigate the adequacy of the functional forms employed by fitting a proportional hazards model including HSV type, treatment, sex, age and duration of primary lesion and scrutinizing martingale residuals.

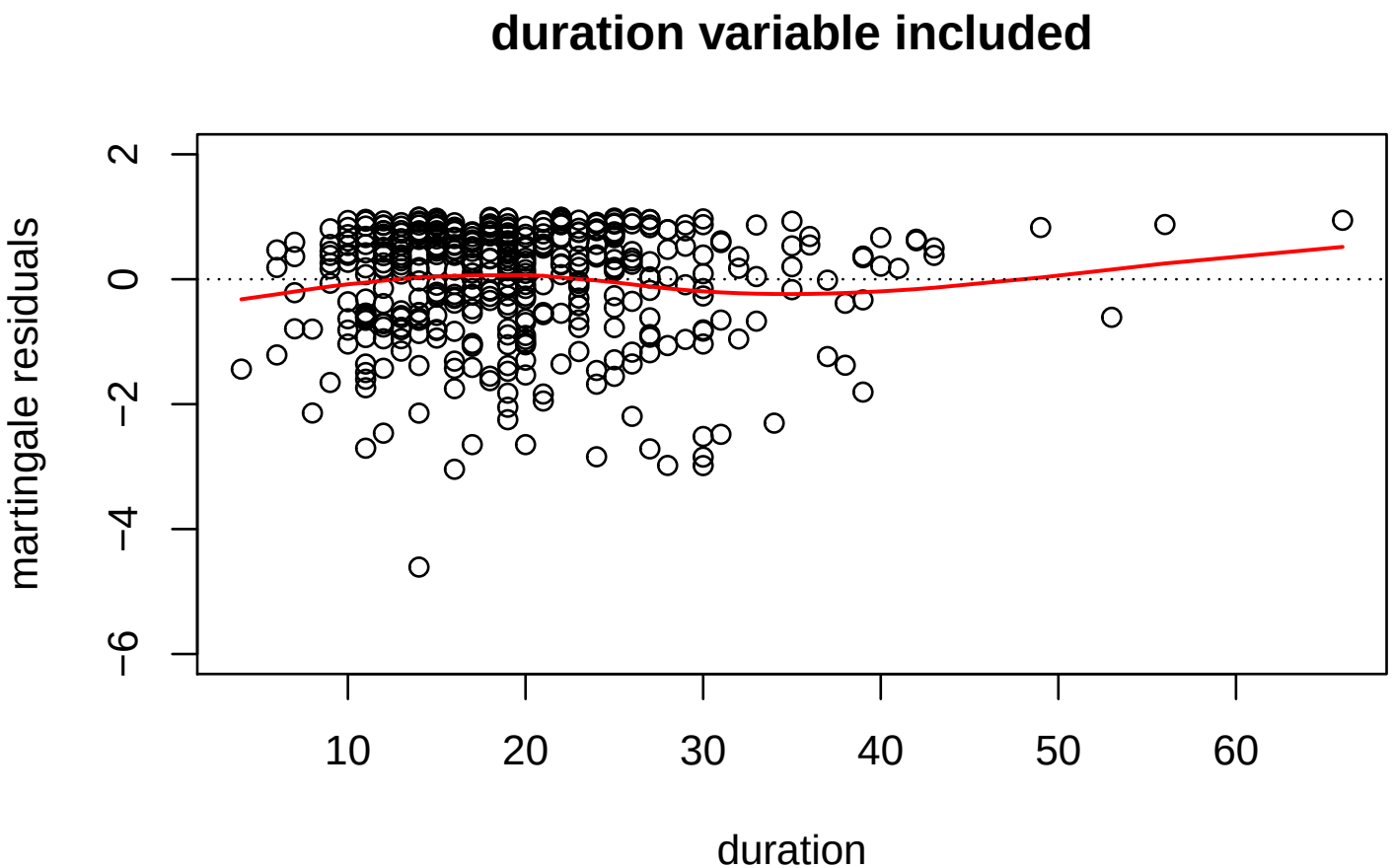
First, we may look at the duration variable.

```
ind = complete.cases(herpes)
coxfit1 = coxph(s.herpes~treat.+type.+male+age,subset=ind,data=herpes)
mgresid1 = residuals(coxfit1,type="martingale")
coxfit2 = coxph(s.herpes~treat.+type.+male+age+duration,subset=ind,data=herpes)
mgresid2 = residuals(coxfit2,type="martingale")
durationvals = herpes$duration[ind]

plot(durationvals,mgresid1,xlab="duration",ylab="martingale residuals",ylim=c(-6,2),
      main="duration variable not included")
mgresid1.loess = loess(mgresid1~durationvals,degree=1)
lines(sort(durationvals),predict(mgresid1.loess,sort(durationvals)),col=2,lwd=1.5)
abline(h=0,lty=3)

plot(durationvals,mgresid2,xlab="duration",ylab="martingale residuals",ylim=c(-6,2),
      main="duration variable included")
mgresid2.loess = loess(mgresid2~durationvals,degree=1)
lines(sort(durationvals),predict(mgresid2.loess,sort(durationvals)),col=2,lwd=1.5)
abline(h=0,lty=3)
```

Diagnostic tools



Diagnostic tools

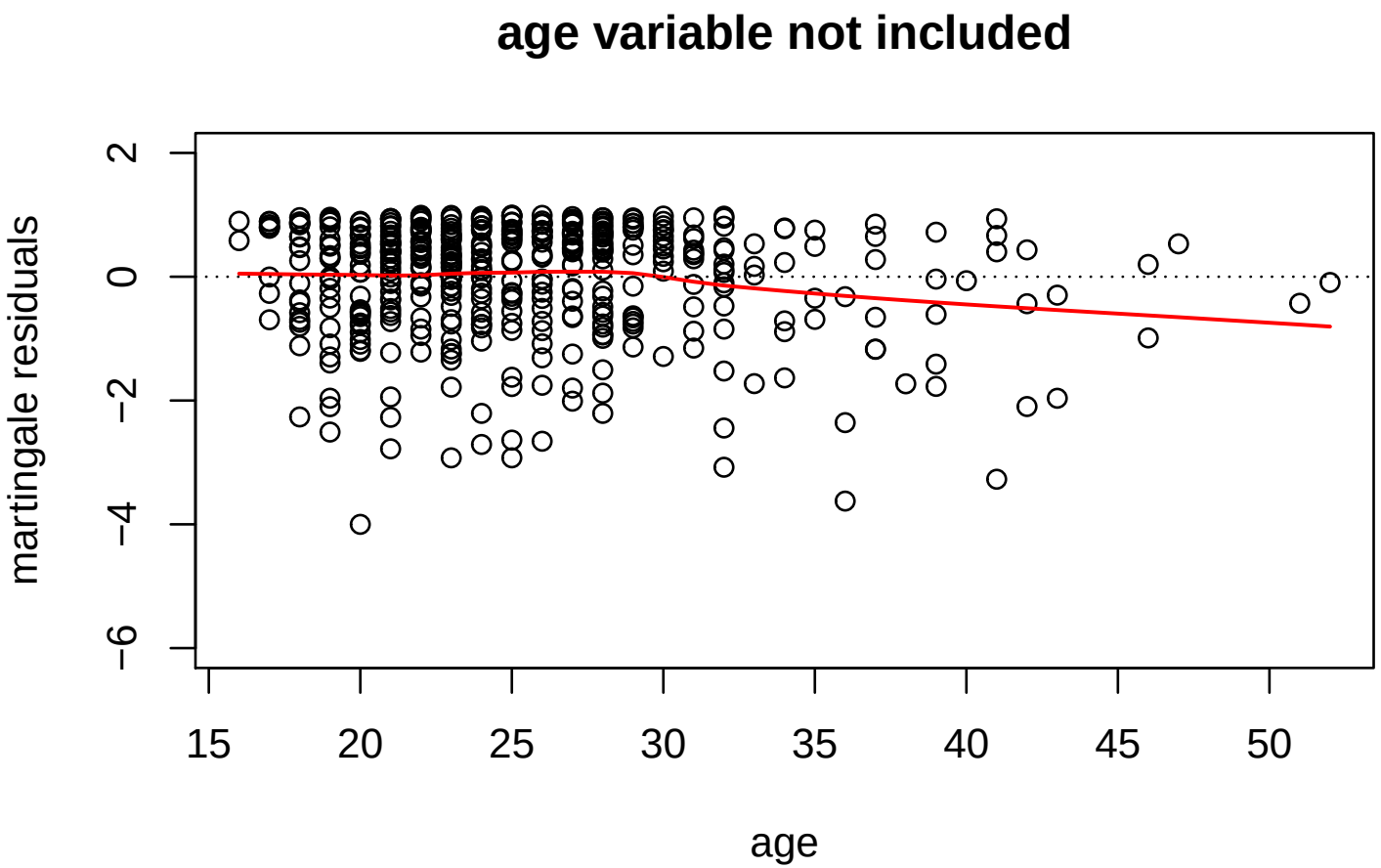


Diagnostic tools

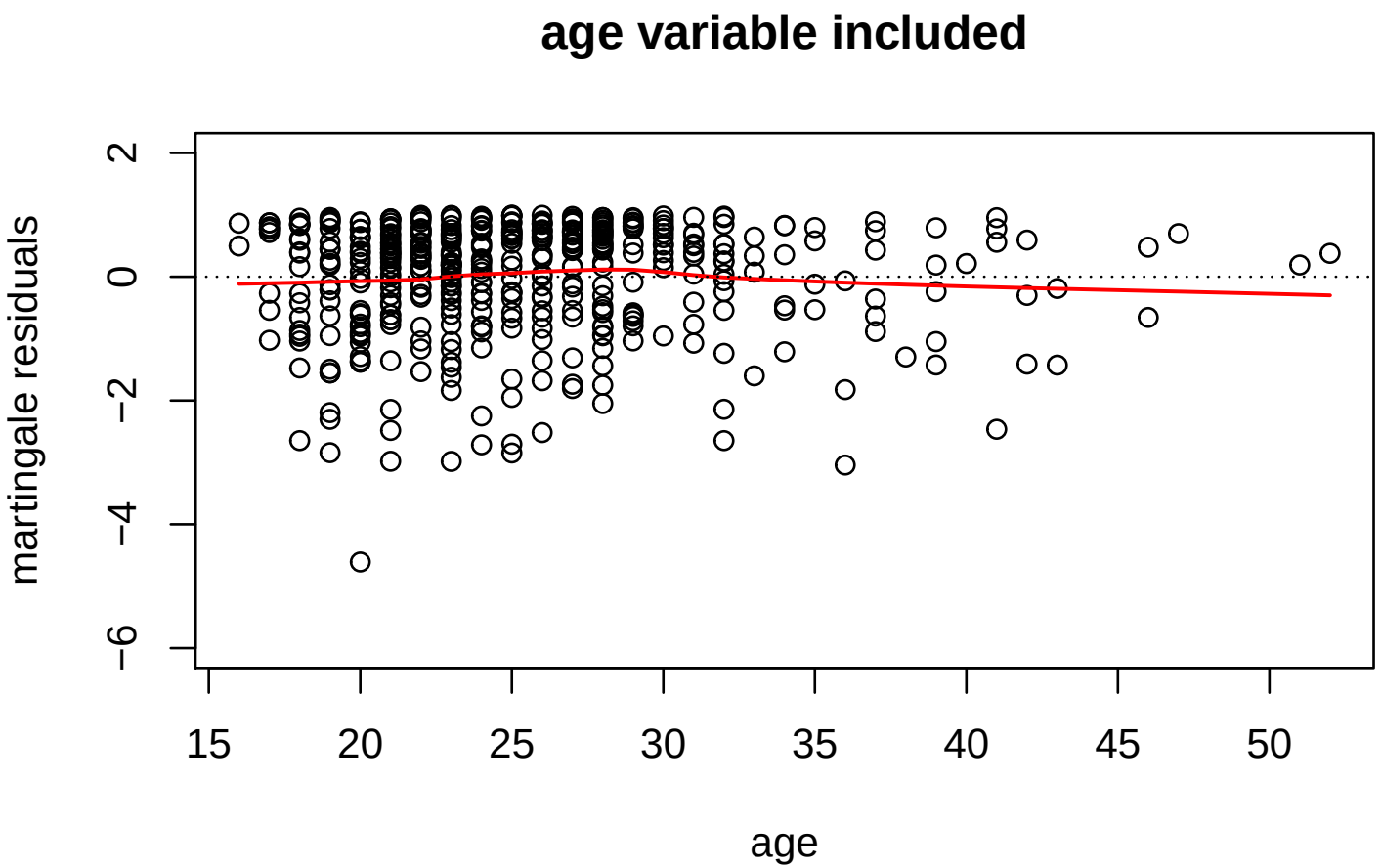
We may do the same for the age variable.

```
ind = complete.cases(herpes)
coxfit1 = coxph(s.herpes~treat.+type.+male+duration,subset=ind,data=herpes)
mgresid1 = residuals(coxfit1,type="martingale")
coxfit2 = coxph(s.herpes~treat.+type.+male+duration+age,subset=ind,data=herpes)
mgresid2 = residuals(coxfit2,type="martingale")
agevals = herpes$age[ind]

plot(agevals,mgresid1,xlab="age",ylab="martingale residuals",ylim=c(-6,2),
      main="age variable not included")
mgresid1.loess = loess(mgresid1~agevals,degree=1)
lines(sort(agevals),predict(mgresid1.loess,sort(agevals)),col=2,lwd=1.5)
abline(h=0,lty=3)

plot(agevals,mgresid2,xlab="age",ylab="martingale residuals",ylim=c(-6,2),
      main="age variable included")
mgresid2.loess = loess(mgresid2~agevals,degree=1)
lines(sort(agevals),predict(mgresid2.loess,sort(agevals)),col=2,lwd=1.5)
abline(h=0,lty=3)
```

Diagnostic tools



Diagnostic tools



Diagnostic tools

Martingale residuals tend to be highly skewed on the interval $(-\infty, 1]$ – this can make it difficult to understand whether a patient is an outlier.

The **deviance residual** for the i^{th} observation is defined as

$$\begin{aligned} D_i &:= \text{sign}(M_i)\sqrt{-2\{M_i + \Delta_i \log(\Delta_i - M_i)\}} \\ &= \text{sign}(M_i)\sqrt{-2\{\Delta_i - \widehat{H}(Y_i \mid Z_i) + \Delta_i \log \widehat{H}(Y_i \mid Z_i)\}} \ . \end{aligned}$$

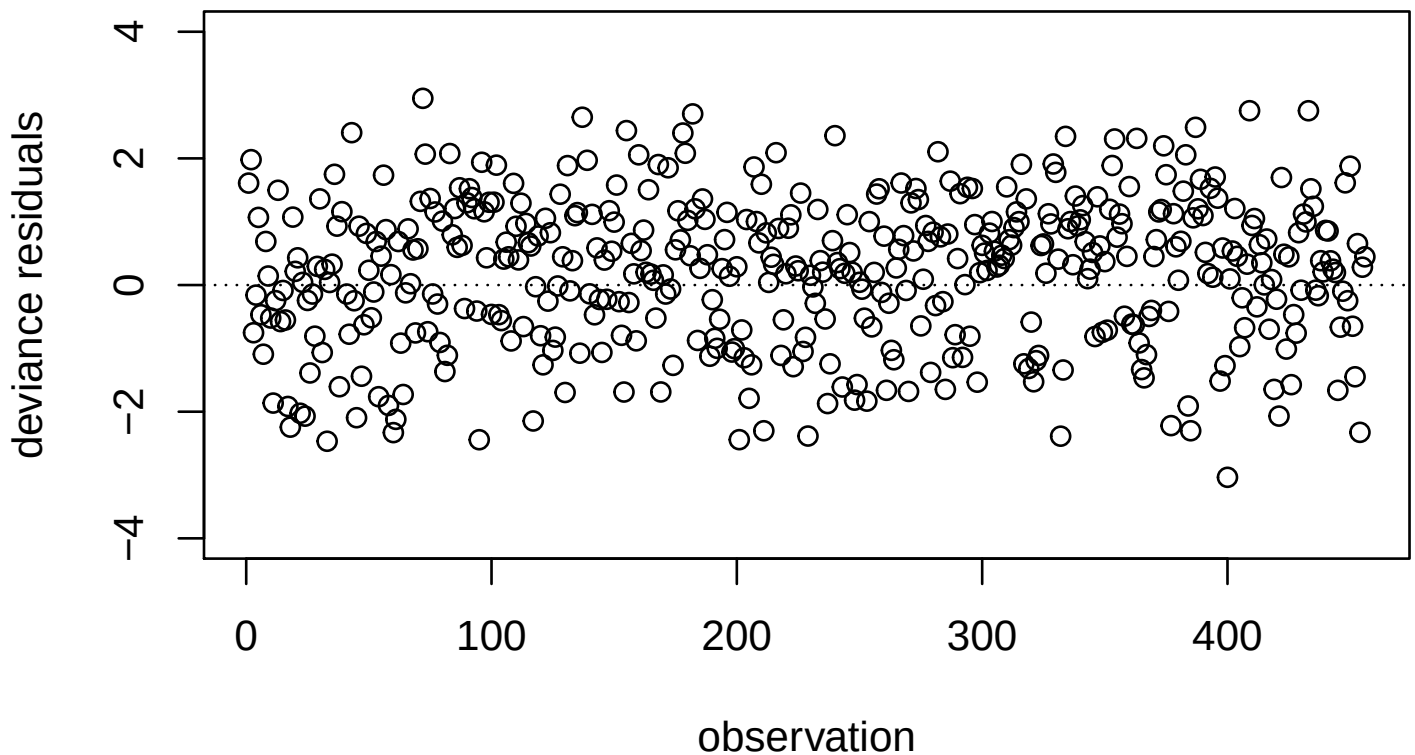
A few observations about deviance residuals:

- the square root shrinks large negative values and the logarithm blows up residuals close to one;
- these residuals are close to symmetric around zero with unit standard deviation;
- positive/negative value = patient died too soon/late;
- these residuals are particularly useful for identifying outliers.

Diagnostic tools

We can look at whether there is any outlier based on the joint model used so far.

```
ind = which(complete.cases(herpes))
coxfit = coxph(s.herpes~treat.+age+duration+type.+male,data=herpes)
devresid = residuals(coxfit,type="deviance")
plot(ind,devresid,xlab="observation",ylab="deviance residuals",ylim=c(-4,4))
abline(h=0,lty=3)
```



Diagnostic tools

The Schoenfeld residual is another type of residual, particularly useful for assessing proportionality of hazards.

For a given event time t and patient j at risk at time t, the expression

$$\pi_j(t;\beta) := \frac{e^{\beta z_j}}{\sum_{\ell \in \mathcal{R}(t)} e^{\beta z_\ell}}$$

is the probability that patient j was the one to experience an event at event time t.

If patient i experienced an event, the expression

$$\bar{z}_{ik}(\beta) := \sum_{j \in \mathcal{R}(y_i)} \pi_j(y_i;\beta) z_{jk}$$

is the average value of the k^{th} covariate for individuals at risk, weighted according to each individual's probability of experiencing the event under the Cox model.

Diagnostic tools

If the i^{th} patient experienced an event, the **Schoenfeld residual** for this patient and the k^{th} covariate is defined as

$$SR_{i,k} := z_{ik} - \bar{z}_{ik}(\hat{\beta}) \ .$$

Heuristically, this has the form "observed – expected" and is a component of the score, much like residuals in the context of classical linear models.

Denoting by $\mathcal{E}(t)$ the set of patients who experienced an event at time t, the Schoenfeld residual at time t corresponding to the k^{th} covariate is then defined as

$$SR_k(t) := \sum_{i \in \mathcal{E}(t)} SR_{i,k} \ .$$

A few observations about Schoenfeld residuals:

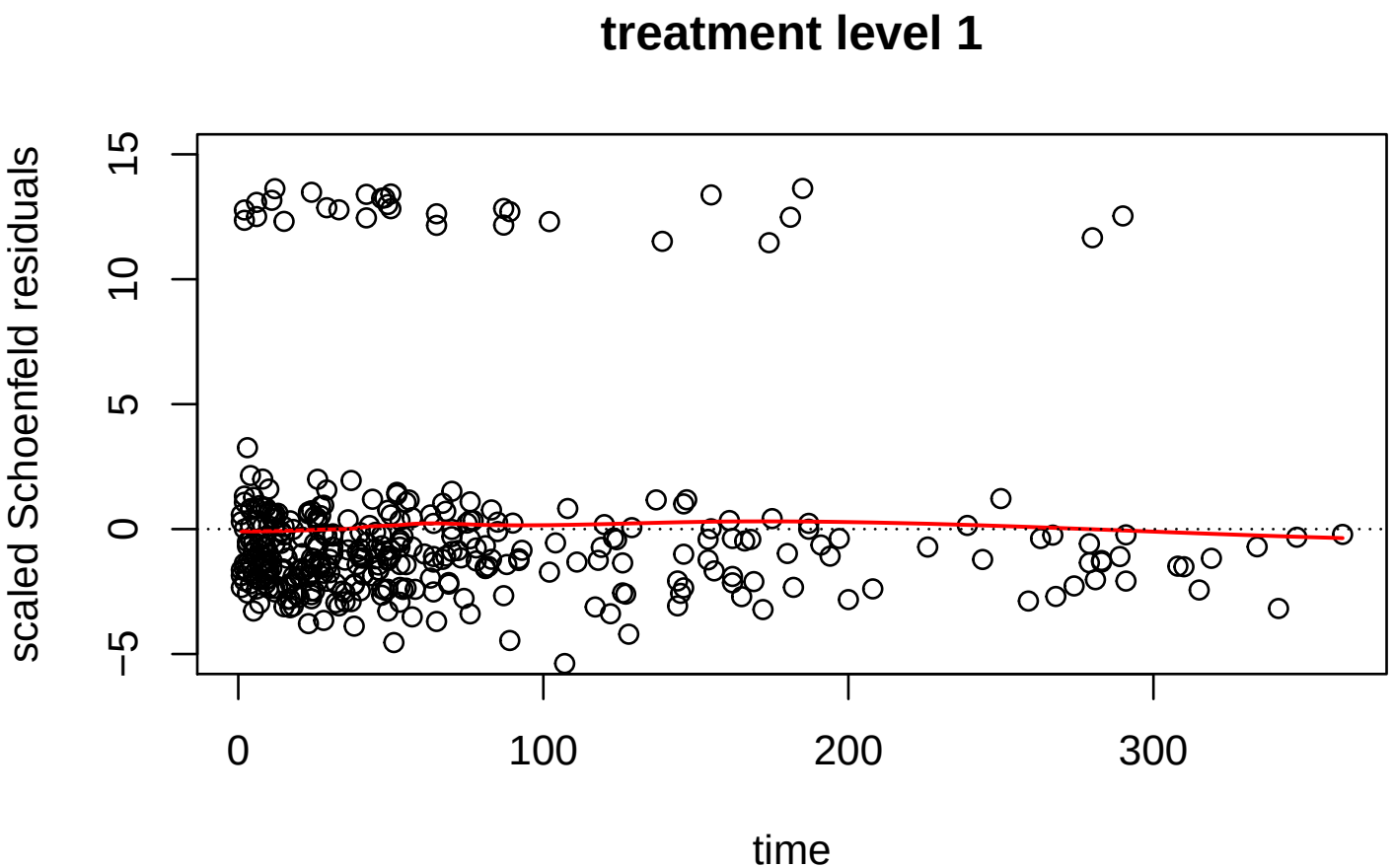
- if the proportional hazards model is correctly specified, across event times, the Schoenfeld residuals are approximately uncorrelated and with mean zero;
- in practice, a variance-scaled version is used to incorporate uncertainty;
- when plotting these residuals against time, trends indicate potential deviation from proportionality of hazards.

Diagnostic tools

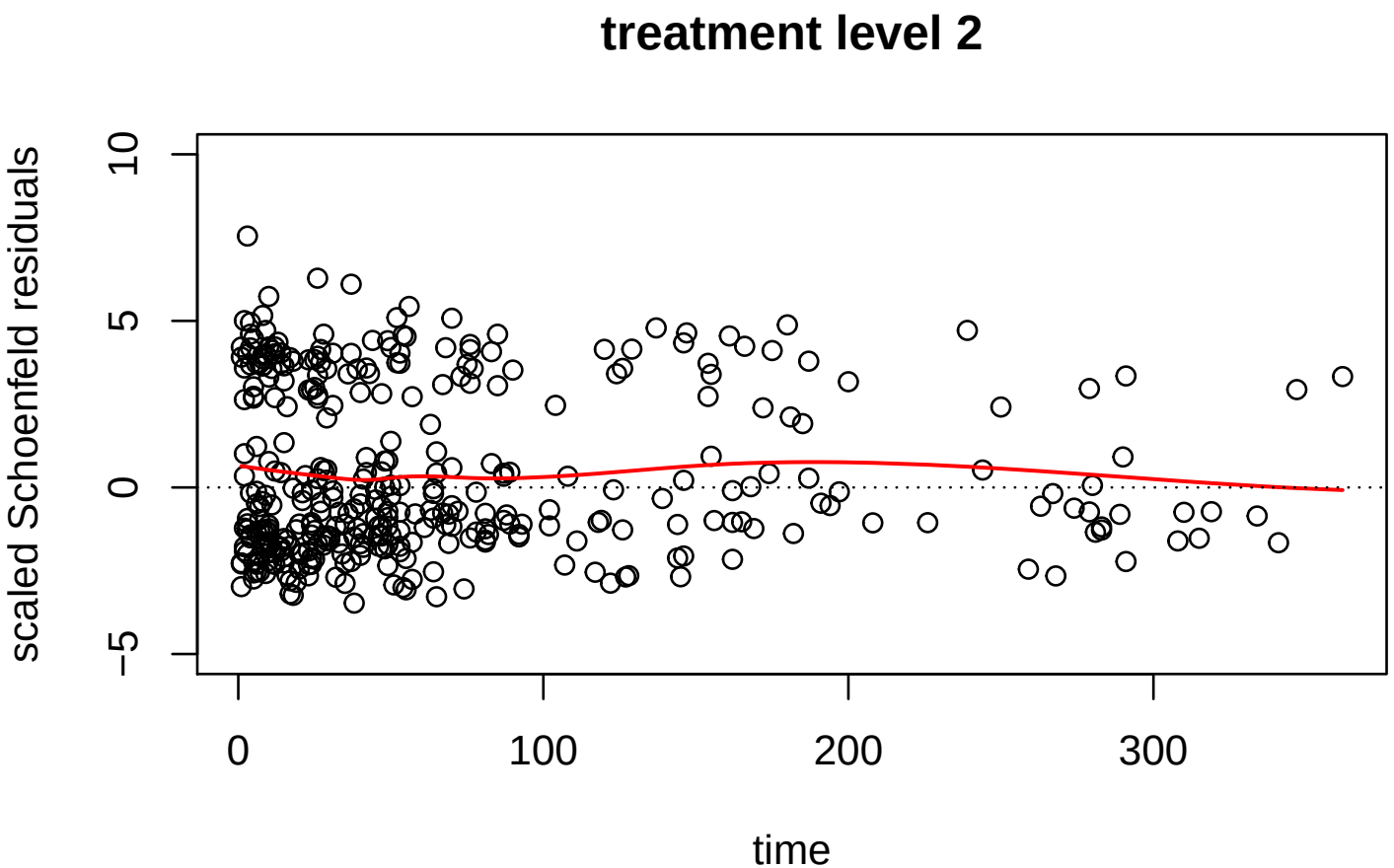
For each covariate, we can plot the scaled Schoenfeld residuals against time to determine whether there is any trend, thus indicating departure from proportional hazards.

```
coxfit = coxph(s.herpes~treat.+type.+age+duration+male,subset=ind,data=herpes)
schoenresid = residuals(coxfit,type="scaledsch")

times = as.numeric(rownames(schoenresid))
plot(times,schoenresid[,1],xlab="time",ylab="scaled Schoenfeld residuals",ylim=c(-5,15))
schoenresid.loess = loess(schoenresid[,1]~times,degree=1)
lines(unique(times),predict(schoenresid.loess,unique(times)),col=2,lwd=1.5)
abline(h=0,lty=3)
```

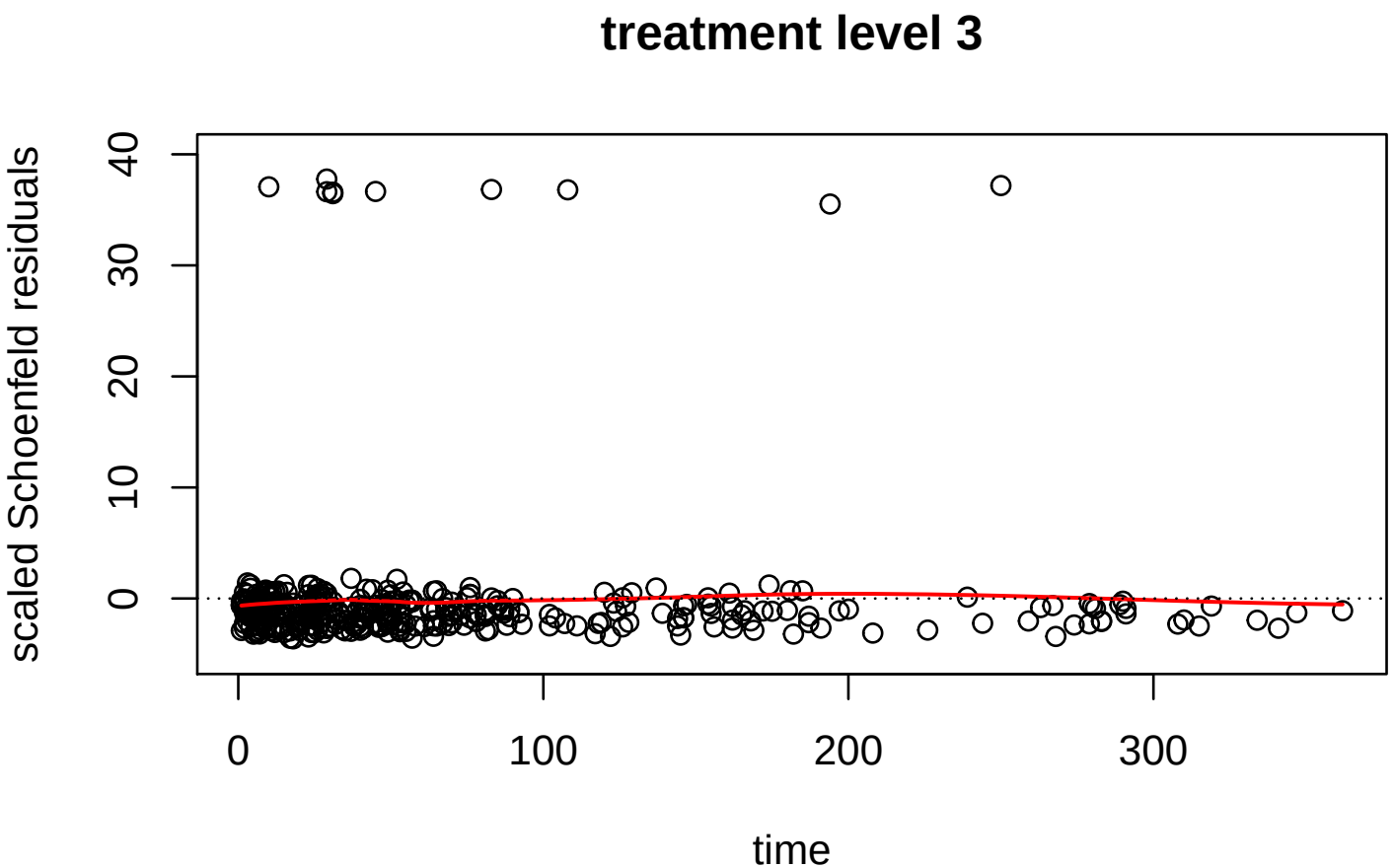
Diagnostic tools



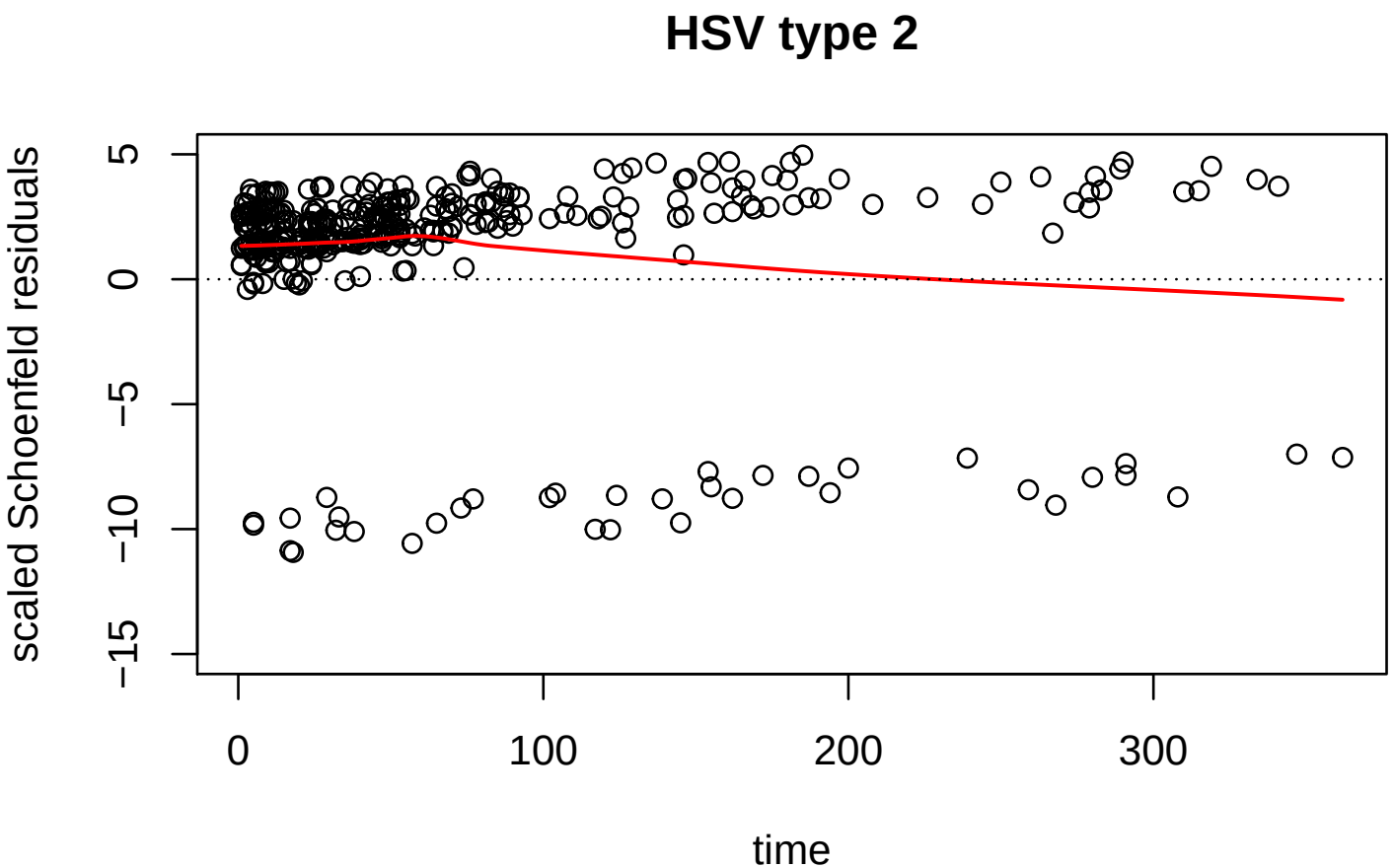
Diagnostic tools



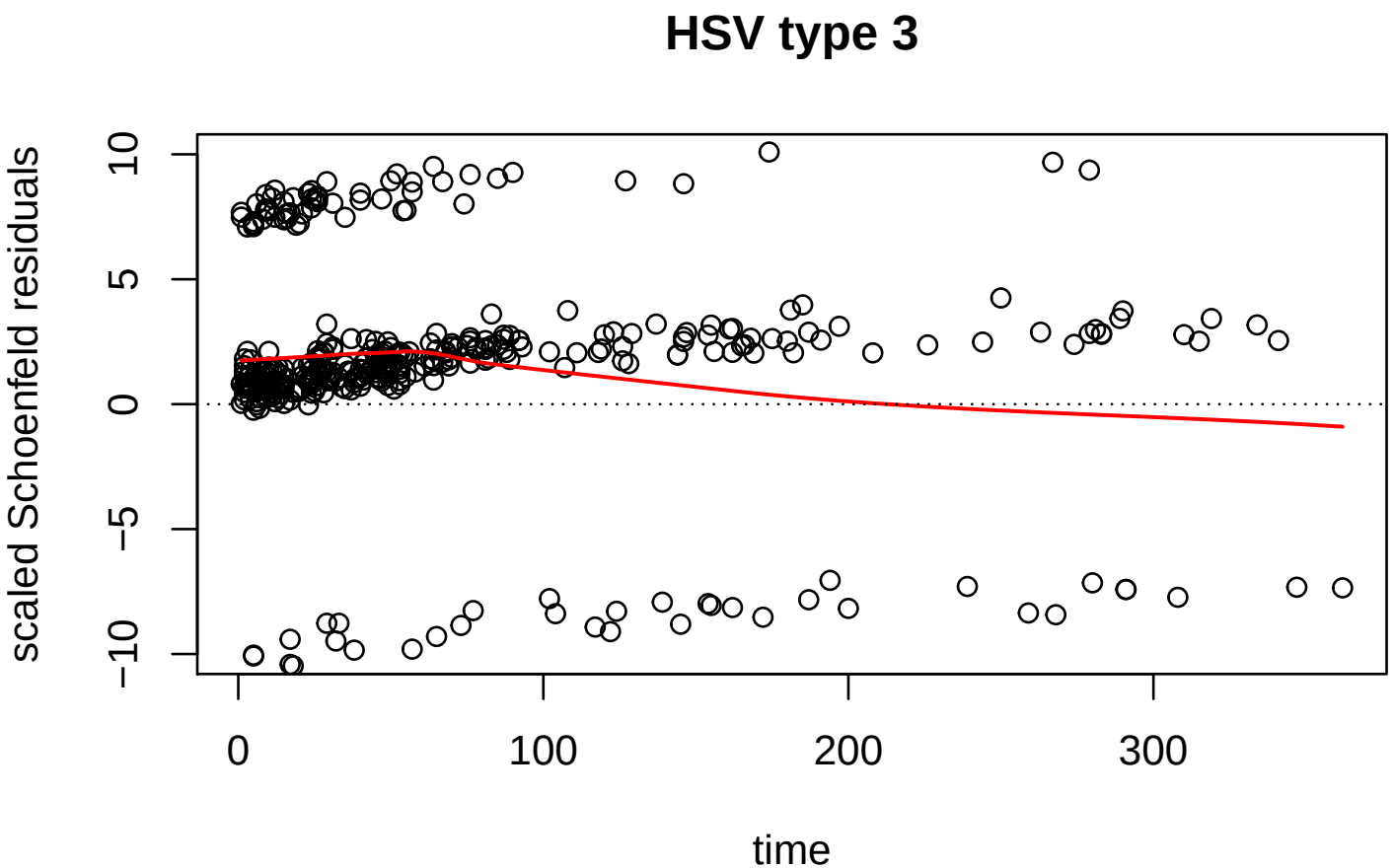
Diagnostic tools



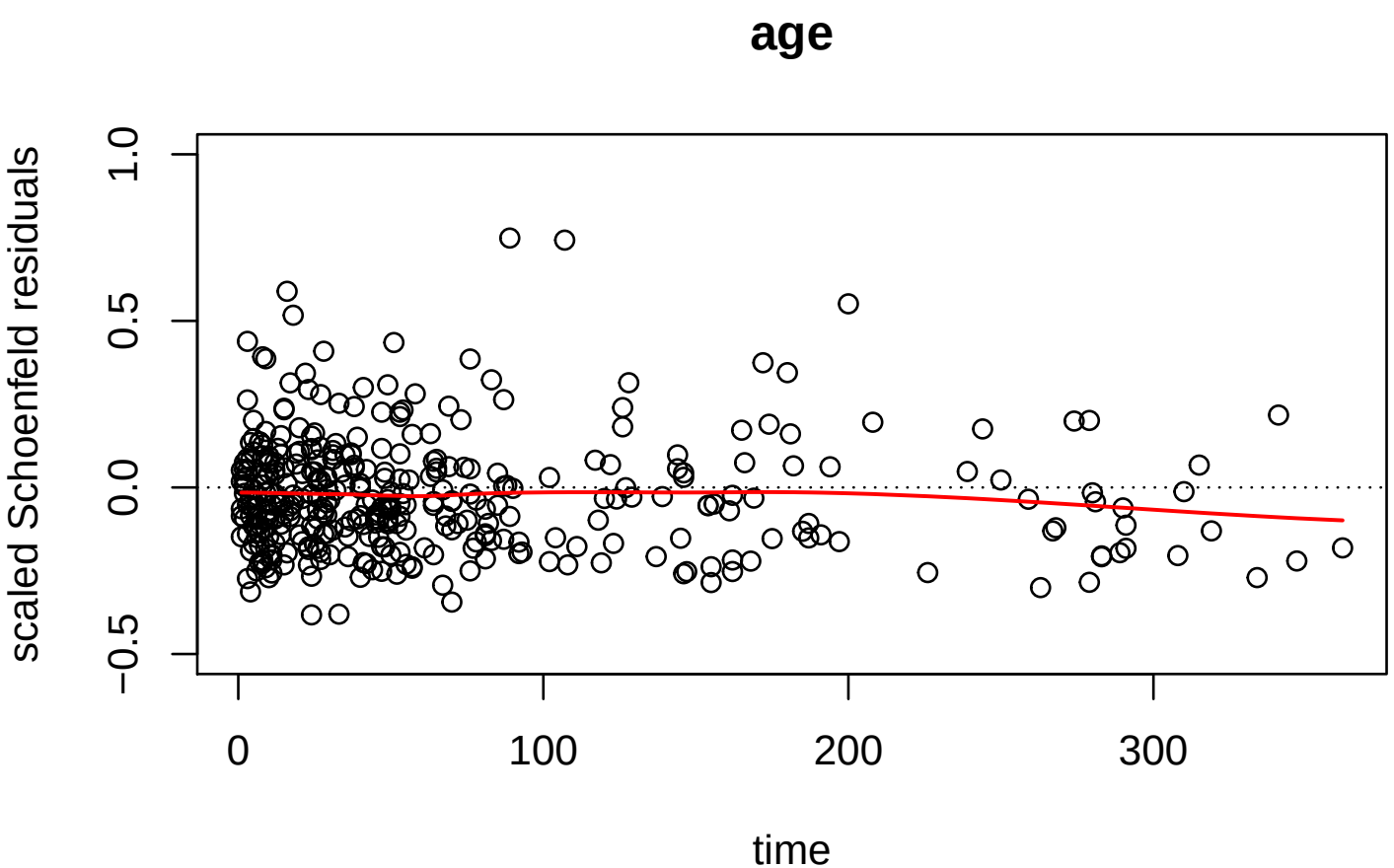
Diagnostic tools



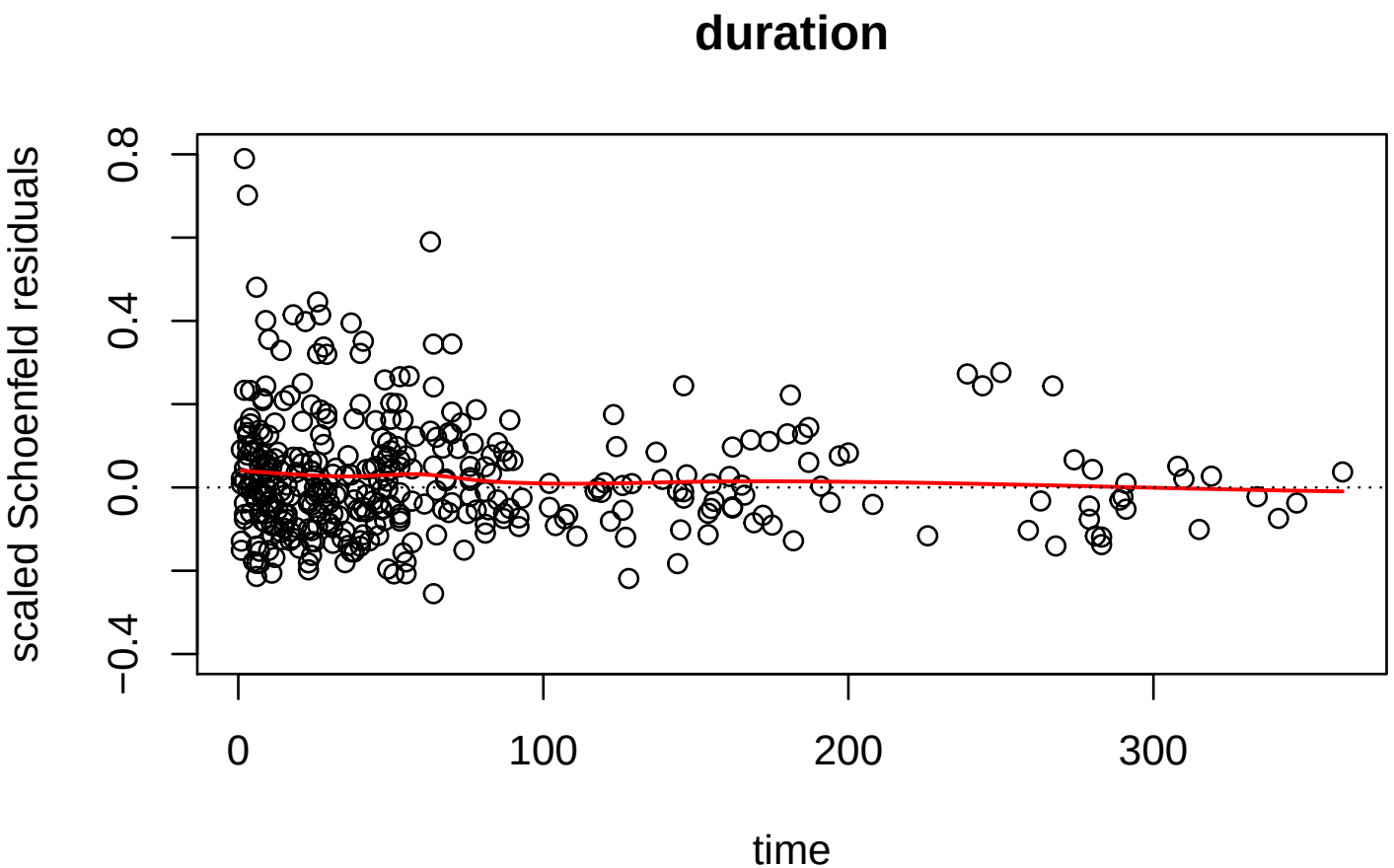
Diagnostic tools



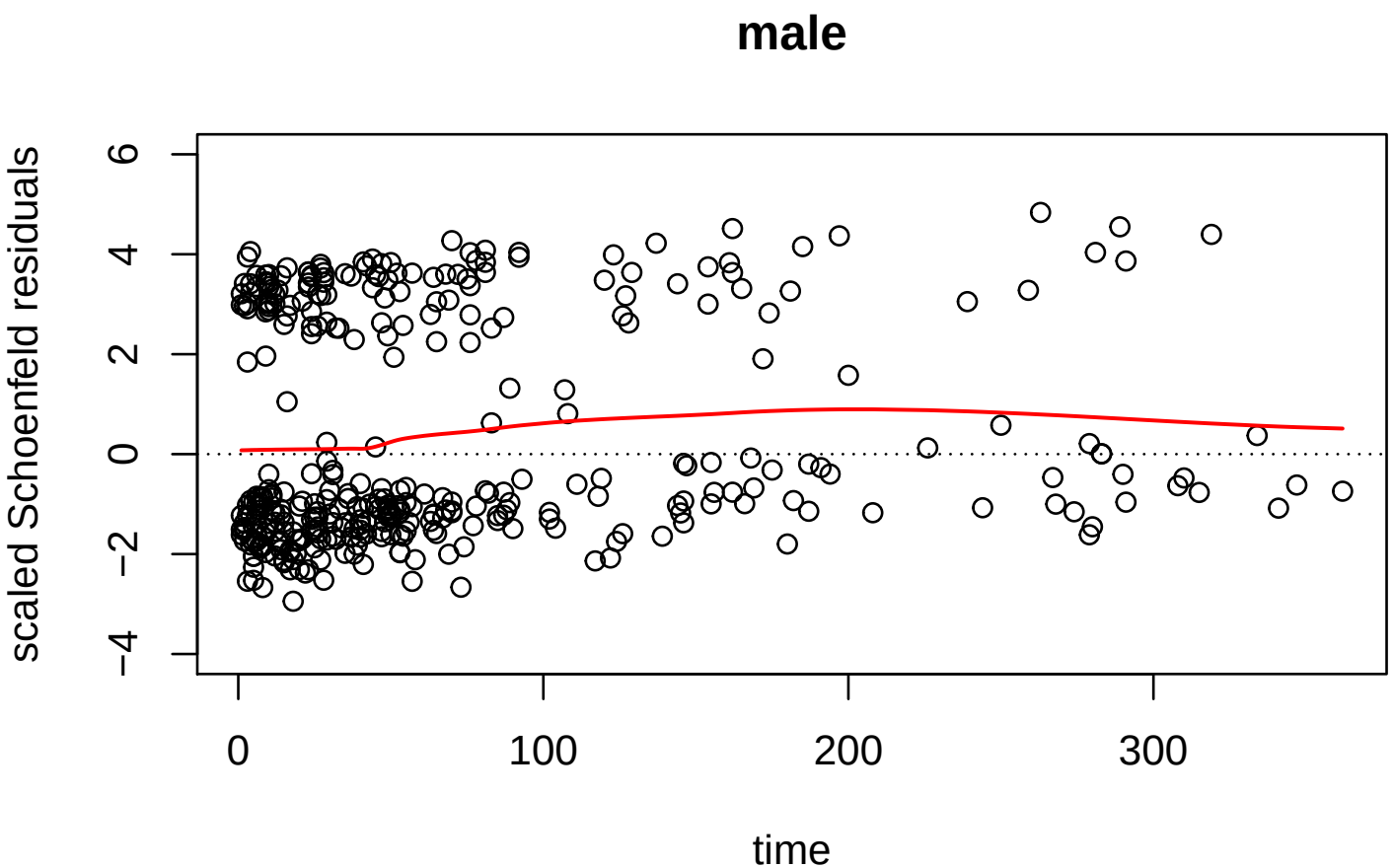
Diagnostic tools



Diagnostic tools



Diagnostic tools



Diagnostic tools

We can formally test whether there is a significant trend in these plots.

```
cox.zph(coxfit)
```

```
##              rho  chisq      p
## treat.1   0.0217  0.183 0.6684
## treat.2  -0.0212  0.182 0.6699
## treat.3   0.0250  0.231 0.6308
## type.2   -0.0932  3.131 0.0768
## type.3   -0.1093  4.274 0.0387
## age      -0.0397  0.553 0.4569
## duration -0.0869  3.145 0.0762
## male      0.1117  4.750 0.0293
## GLOBAL        NA 13.366 0.0999
```

Diagnostic tools

For simplicity, suppose there is a single covariate Z.

If rather than the proportional hazards model $h(t \mid z) = h_0(t)e^{\beta z}$, the true conditional hazard function is given by

$$h(t \mid z) = h_0(t)e^{\beta(t)z} \ ,$$

which we refer to as an **extended Cox model**. This is no longer a proportional hazards model since hazard ratios vary in time.

An extended Cox model with time-varying coefficient $\beta(t) = \beta + \gamma g(t)$, where g is a fixed and pre-determined function of time, can be fit very simply:

1. create a new variable $Z_i^*(t) := g(t) \cdot Z_i$;
2. fit a proportional hazards model with time-varying covariates Z_i and $Z_i^*(t)$; (we will see how this is done in the next chapter)
3. the estimated coefficient associated to Z^* is an estimate of γ.

Diagnostic tools

Time-varying coefficients allow relaxation of the proportional hazards assumption.

Common forms for g include $g(t) = t - t_0$, $g(t) = \log\left(\frac{t}{t_0}\right)$ and $g(t) = I(t > t_0)$ for some fixed $t_0 \geqslant 0$.

Time-varying coefficients can be an extremely useful tool in practice.

1. they can be used to test for lack of proportional hazards:

   if we postulate $\beta(t) = \beta + \gamma g(t)$ *for some function* g *and we reject* $\mathcal{H}_0 : \gamma = 0$ *vs* $\mathcal{H}_1 : \gamma \neq 0$, *proportionality of hazards can also be rejected;*

2. they allow a more appropriate modeling of situations where proportional hazards are unlikely (e.g., vaccine trials);

If $\beta(t) = \beta + \gamma g(t)$, scaled Schoenfeld residual at time t have mean $\approx \gamma g(t)$.

After regressing scaled Schoenfeld residuals linearly against time, if the slope is significantly different from zero, we may reject proportionality of hazards.

Diagnostic tools

```
herpes$treat1 = as.numeric(herpes$treat==1)
herpes$treat2 = as.numeric(herpes$treat==2)
herpes$treat3 = as.numeric(herpes$treat==3)

coxfit.tv1 = coxph(s.herpes~tt(treat1)+tt(treat2)+tt(treat3)+
                           treat1+treat2+treat3+type.+age+duration+male,
                   tt=function(x,t,...) x*t,data=herpes)
summary(coxfit.tv1)$coef
```

```
##                     coef exp(coef)    se(coef)           z     Pr(>|z|)
## tt(treat1)  0.0008972662 1.0008977 0.002320120  0.38673265 6.989541e-01
## tt(treat2)  0.0007035425 1.0007038 0.001552665  0.45311921 6.504629e-01
## tt(treat3)  0.0009072752 1.0009077 0.003932172  0.23073131 8.175235e-01
## treat1     -0.0066243367 0.9933976 0.267907943 -0.02472617 9.802734e-01
## treat2      0.3517615396 1.4215695 0.158508479  2.21919699 2.647333e-02
## treat3     -0.3176210951 0.7278785 0.455300281 -0.69760795 4.854224e-01
## type.2      1.2429659076 3.4658777 0.184505354  6.73674708 1.619715e-11
## type.3      1.5882858177 4.8953502 0.220028844  7.21853457 5.254686e-13
## age        -0.0236519298 0.9766256 0.009294621 -2.54469007 1.093748e-02
## duration    0.0235134019 1.0237920 0.006998682  3.35969005 7.802996e-04
## male        0.3065983621 1.3587951 0.114468585  2.67844983 7.396381e-03
```

Diagnostic tools

```
herpes$type2 = as.numeric(herpes$type==2)
herpes$type3 = as.numeric(herpes$type==3)

coxfit.tv2 = coxph(s.herpes~tt(type2)+tt(type3)+
                           type2+type3+treat.+age+duration+male,
                   tt=function(x,t,...) x*t,data=herpes)
summary(coxfit.tv2)$coef

##                   coef exp(coef)    se(coef)          z     Pr(>|z|)
## tt(type2) -0.004275033 0.9957341 0.001855010 -2.3045872 2.118970e-02
## tt(type3) -0.006849077 0.9931743 0.002896231 -2.3648241 1.803863e-02
## type2      1.729167667 5.6359610 0.297637084  5.8096513 6.260310e-09
## type3      2.204021144 9.0613774 0.336099358  6.5576476 5.466316e-11
## treat.1    0.032451981 1.0329843 0.201304509  0.1612084 8.719293e-01
## treat.2    0.383405354 1.4672727 0.125037154  3.0663314 2.167030e-03
## treat.3   -0.262366098 0.7692294 0.326158580 -0.8044127 4.211587e-01
## age       -0.021418859 0.9788089 0.009287549 -2.3061907 2.109998e-02
## duration   0.022920860 1.0231856 0.006982247  3.2827343 1.028055e-03
## male       0.291136925 1.3379478 0.114423048  2.5443906 1.094686e-02
```

Diagnostic tools

```
coxfit.tv3 = coxph(s.herpes~tt(duration)+
                           duration+treat.+type.+age+male,
                   tt=function(x,t,...) x*t,data=herpes)
summary(coxfit.tv3)$coef
```

```
##                       coef exp(coef)     se(coef)          z     Pr(>|z|)
## tt(duration) -0.0001521433 0.9998479 9.560394e-05 -1.5913915 1.115215e-01
## duration      0.0319501430 1.0324660 8.600782e-03  3.7147951 2.033684e-04
## treat.1       0.0511178138 1.0524469 2.013069e-01  0.2539297 7.995498e-01
## treat.2       0.3904507868 1.4776467 1.245684e-01  3.1344279 1.721896e-03
## treat.3      -0.2456221849 0.7822177 3.261981e-01 -0.7529847 4.514591e-01
## type.2        1.2253624254 3.4054001 1.828452e-01  6.7016372 2.060974e-11
## type.3        1.5690512802 4.8020902 2.187481e-01  7.1728678 7.344125e-13
## age          -0.0242605817 0.9760313 9.302436e-03 -2.6079816 9.107785e-03
## male          0.3051217027 1.3567901 1.143449e-01  2.6684338 7.620579e-03
```

Diagnostic tools

```
coxfit.tv4 = coxph(s.herpes~tt(male)+
                           male+treat.+type.+age+duration,
                   tt=function(x,t,...) x*t,data=herpes)
summary(coxfit.tv4)$coef
```

```
##                  coef exp(coef)    se(coef)          z     Pr(>|z|)
## tt(male)  0.002957411 1.0029618 0.001423233  2.0779526 3.771372e-02
## male      0.119482723 1.1269138 0.147593131  0.8095412 4.182039e-01
## treat.1   0.053680253 1.0551472 0.200845432  0.2672715 7.892602e-01
## treat.2   0.407985845 1.5037859 0.124921127  3.2659475 1.090985e-03
## treat.3  -0.261466406 0.7699217 0.326292289 -0.8013257 4.229431e-01
## type.2    1.285915490 3.6179787 0.186580309  6.8920215 5.500489e-12
## type.3    1.609680345 5.0012123 0.220272040  7.3076925 2.717826e-13
## age      -0.023930090 0.9763540 0.009301079 -2.5728294 1.008709e-02
## duration  0.023911101 1.0241993 0.007000324  3.4157136 6.361513e-04
```

Diagnostic tools

```
coxfit.tv5 = coxph(s.herpes~tt(age)+
                           age+treat.+type.+duration+male,
                   tt=function(x,t,...) x*t,data=herpes)
summary(coxfit.tv5)$coef
```

```
##                   coef exp(coef)     se(coef)          z     Pr(>|z|)
## tt(age)  -6.366566e-05 0.9999363 0.0001069815 -0.5951092 5.517705e-01
## age      -1.836830e-02 0.9817994 0.0122829961 -1.4954248 1.348036e-01
## treat.1   6.394335e-02 1.0660320 0.2013209394  0.3176190 7.507740e-01
## treat.2   3.980357e-01 1.4888972 0.1249358818  3.1859198 1.442946e-03
## treat.3  -2.419612e-01 0.7850867 0.3264213194 -0.7412542 4.585393e-01
## type.2    1.235069e+00 3.4386170 0.1829228357  6.7518599 1.459621e-11
## type.3    1.585215e+00 4.8803396 0.2192928659  7.2287569 4.873879e-13
## duration  2.351831e-02 1.0237971 0.0070124982  3.3537711 7.971827e-04
## male      3.053576e-01 1.3571102 0.1144953437  2.6669871 7.653462e-03
```

Diagnostic tools

Suppose that Z is a categorical covariate.

A simple visualization technique, referred to as log-log plots, builds on the fact that

$$\log\{-\log S(t \mid z)\} = \log\{-\log S_0(t)\} + \beta z \quad :$$

1. compute the KM estimator for each level of Z;
2. plot the complementary log-log transform of each estimate on the same graph (often against the logarithm of time);
3. if the curves fail to be parallel, proportionality of hazards fails to hold.

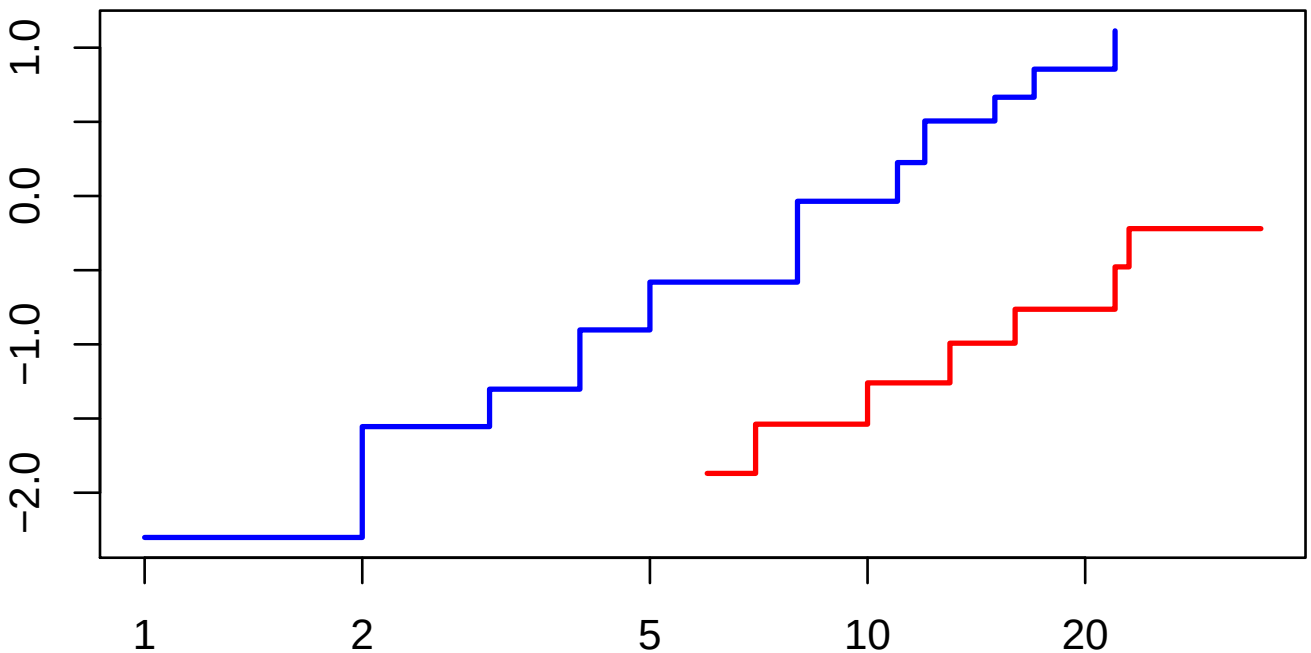
If proportionality of hazards holds for a continuous covariate, it need not hold for a discretization of this covariate and vice-versa. This limits the applicability of this tool.

Slightly more complex procedures are possible to deal with the need to adjust for many covariates (including, possibly, some continuous covariates).

Diagnostic tools

We can illustrate this using data from the original 6MP trial.

```
plot(survfit(s.mp~tx,data=mp),col=c(4,2),lwd=2,fun="cloglog")
```



Diagnostic tools

It may be of interest to determine whether any particular observation inordinately drives coefficient estimates.

How can we do this?

Estimate model coefficient β using all the data – leading to $\hat{\beta}$ – and using all the data except for the i^{th} observation – leading to $\hat{\beta}_{(i)}$.

The standardized **delta-beta** for the i^{th} observation, also called DFBETA, is defined as

$$\Delta\beta_i := \frac{\hat{\beta} - \hat{\beta}_{(i)}}{\widehat{SE}(\hat{\beta})}$$

If $|\Delta\hat{\beta}_i|$ is relatively large, the i^{th} observation is considered influential.

A few observations:

- rerunning the model n times to get all delta-betas is unwieldy, so an approximation based on so-called influence functions is used in practice;
- both censored and uncensored observations impact coefficient estimates.

Diagnostic tools

Based on the diagnostics obtained so far, we may choose to fit a proportional hazards model including treatment, age and duration, stratified by sex and HSV type.

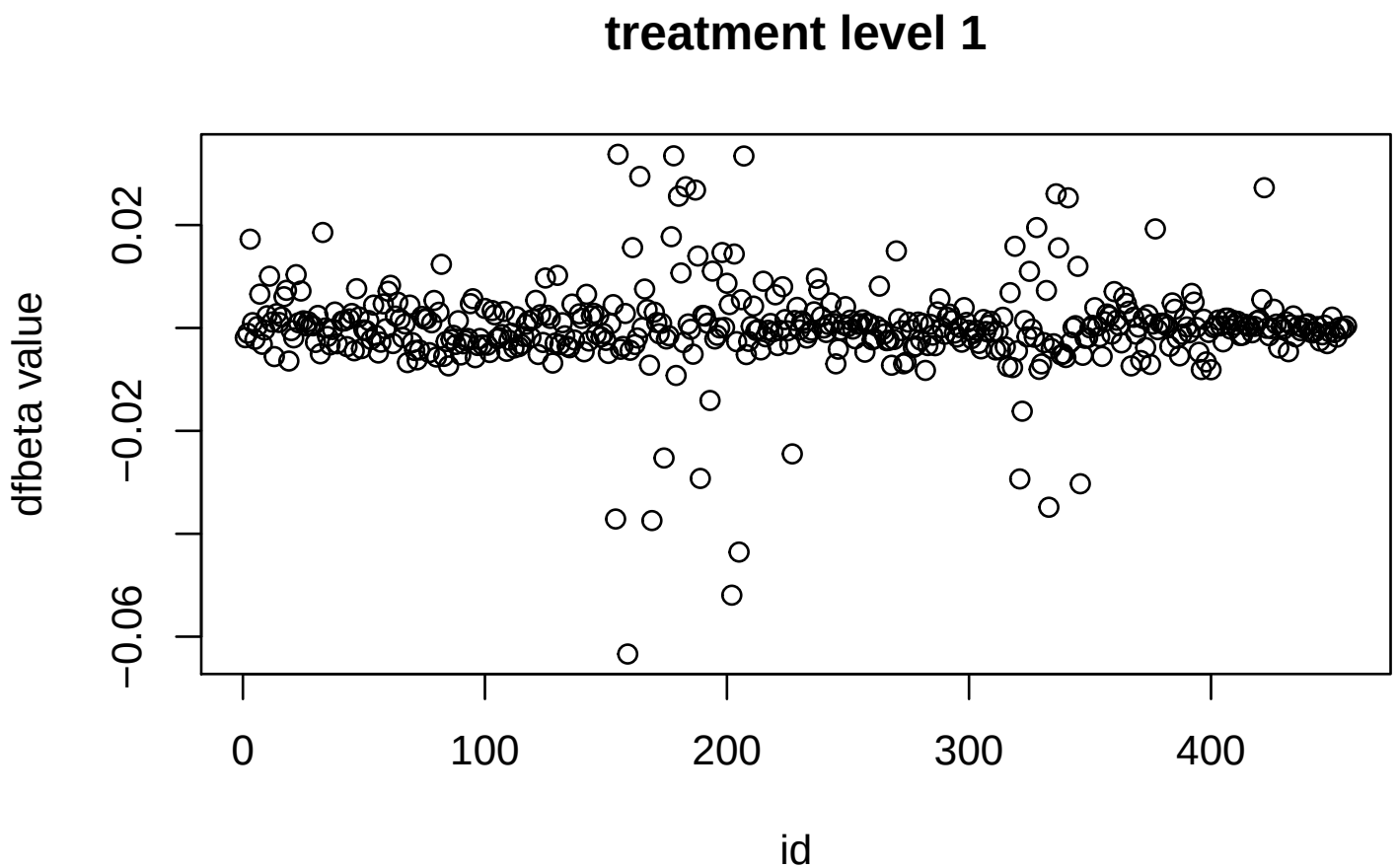
```
ind = which(complete.cases(herpes))
coxfit = coxph(s.herpes~treat.+duration+age+strata(type2)+strata(type3)+strata(male),subset=ind,data=herpe
dfbetas = residuals(coxfit,type="dfbeta")
colnames(dfbetas) = names(coef(coxfit))
summary(coxfit)$coef
```

```
##                  coef exp(coef)    se(coef)           z     Pr(>|z|)
## treat.1   0.007807631 1.0078382 0.202163493  0.03862038 0.9691930537
## treat.2   0.365897579 1.4418076 0.126321758  2.89655231 0.0037728774
## treat.3  -0.225226526 0.7983354 0.327370540 -0.68798654 0.4914612550
## duration  0.023272415 1.0235453 0.006997094  3.32601133 0.0008809835
## age      -0.017954052 0.9822062 0.009489252 -1.89204076 0.0584855501
```

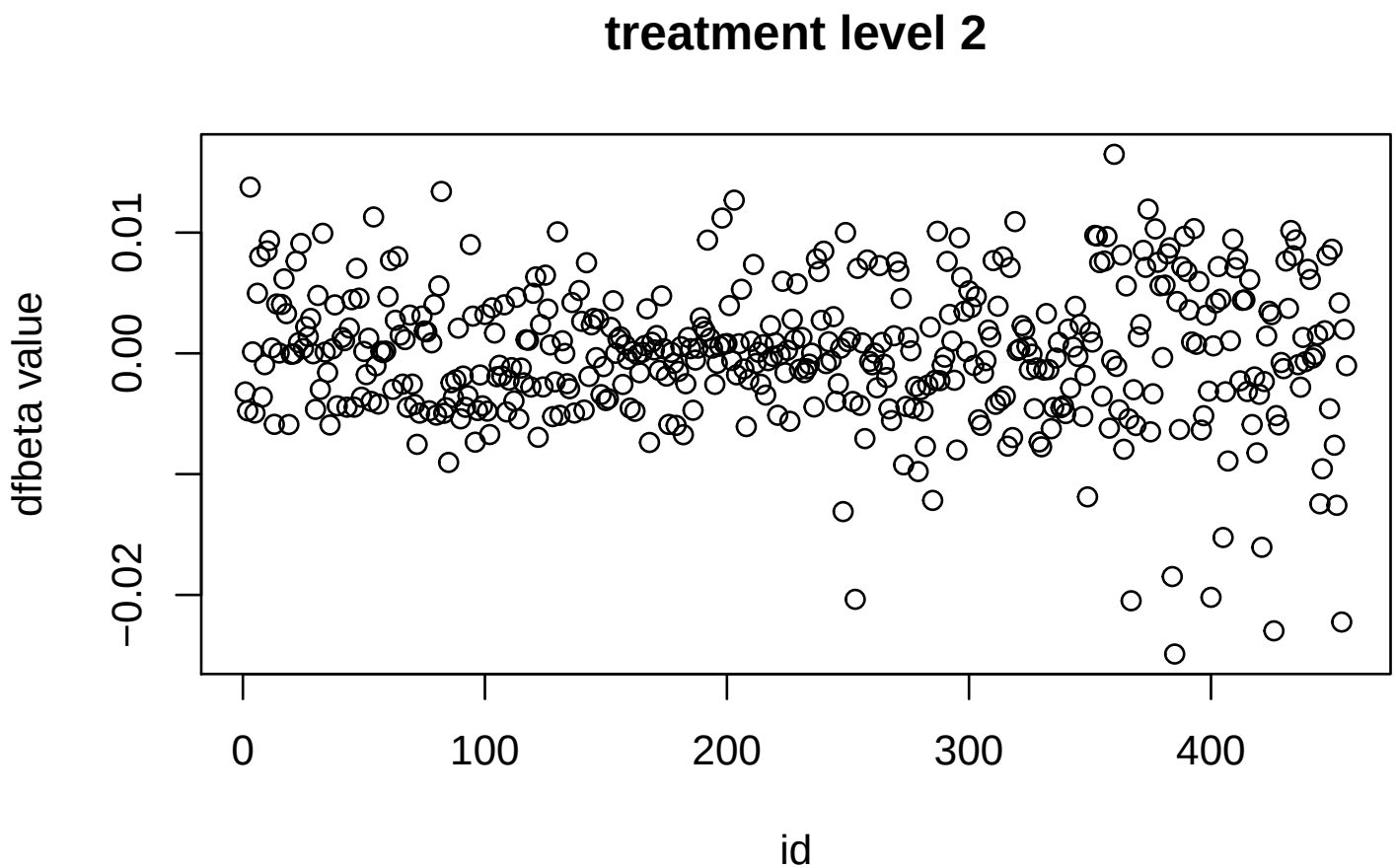
We can plot the DFBETA values for each main covariate in this stratified model.

```
plot(ind,dfbetas[,1],ylab="dfbeta value",xlab="id",main="treatment level 1")
plot(ind,dfbetas[,2],ylab="dfbeta value",xlab="id",main="treatment level 2")
plot(ind,dfbetas[,3],ylab="dfbeta value",xlab="id",main="treatment level 3")
plot(ind,dfbetas[,4],ylab="dfbeta value",xlab="id",main="duration")
plot(ind,dfbetas[,5],ylab="dfbeta value",xlab="id",main="age")
```

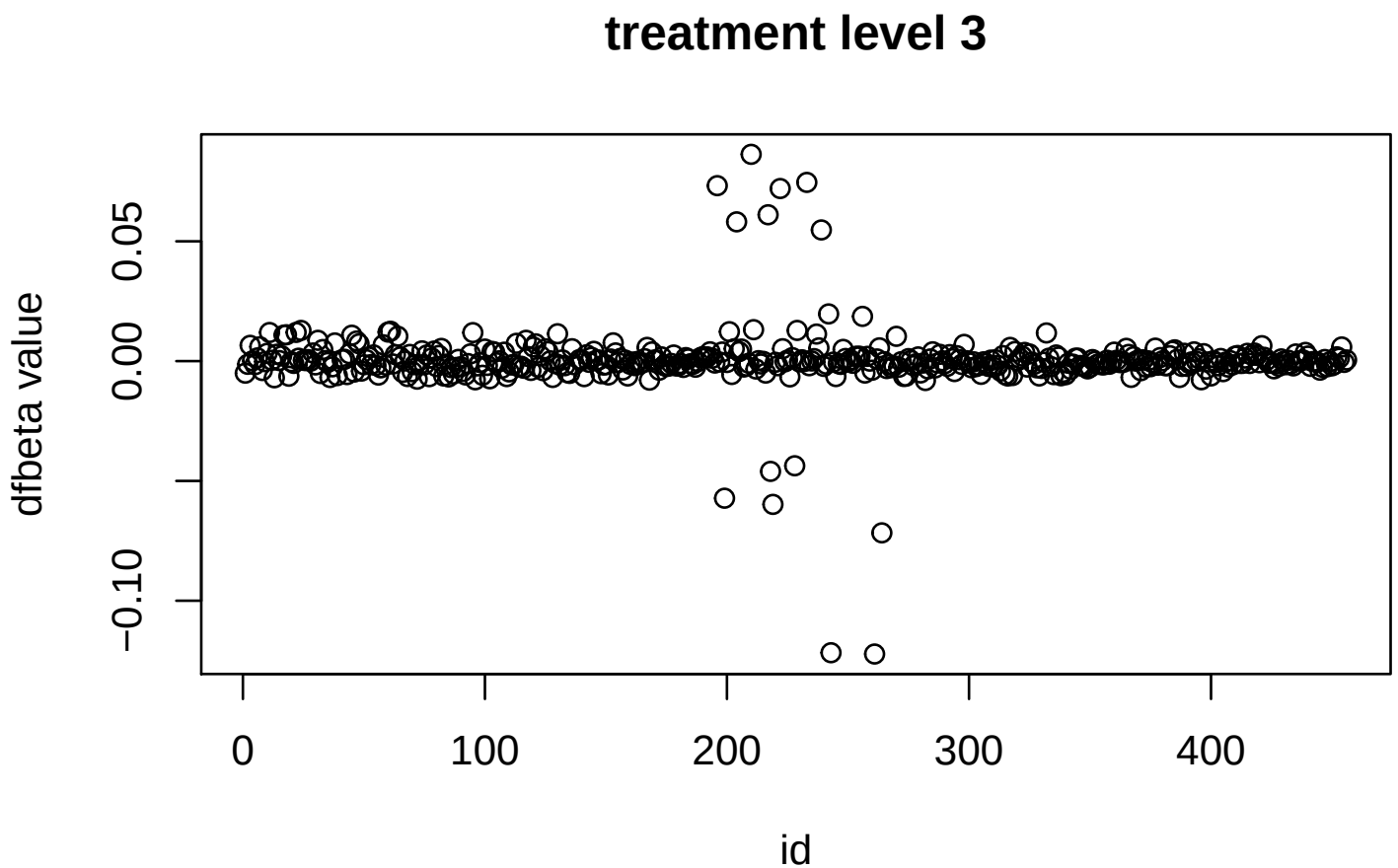
Diagnostic tools



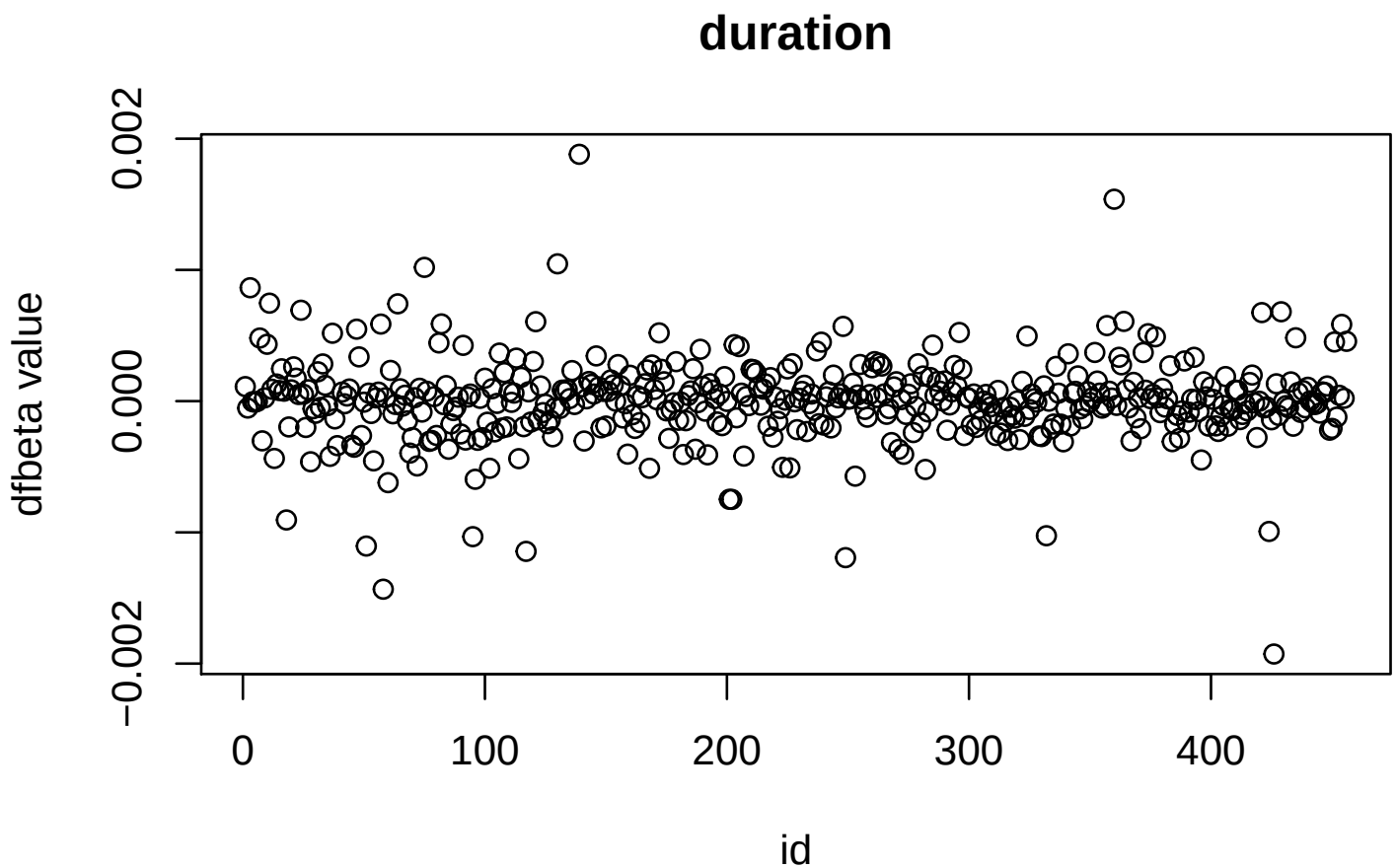
Diagnostic tools



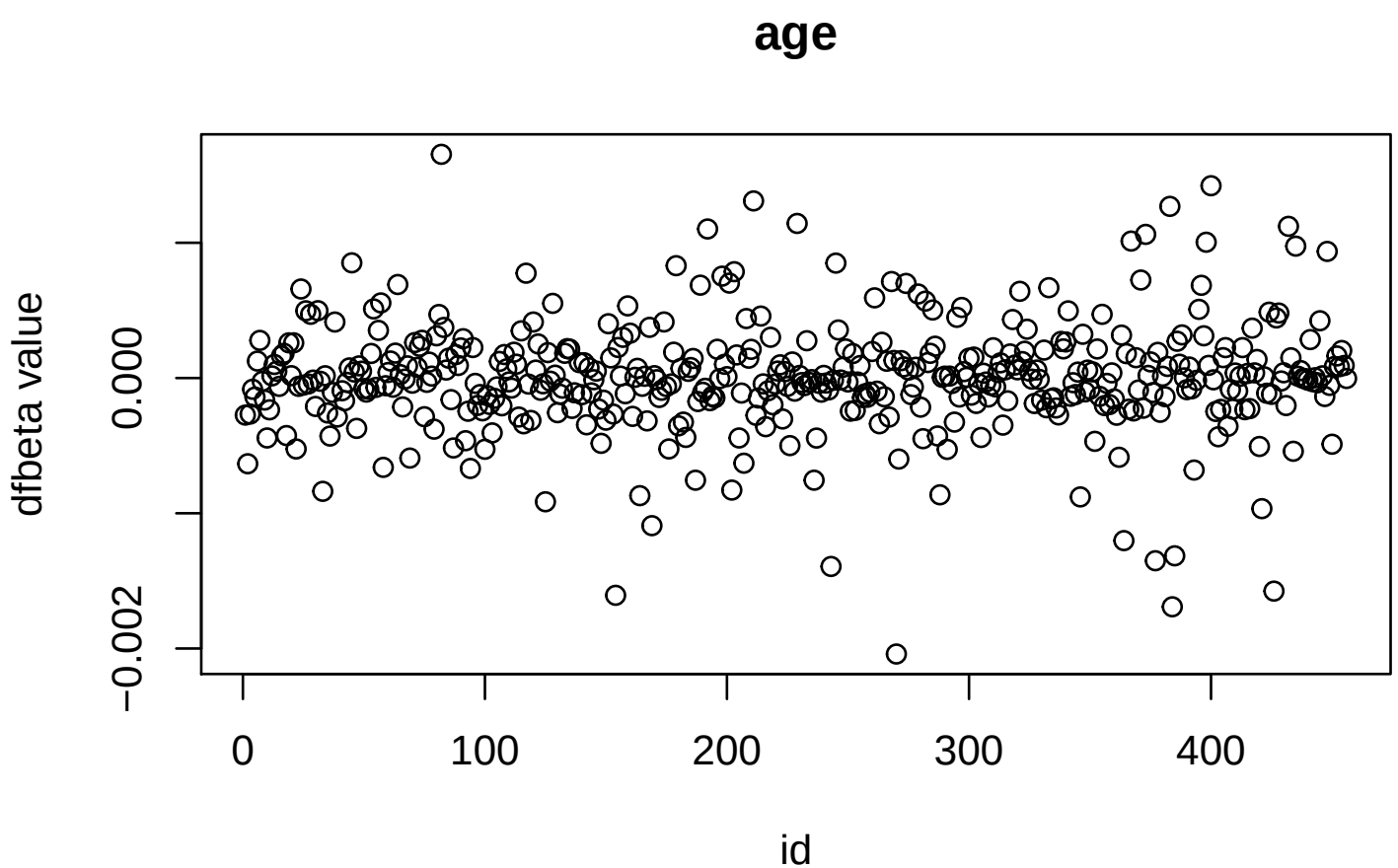
Diagnostic tools



Diagnostic tools



Diagnostic tools



Additional reading and references

ADDITIONAL READING:

- Moore, D.F. *Applied Survival Analysis Using R.* Chapters 5, 6 and 7.
- Hosmer, DW, Lemeshow, S, May, S. *Applied survival analysis.* (2nd ed.) Chapters 3, 4, 6 and 7.2.
- Kleinbaum, DG, Klein, M. *Survival analysis: a self-learning text.* (3rd ed.) Chapters 3, 4 and 5.

REFERENCES:

- Cox, D.R., Snell, E.J. 1968 *A general definition of residuals.* Journal of the Royal Statistical Society, Series B.
- Cox, D.R. 1972 *Regression models and life-tables. (+ discussion)* Journal of the Royal Statistical Society, Series B.
- Therneau, T.M., Grambsch, P.M., Fleming, T.R. 1990 *Martingale-based residuals for survival models.* Biometrika.
- Grambsch, P.M., Therneau, T.M. 1994 *Proportional hazards tests and diagnostics based on weighted residuals.* Biometrika.