GR5291 - HW9

Mathieu Sauterey - UNI: mjs2364

November 8, 2018

Consider the Mayo Clinic Lung Cancer Data including the variables inst (institution code), time (survival time in days), status (censoring status 1=censored, 2=dead), age (age in years) and sex (male=1 and female=2)

Define 'AGE GROUP' as 'YOUNG" if 'age < 65", and OLD, otherwise.

```
# Loads the survival package
library(survival)
```

```
## Warning: package 'survival' was built under R version 3.4.4
```

```
# Loads the Mayo Clinic Lung Cancer Data
data(cancer)

#Defines AGE GROUP
cancer$age_group <- ifelse(cancer$age < 65, "YOUNG", "OLD")

# Prints the first rows of the dataset
head(cancer[,c(2,3,4,11)])
```

```
##   time status age age_group
## 1  306      2  74       OLD
## 2  455      2  68       OLD
## 3 1010      1  56     YOUNG
## 4  210      2  57     YOUNG
## 5  883      2  60     YOUNG
## 6 1022      1  74       OLD
```

1. Using a Cox proportional hazards model, estimate the hazard rate for old relative to young.

```
# Fits the Cox proportional hazards model
fitcox <- coxph(formula = Surv(time, status) ~ age_group, data = cancer)

summary(fitcox)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age_group, data = cancer)
##
##   n= 228, number of events= 165
##
##                   coef exp(coef) se(coef)     z Pr(>|z|)
## age_groupYOUNG -0.2985    0.7419   0.1562 -1.91   0.0561 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## age_groupYOUNG    0.7419      1.348    0.5462     1.008
##
## Concordance= 0.538  (se = 0.022 )
## Rsquare= 0.016   (max possible= 0.999 )
## Likelihood ratio test= 3.62  on 1 df,   p=0.06
## Wald test            = 3.65  on 1 df,   p=0.06
## Score (logrank) test = 3.68  on 1 df,   p=0.06
```

Hazard of rate for young is 1-0.7419 = 26% smaller relative to old. However it is not statistically significant.

2. Assess the validity of the proportional hazards assumption in (1).

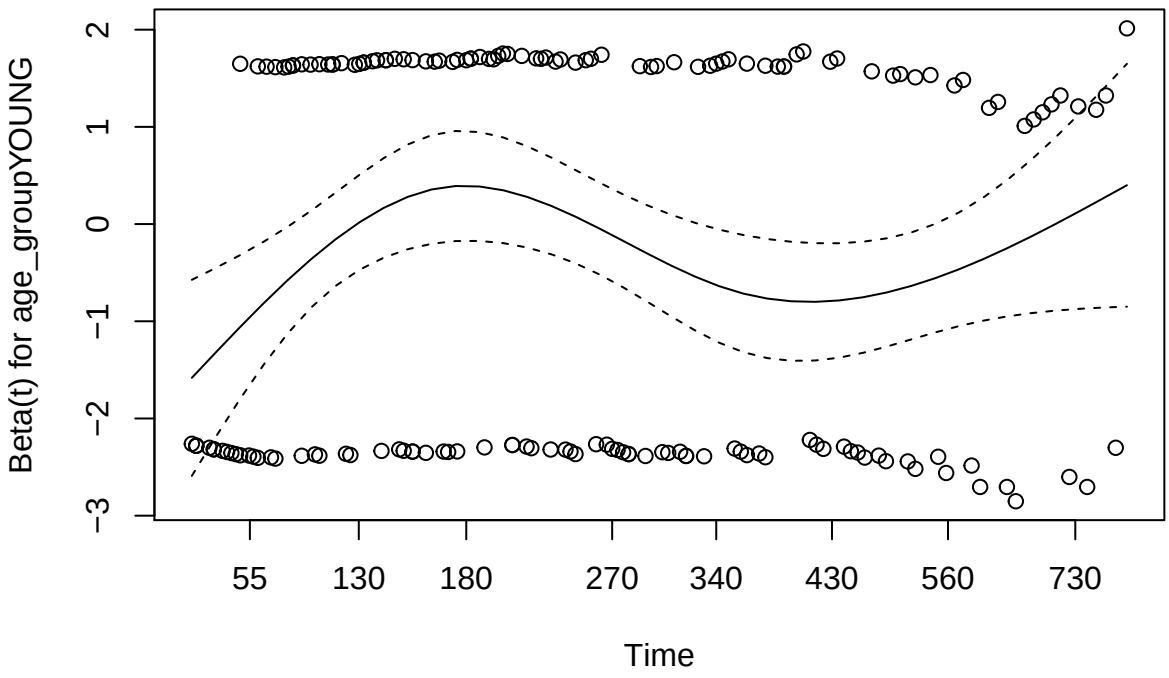
We must check the 3 assumptions of the model:

- Testing the proportional hazards assumption (ratio of hazards does not depend on t) using Schoenfeld residuals.
- Examining influential observations (or outliers) using dfbetas.
- Detecting nonlinearity in relationship between the log hazard and the covariates using martingale residuals. BUT nonlinearity is not an issue for categorical variables like age_group or sex.

```
# Proportional hazards assumption
test.ph <- cox.zph(fitcox)
test.ph
```

```
##                   rho  chisq     p
## age_groupYOUNG 0.0211 0.0727 0.787
```

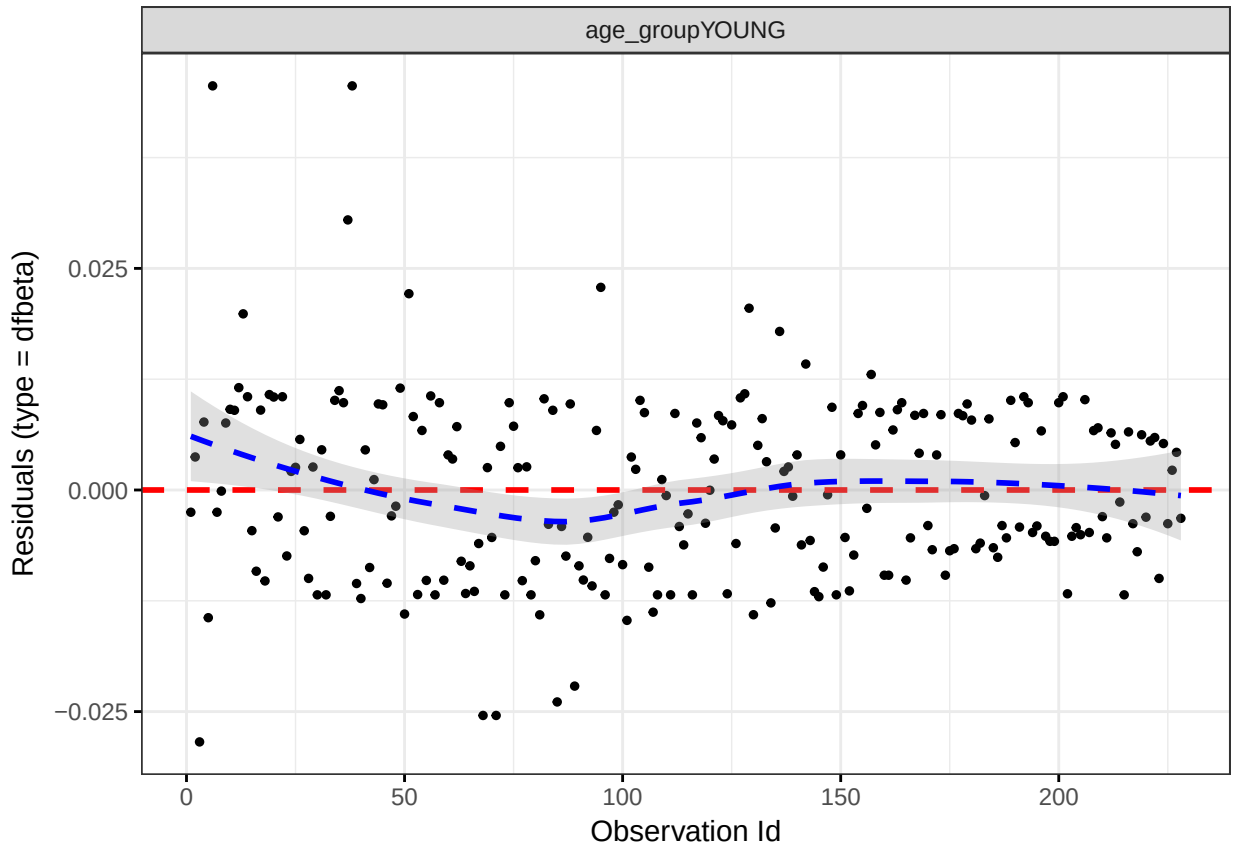
```
plot(test.ph)
```




```
# Outliers
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 3.4.4
## Loading required package: ggplot2
## Loading required package: ggpubr
## Warning: package 'ggpubr' was built under R version 3.4.3
## Loading required package: magrittr
```

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



-Based on the first ouput above, the test is not statistically significant for the age_group covariate BUT the beta line versus time is clearly oscillating over time. Therefore, we can NOT assume the proportional hazards.
- The second graph of the dfbeta residuals shows that there is not really any influential observations.

3. Repeat 1, adjusting for "Sex".

```
# Fits the Cox proportional hazards model
fitcox <- coxph(formula = Surv(time, status) ~ sex, data = cancer)

summary(fitcox)
```

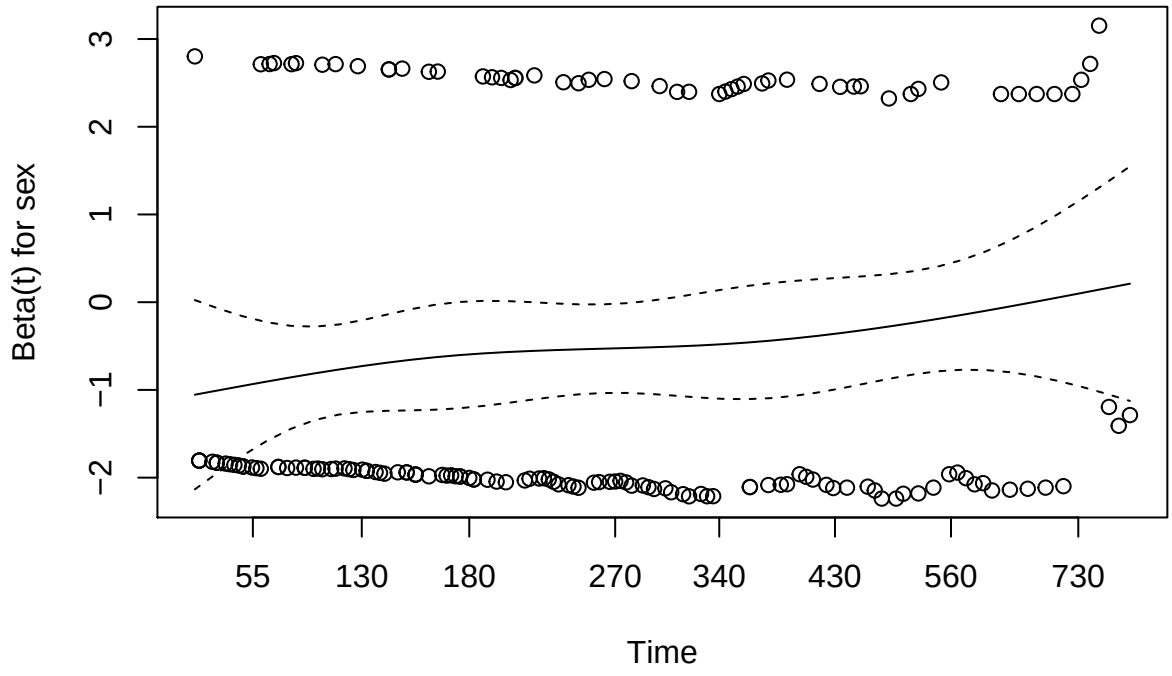
```
## Call:
## coxph(formula = Surv(time, status) ~ sex, data = cancer)
##
##   n= 228, number of events= 165
##
##        coef exp(coef) se(coef)      z Pr(>|z|)
## sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## sex     0.588      1.701    0.4237     0.816
##
## Concordance= 0.579  (se = 0.022 )
## Rsquare= 0.046   (max possible= 0.999 )
```

```
## Likelihood ratio test= 10.63  on 1 df,   p=0.001
## Wald test            = 10.09  on 1 df,   p=0.001
## Score (logrank) test = 10.33  on 1 df,   p=0.001
```

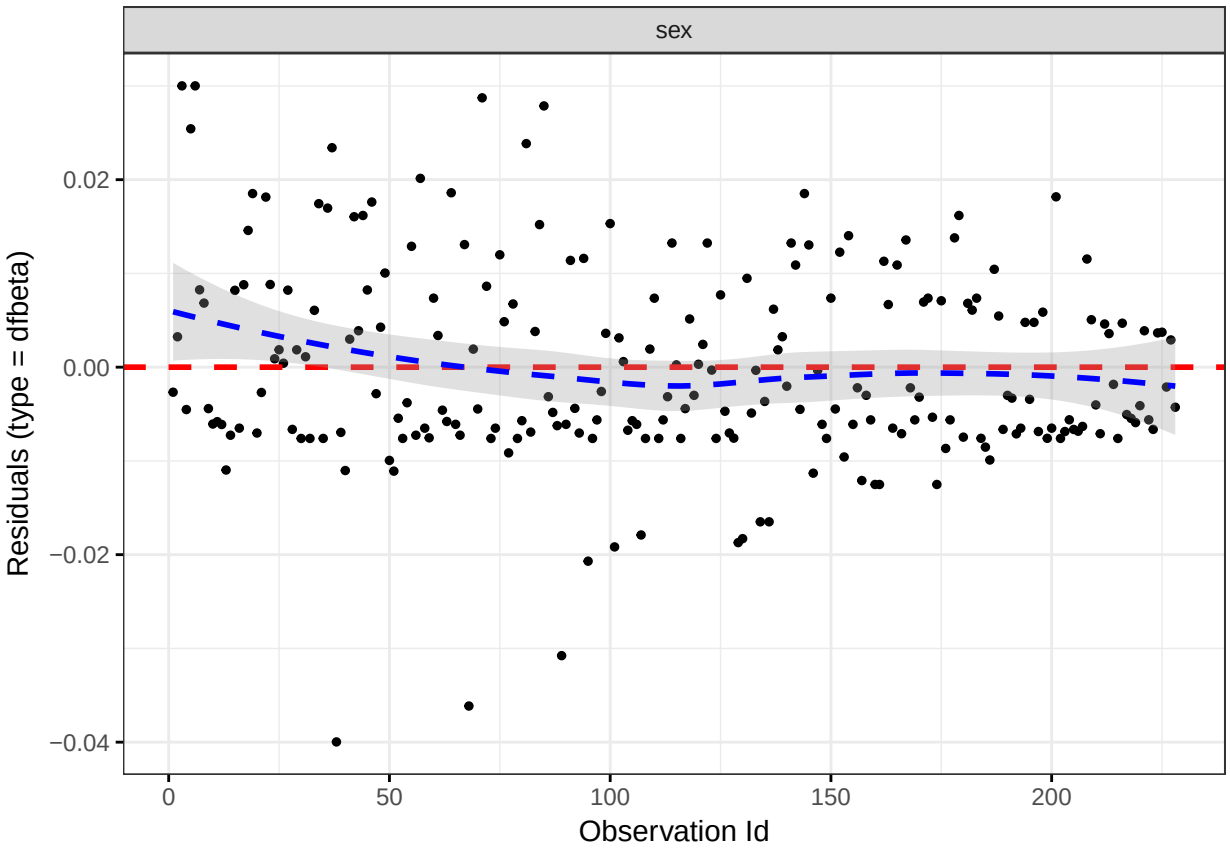
```
# Proportional hazards assumption
test.ph <- cox.zph(fitcox)
test.ph
```

```
##       rho chisq      p
## sex 0.131  2.77 0.0962
```

```
plot(test.ph)
```




```
# Outliers
library(survminer)
ggcoxdiagnostics(fitcox, type = "dfbeta", linear.predictions = FALSE, ggtheme = theme_bw())
```



Hazard of rate for females is 1-0.588 = 41% smaller relative to males, and it is statistically significant.

-Based on the first ouput above, the test is not statistically significant for the sex covariate (p-value of 0.0962). Additionally, the beta line versus time is approximately horizontal. Therefore, we can assume the proportional hazards. - The second graph of the dfbeta residuals shows that there is not really any influential observations.