



# STHDA

Statistical tools for high-throughput data analysis

Licence:

[Home](#)[Basics +](#)[Data +](#)[Visualize +](#)[Analyze +](#)[Products +](#)[Contribute](#)[Support +](#)[About](#)[Sign in](#)

[Home](#) / [Easy Guides](#) / [R software](#) / [Survival Analysis](#) / [Cox Model Assumptions](#)



Login

Password

Auto connect ☒

Sign in

Register



Forgotten password

## Cox Model Assumptions

Previously, we described the [basic methods for analyzing survival data](#), as well as, the [Cox proportional hazards methods](#) to deal with the situation where several factors impact on the survival process.

In the current article, we continue the series by describing methods to evaluate the validity of the **Cox model assumptions**.



Note that, when used inappropriately, statistical models may give rise to misleading conclusions. Therefore, it's important to check that a given model is an appropriate representation of the data.

### Welcome!

Want to Learn More on R  
Programming and Data  
Science?

Follow us [by Email](#)

by [FeedBurner](#)

### Contents

- [Diagnostics for the Cox model](#)
- [Assessing the validity of a Cox model in R](#)
  - [Installing and loading required R packages](#)
  - [Computing a Cox model](#)
  - [Testing proportional Hazards assumption](#)
  - [Testing influential observations](#)
  - [Testing non linearity](#)

on Social Networks



- [Summary](#)
- [Infos](#)

## Diagnostics for the Cox model

The Cox proportional hazards model makes several assumptions. Thus, it is important to assess whether a fitted Cox regression model adequately describes the data.

Here, we'll discuss three types of diagnostics for the Cox model:

- Testing the proportional hazards assumption.
- Examining influential observations (or outliers).
- Detecting nonlinearity in relationship between the log hazard and the covariates.

In order to check these model assumptions, *Residuals* method are used. The common residuals for the Cox model include:

- *Schoenfeld residuals* to check the proportional hazards assumption
- *Martingale residual* to assess nonlinearity
- *Deviance residual* (symmetric transformation of the Martingale residuals), to examine influential observations

## Assessing the validity of a Cox model in R

### Installing and loading required R packages






We'll use two R packages:

- **survival** for computing survival analyses
- **survminer** for visualizing survival analysis results
- Install the packages

```
install.packages(c("survival", "survminer"))
```

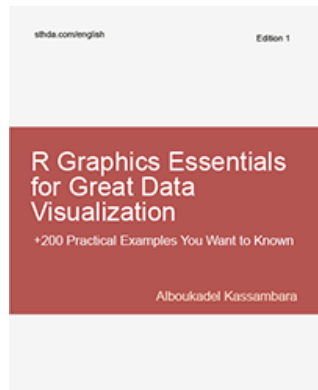
- Load the packages

### R Packages

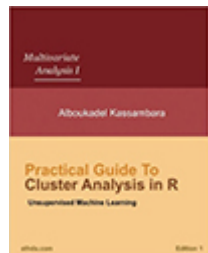
 <b>factoextra</b>	+
 <b>survminer</b>	+
 <b>ggpubr</b>	+
 <b>ggcorrplot</b>	+
 <b>fastqcr</b>	+

### Our Books

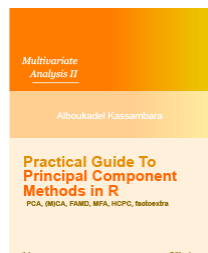
### 3D Plots in R



R Graphics Essentials for  
Great Data Visualization:  
200 Practical Examples  
You Want to Know for  
Data Science  
★ **NEW!!**



Practical Guide to Cluster  
Analysis in R



Practical Guide to

```
library("survival")
library("survminer")
```

## Computing a Cox model

We'll use the lung data sets and the `coxph()` function in the survival package.

To compute a Cox model, type this:

```
library("survival")
res.cox <- coxph(Surv(time, status) ~ age + sex + wt.loss, data = lung)
res.cox
```

Call:

```
coxph(formula = Surv(time, status) ~ age + sex + wt.loss, data = lung)
              coef exp(coef) se(coef)      z      p
age      0.02009   1.02029  0.00966   2.08 0.0377
sex     -0.52103   0.59391  0.17435  -2.99 0.0028
wt.loss  0.00076   1.00076  0.00619   0.12 0.9024
Likelihood ratio test=14.7 on 3 df, p=0.00212
n= 214, number of events= 152
(14 observations deleted due to missingness)
```

## Testing proportional Hazards assumption

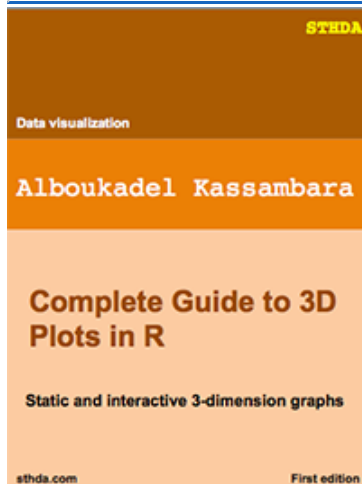
The proportional hazards (PH) assumption can be checked using statistical tests and graphical diagnostics based on the *scaled Schoenfeld residuals*.



In principle, the *Schoenfeld residuals* are independent of time. A plot that shows a non-random pattern against time is evidence of violation of the PH assumption.

The function `cox.zph()` [in the *survival* package] provides a convenient solution to test the proportional hazards assumption for each covariate included in a Cox regression model fit.

## Principal Component Methods in R



### Blogroll

 **Datanovia: Online Data Science Courses**

 **R-Bloggers**

For each covariate, the function `cox.zph()` correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole.

✓ The proportional hazard assumption is supported by a non-significant relationship between residuals and time, and refuted by a significant relationship.

To illustrate the test, we start by computing a Cox regression model using the lung data set [in survival package]:

```
library("survival")
res.cox <- coxph(Surv(time, status) ~ age + sex + wt.loss, data = lung)
res.cox
```

Call:

```
coxph(formula = Surv(time, status) ~ age + sex + wt.loss, data = lung)
      coef exp(coef) se(coef)      z      p
age      0.02009   1.02029  0.00966   2.08 0.0377
sex     -0.52103   0.59391  0.17435  -2.99 0.0028
wt.loss  0.00076   1.00076  0.00619   0.12 0.9024
Likelihood ratio test=14.7 on 3 df, p=0.00212
n= 214, number of events= 152
(14 observations deleted due to missingness)
```

To test for the proportional-hazards (PH) assumption, type this:

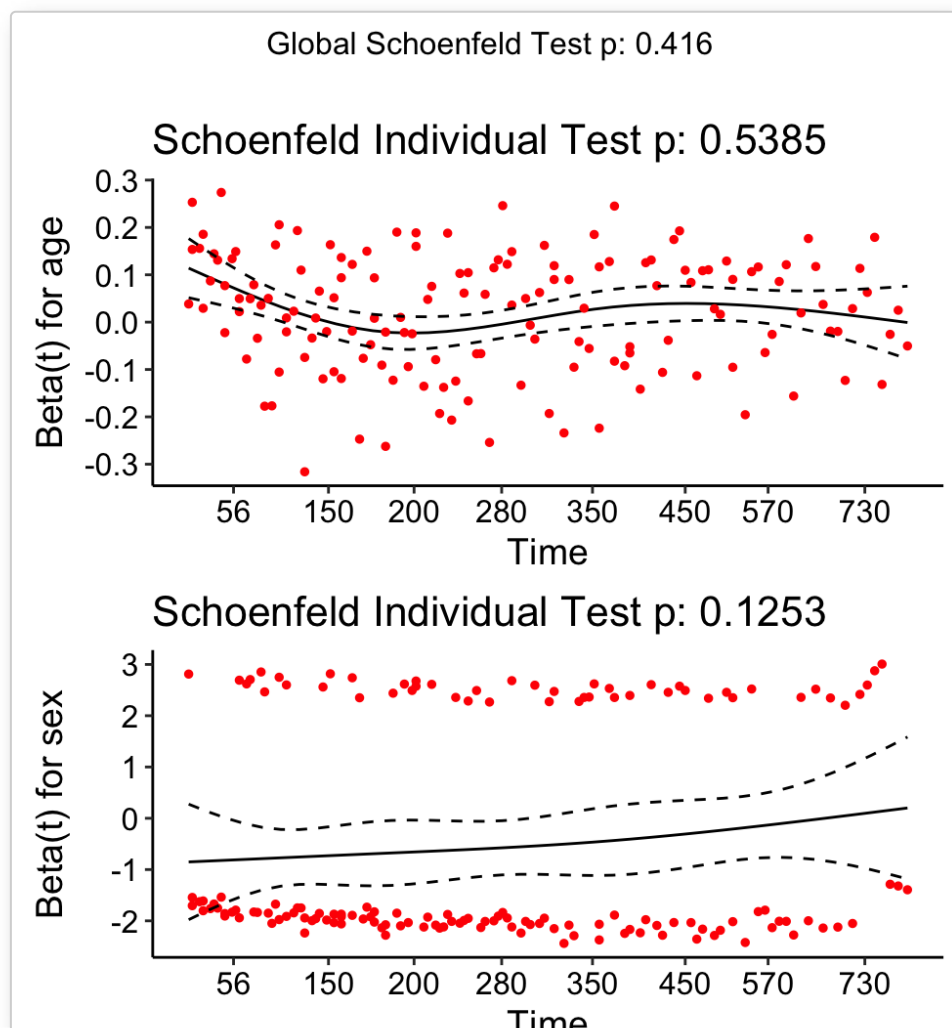
```
test.ph <- cox.zph(res.cox)
test.ph
```

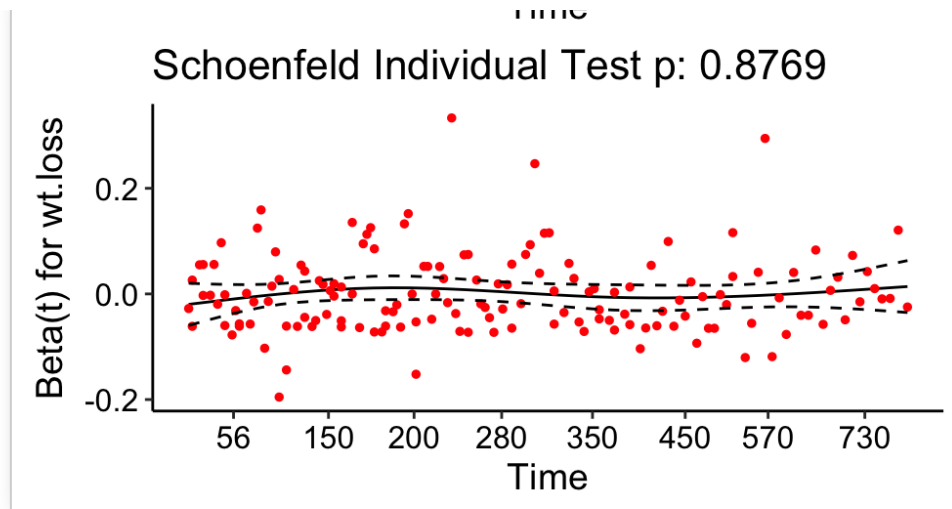
```
      rho chisq      p
age    -0.0483 0.378 0.538
sex      0.1265 2.349 0.125
wt.loss  0.0126 0.024 0.877
GLOBAL      NA 2.846 0.416
```

✓ From the output above, the test is not statistically significant for each of the covariates, and the global test is also not statistically significant. Therefore, we can assume the proportional hazards.

It's possible to do a graphical diagnostic using the function `ggcoxzph()` [in the *survminer* package], which produces, for each covariate, graphs of the scaled Schoenfeld residuals against the transformed time.

```
ggcoxzph(test.ph)
```





In the figure above, the solid line is a smoothing spline fit to the plot, with the dashed lines representing a  $\pm 2$ -standard-error band around the fit.



Note that, systematic departures from a horizontal line are indicative of non-proportional hazards, since proportional hazards assumes that estimates  $\beta_1, \beta_2, \beta_3$  do not vary much over time.

From the graphical inspection, there is no pattern with time. The assumption of proportional hazards appears to be supported for the covariates sex (which is, recall, a two-level factor, accounting for the two bands in the graph), wt.loss and age.

Another graphical methods for checking proportional hazards is to plot  $\log(-\log(S(t)))$  vs.  $t$  or  $\log(t)$  and look for parallelism. This can be done only for categorical covariates.

A violations of proportional hazards assumption can be resolved by:

- Adding covariate\*time interaction
- Stratification

Stratification is usefull for “nuisance” confounders, where you do not care to estimate the effect. You cannot examine the effects of the stratification variable (John Fox & Sanford Weisberg).

To read more about how to accomodate with non-proportional hazards, read the following articles:

- Jadwiga Borucka, PAREXEL, Warsaw, Poland. [Extensions of cox model for non-proportional hazards purpose](#). 2013.
- John Fox & Sanford Weisberg. [Cox Proportional-Hazards Regression for Survival Data in R](#).
- Max Gordon. [Dealing with non-proportional hazards in R](#). March 29, 2016.

## Testing influential observations

To test influential observations or outliers, we can visualize either:

- the *deviance residuals* or
- the *dfbeta* values

The function `ggcoxdiagnostics()` [in *survminer* package] provides a convenient solution for checking influential observations. The simplified format is as follow:

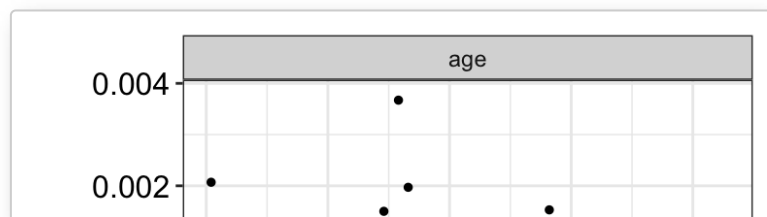
```
ggcoxdiagnostics(fit, type = , linear.predictions = TRUE)
```

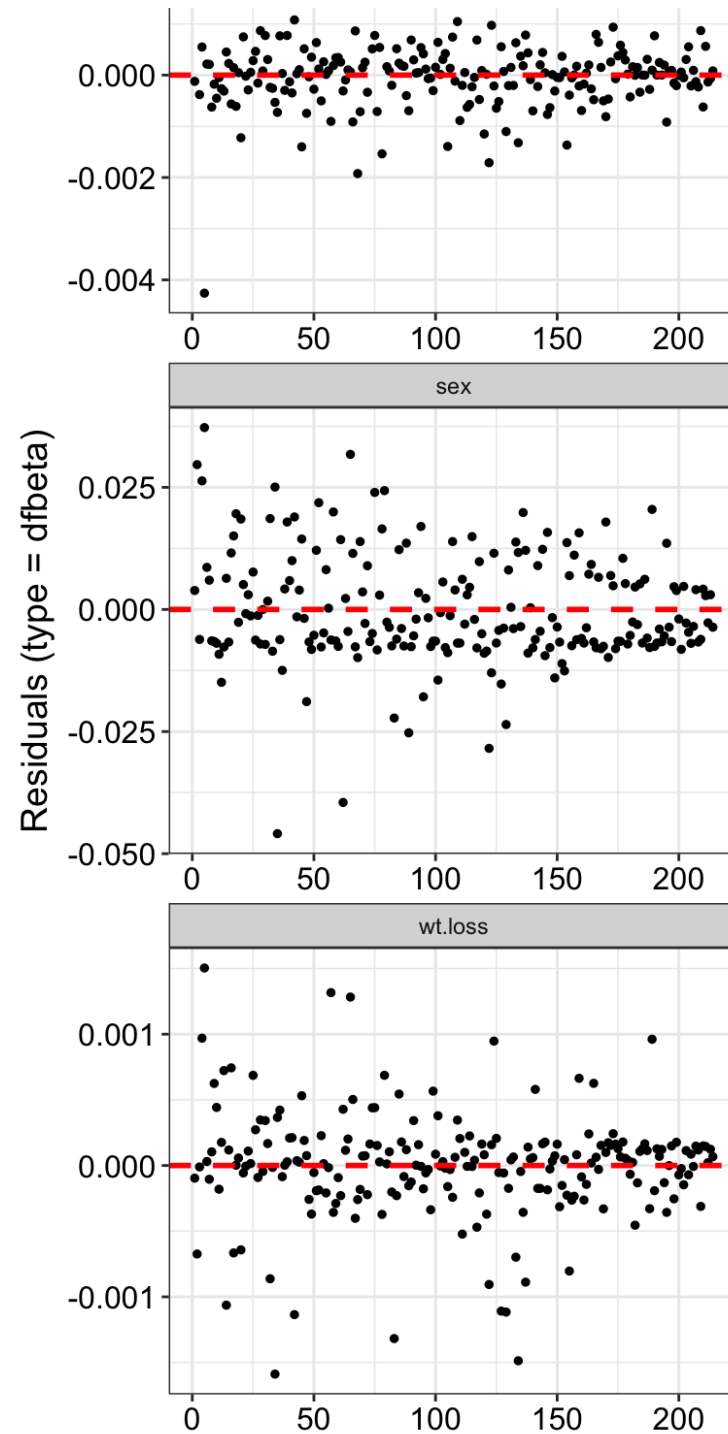
- `fit`: an object of class `coxph.object`
- `type`: the type of residuals to present on Y axis. Allowed values include one of `c("martingale", "deviance", "score", "schoenfeld", "dfbeta", "dfbetas", "scaledsch", "partial")`.
- `linear.predictions`: a logical value indicating whether to show linear predictions for observations (TRUE) or just indexed of observations (FALSE) on X axis.

Specifying the argument `type = "dfbeta"`, plots the estimated changes in the regression coefficients upon deleting each observation in turn; likewise, `type="dfbetas"` produces the estimated changes in the coefficients divided by their standard errors.

For example:

```
ggcoxdiagnostics(res.cox, type = "dfbeta",  
                 linear.predictions = FALSE, ggtheme = theme_bw())
```







## The index number of observations

(Index plots of dfbeta for the Cox regression of time to death on age, sex and wt.loss)

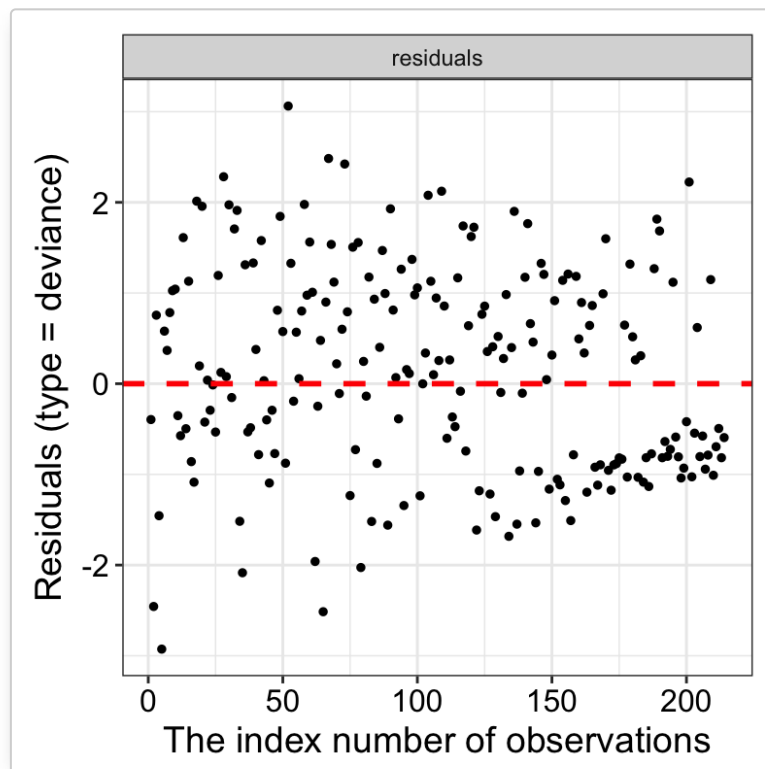
The above index plots show that comparing the magnitudes of the largest dfbeta values to the regression coefficients suggests that none of the observations is terribly influential individually, even though some of the dfbeta values for age and wt.loss are large compared with the others.

It's also possible to check outliers by visualizing the deviance residuals. The deviance residual is a normalized transform of the martingale residual. These residuals should be roughly symmetrically distributed about zero with a standard deviation of 1.

- Positive values correspond to individuals that “died too soon” compared to expected survival times.
- Negative values correspond to individual that “lived too long”.
- Very large or small values are outliers, which are poorly predicted by the model.

Example of deviance residuals:

```
ggcoxdiagnostics(res.cox, type = "deviance",  
                 linear.predictions = FALSE, ggtheme = theme_bw())
```



✓ The pattern looks fairly symmetric around 0.

## Testing non linearity

Often, we assume that continuous covariates have a linear form. However, this assumption should be checked.

Plotting the *Martingale residuals* against continuous covariates is a common approach used to detect *nonlinearity* or, in other words, to assess the functional form of a covariate. For a given continuous covariate, patterns in the plot may suggest that the variable is not properly fit.

Nonlinearity is not an issue for categorical variables, so we only examine plots of martingale residuals and partial residuals against a continuous variable.

Martingale residuals may present any value in the range  $(-\infty, +1)$ :

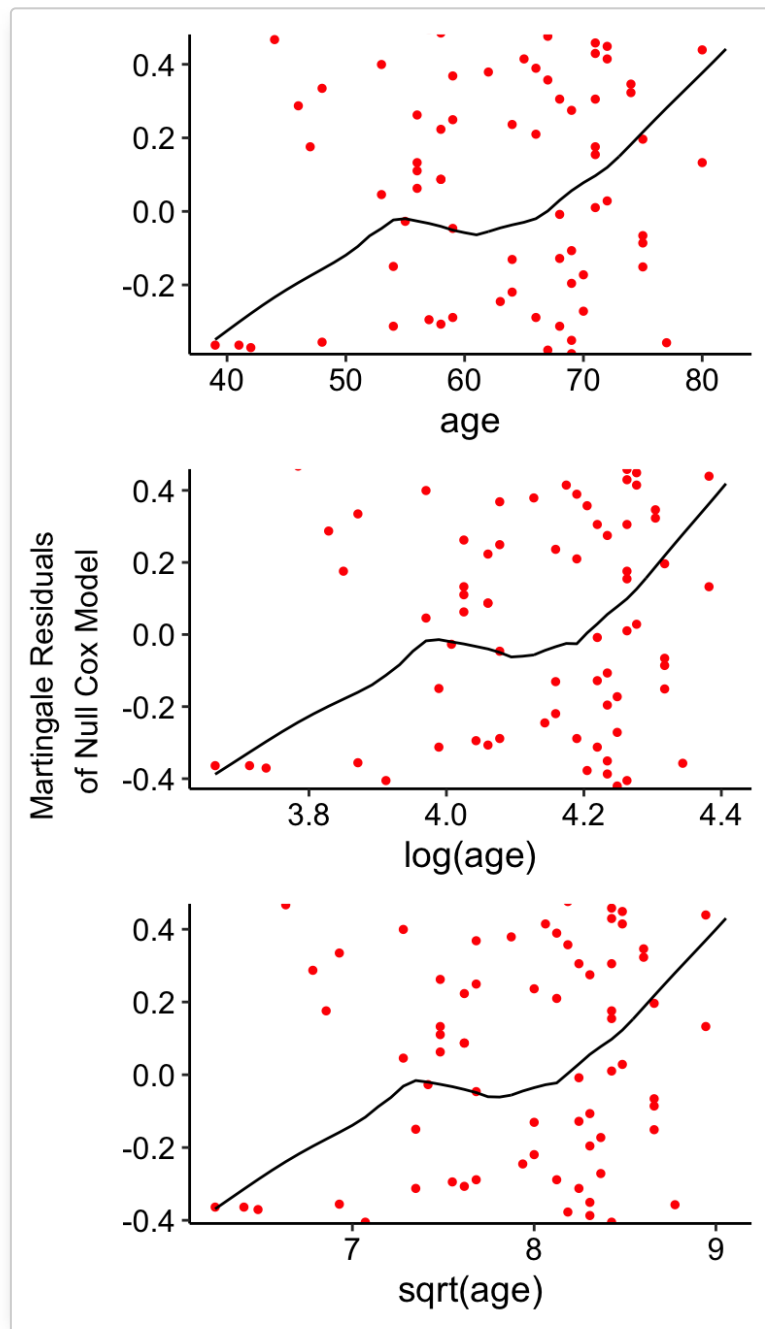
- A value of martingale residuals near 1 represents individuals that “died too soon”,
- and large negative values correspond to individuals that “lived too long”.

To assess the functional form of a continuous variable in a Cox proportional hazards model, we'll use the function *ggcoxfunctional()* [in the *survminer* R package].

The function *ggcoxfunctional()* displays graphs of continuous covariates against martingale residuals of null cox proportional hazards model. This might help to properly choose the functional form of continuous variable in the Cox model. Fitted lines with lowess function should be linear to satisfy the Cox proportional hazards model assumptions.

For example, to assess the functional forme of age, type this:

```
ggcoxfunctional(Surv(time, status) ~ age + log(age) + sqrt(age), data = lung)
```



It appears that, nonlinearity is slightly here.