check assumption of cox regression

reference: https://www.coursera.org/learn/survival-analysis-r-public-health/supplement/5wr8w/feedback-on-practice-quiz

Preparation:

```
library(survival)
```

```
## Warning: package 'survival' was built under R version 3.5.3
```

```
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 3.5.3
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
## Warning: package 'ggpubr' was built under R version 3.5.3
```

```
## Loading required package: magrittr
```

```
data_link = 'https://raw.githubusercontent.com/AilingLiu/Survival_analysis/master/Data/simulated%20HF%2

hf =read.csv(data_link)
head(hf)
```

```
##   id death los age gender cancer cabg crt defib dementia diabetes hypertension
## 1  1     0   2  90      2      0    0   0     0        0        0            0
## 2  2     0  10  74      1      0    0   0     0        0        0            1
## 3  3     0   3  83      2      0    0   0     0        0        0            1
## 4  4     0   1  79      1      0    0   0     0        0        1            1
## 5  5     0  17  94      2      0    0   0     0        0        1            1
## 6  6     0  47  89      1      0    0   0     0        0        0            0
##   ihd mental_health arrhythmias copd obesity pvd renal_disease valvular_disease
## 1   0             0           1    0       0   0             0                1
## 2   1             0           0    0       0   0             1                1
## 3   0             0           1    0       0   0             0                0
## 4   1             0           0    1       0   0             0                0
## 5   0             0           0    0       0   0             0                0
## 6   0             0           0    0       0   1             0                1
##   metastatic_cancer pacemaker pneumonia prior_appts_attended prior_dnas pci
## 1                 0         0         0                    4          0   0
## 2                 0         0         1                    9          1   0
## 3                 0         0         0                    1          0   0
## 4                 0         1         0                    9          2   1
## 5                 0         0         0                    3          0   0
## 6                 0         0         1                    3          0   0
##   stroke senile quintile ethnicgroup fu_time
## 1      0      0        2          NA     416
## 2      0      0        4           1     648
## 3      0      0        3           1     466
```

```
## 4      1      0        5           1     441
## 5      0      0        2           1     371
## 6      0      0        3          NA      47
```

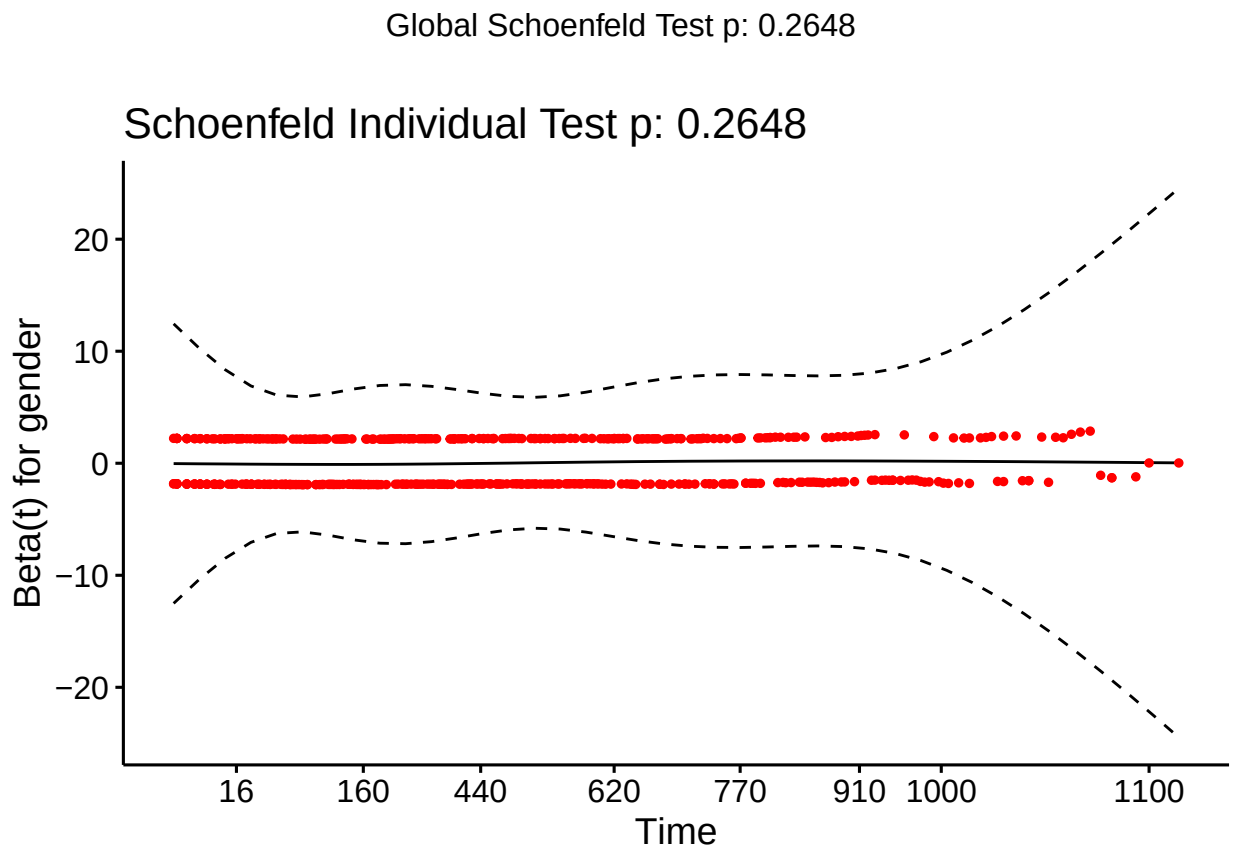
Build a cox regression model with gender as predictor

```
fit <- coxph(Surv(fu_time, death) ~ gender, data=hf) # fit the desired model

temp <- cox.zph(fit)# apply the cox.zph function to the desired model

print(temp, digits = 4) # display the results
```

```
##        chisq df     p
## gender 1.244  1 0.265
## GLOBAL 1.244  1 0.265
```

The pvalue is higher than the conventional 0.05, meaning there is no evidence that time and residuals are related.

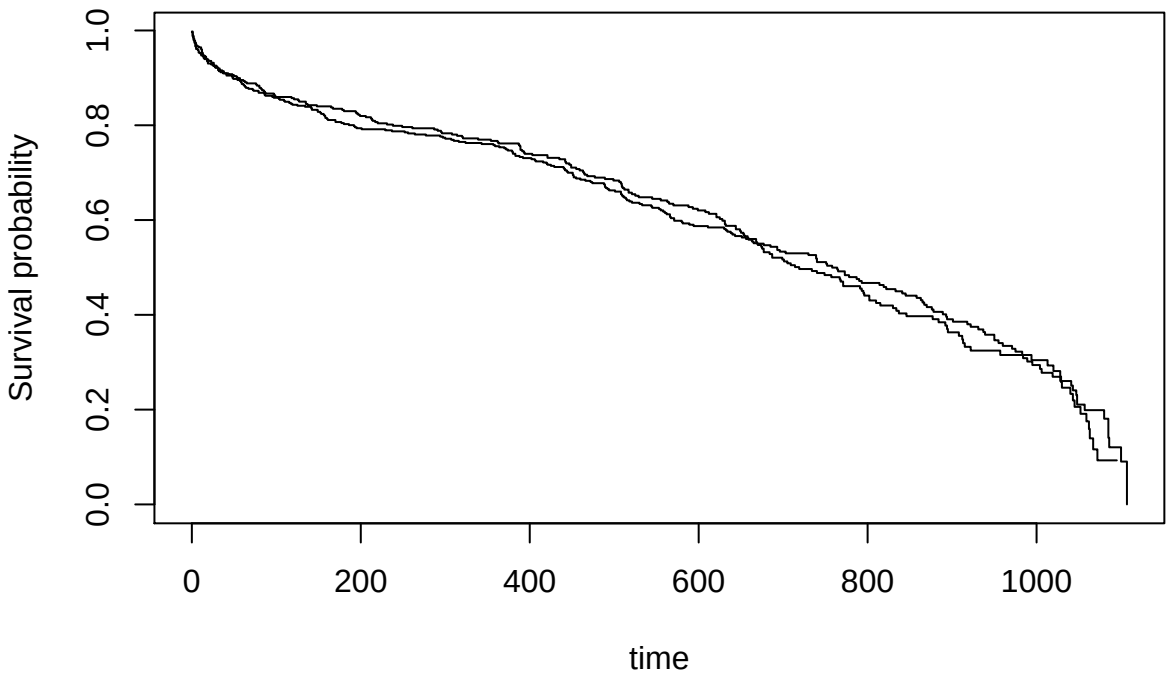
```
ggcoxzph(temp) # plot the curves
```



Technically speaking, the function cox.zph() correlates for each predictor the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. From this plot, it proves the statistics above for no relation between time and residuals.

#KM plot for gender

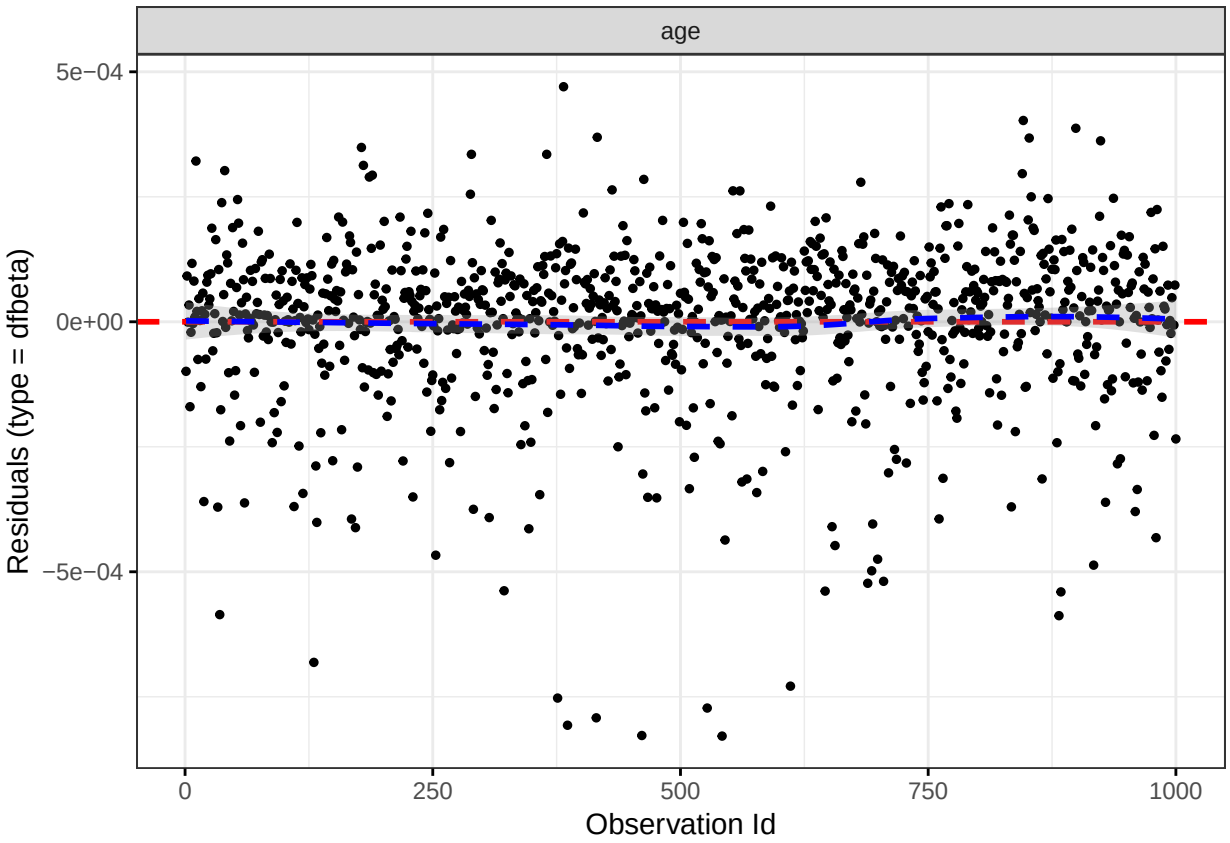
```
km_fit <- survfit(Surv(fu_time, death) ~ gender, data=hf)

plot(km_fit, xlab = "time", ylab = "Survival probability") # label the axes
```



The lines give the survival probability for each gender at each time point. This plot is very no-frills, but in this instance it does the job - it's pretty clear that the two lines are pretty parallel over time throughout the follow-up period, though in this instance the fact that they're almost on top of one another makes it easy to judge.

Use martingale residuals against covariate to look for outliers

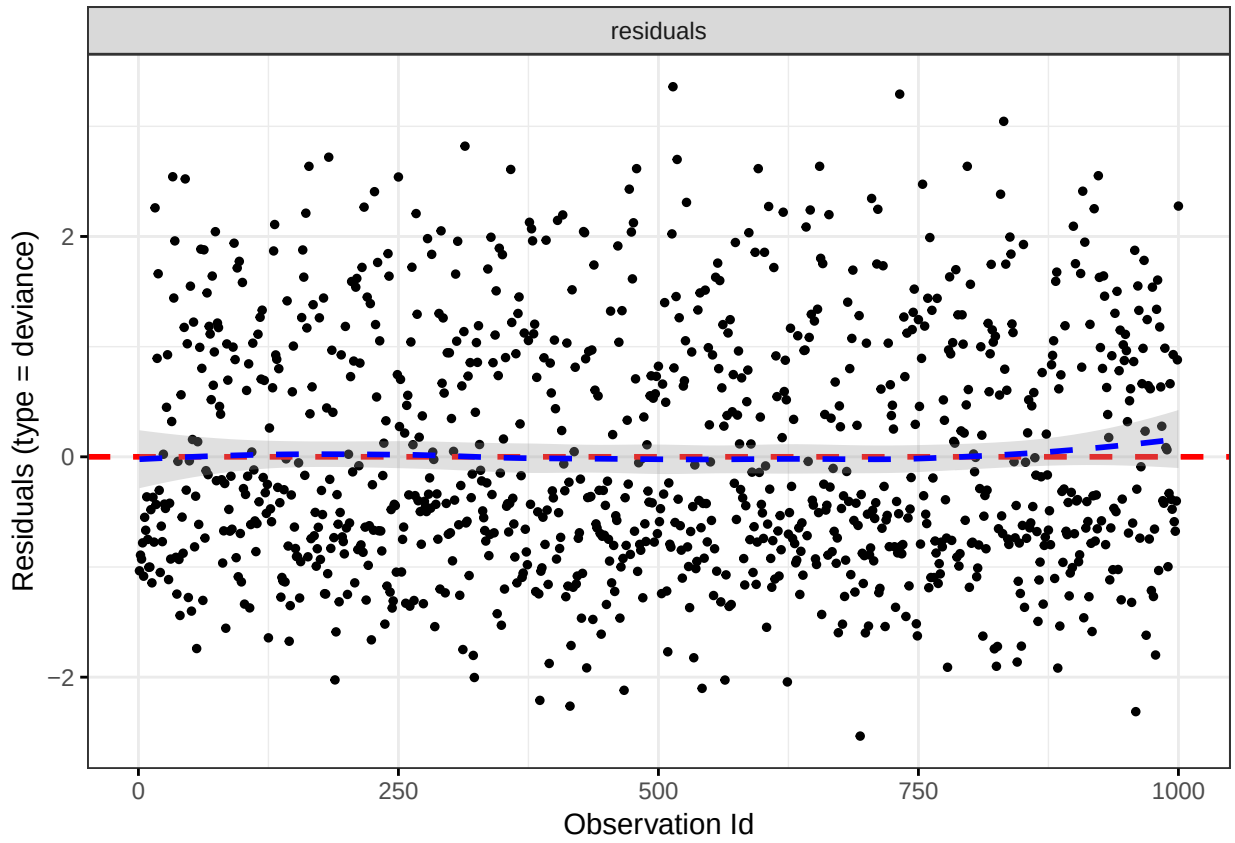
```
res.cox <- coxph(Surv(fu_time, death) ~ age, data=hf)
ggcoxdiagnostics(res.cox, type = "dfbeta",
                 linear.predictions = FALSE, ggtheme = theme_bw())
```



The residuals are sitting around 0

Check outliers with standardized martingale residuals, for data points outside of [-1.96, 1.96], they are potentional outliers (5% of normal distribution)

```
res.cox <- coxph(Surv(fu_time, death) ~ age, data=hf)
ggcoxdiagnostics(res.cox, type = "deviance",
                 linear.predictions = FALSE, ggtheme = theme_bw())
```

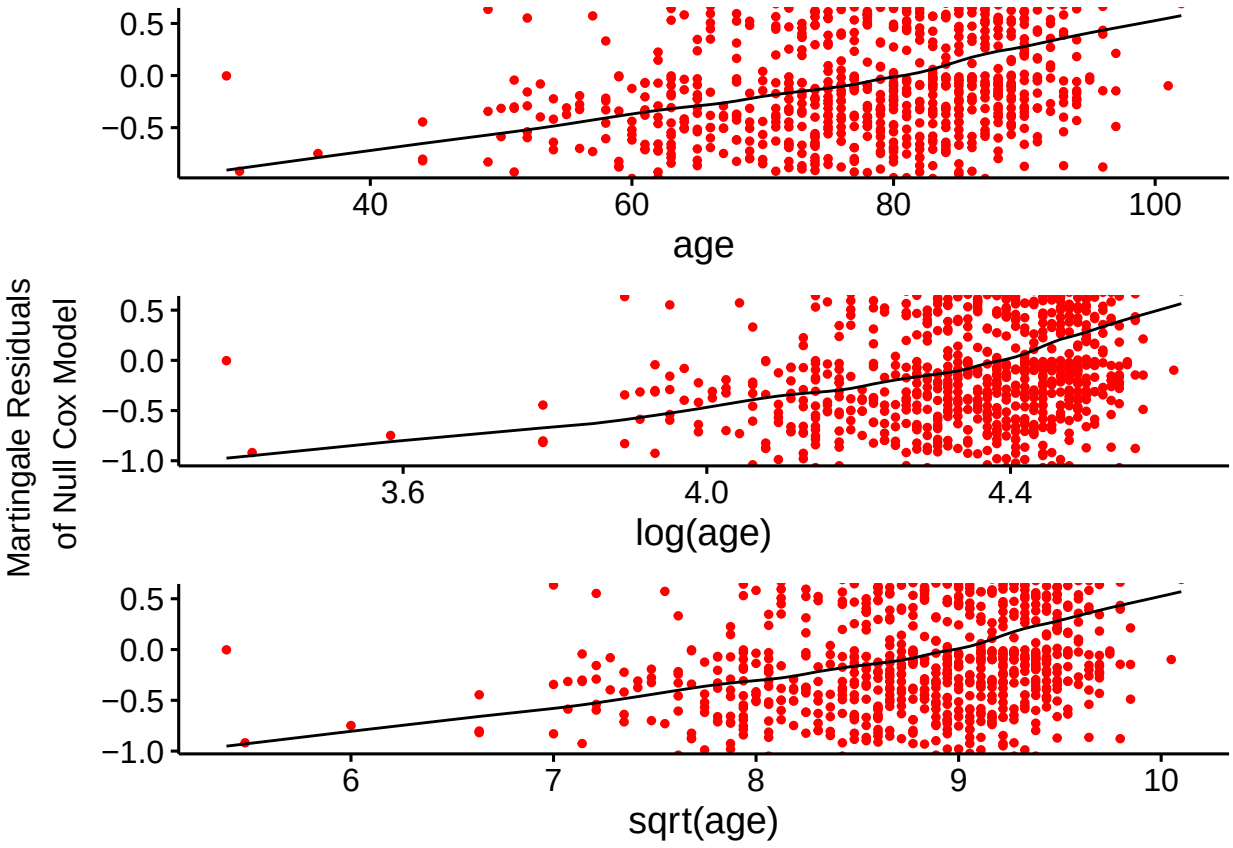


The deviance residuals, which are normalized transformations of the martingale residual and should be roughly symmetrically distributed about zero with a standard deviation of 1. In normal distribution, 5% of observations are more than 1.96 standard deviations from the mean. So if the SD is 1, then only 5% of observations should be bigger than 1.96 or more negative than -1.96. If we observe more than that proportion, then the model doesn't fit the data as well as it should and some observations are a problem.

check linearity between predictor and outcome

```
ggcoxfunctional(Surv(fu_time, death) ~ age + log(age) + sqrt(age), data=hf)
```

```
## Warning: arguments formula is deprecated; will be removed in the next version;
## please use fit instead.
```



ggcoxfunctional() is part of the survminer R package. Martingale residuals may present any value between minus infinity and 1) and have a mean of zero:

Martingale residuals near 1 represent individuals that “died too soon” Large negative values correspond to individuals that “lived too long” The plots should give you nice straight line if the assumption is valid.