PhD PROGRAMME IN HEALTH DATA SCIENCE

# Cox Regression

*STATS – Modelação Estatística*

*PhD Programme in Health Data Science*

**Cristina Costa Santos & Andreia Teixeira**

FACULDADE DE MEDICINA UNIVERSIDADE DO PORTO

MEDCIDS DEPARTAMENTO DE MEDICINA DA COMUNIDADE, INFORMAÇÃO E DECISÃO EM SAÚDE

CINTESIS Health. Research.

# Cox regression

- Models time-to-event data in the presence of censored cases.

- Allows the inclusion of predictor variables (covariates). These can be categorical or continuous.

- Also known as Cox Proportional Hazards model or Cox model.

# Cox regression

$$h(t, X) = h_0(t) e^{\sum_{i=1}^{p} \beta_i X_i}$$

where X = ($X_1$, $X_2$, …, $X_p$) are explanatory/predictor variables.

- $h_0(t)$ is called the **baseline hazard** function.
- Proportional hazards (PH) assumption:
  - the baseline hazard is a function of t, but does not involve X's
  - the exponential expression involves the X's, but does not involve t
  - the X's are called **time-independent** predictors (variable whose value for a given individual does not change over time)
  - It is possible to consider time-dependent variables and then an **extended Cox model** needs to be considered
- As $h_0(t)$ is an unspeficied function, Cox model is a **semiparametric** model

# Hazard rates & ratios

- The **hazard rate** is the probability that if the event in question has not already occurred, it will occur in the next time interval, divided by the length of that interval. This time interval is made very short, so that in effect the hazard rate represents an instantaneous rate.

- **Hazard ratio** (HR) is defined as the hazard for one individual divided by the hazard for a different individual. The two individuals being compared can be distinguished by their values for the set of predictors, X's

$$\widehat{HR} = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)}$$

where $X^* = (X_1^*, X_2^*, \ldots, X_p^*)$ and $X = (X_1, X_2, \ldots, X_p)$ denote the set of X's for two individuals.

# Hazard ratios

HEADS
PhD PROGRAMME IN HEALTH DATA SCIENCE

$$\widehat{HR} = \frac{\hat{h}(t,\mathbf{X}^*)}{\hat{h}(t,\mathbf{X})} = \frac{\cancel{\hat{h}_0(t)}\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i^*}}{\cancel{\hat{h}_0(t)}\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i}}$$

$$\widehat{HR} = \frac{\hat{h}_0(t)\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i^*}}{\hat{h}_0(t)\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i}} = e^{\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)}$$

$$\boxed{\widehat{HR} = \exp\left[\sum_{i=1}^{p} \beta_i (X_i^* - X_i)\right]}$$

# Interpretation of HR

- HR = 1 $\Rightarrow$ **no relationship**
- HR > 1 $\Rightarrow$ “exposed” with **higher hazard** comparing with “unexposed”
- HR < 1 $\Rightarrow$ “exposed” with **lower hazard** comparing with “unexposed”

# Cox regression Assumptions

- **Assumption of Proportional Hazards (PH):**
  The hazards are consistent and do not vary differently over time.

Can **examine the residuals (Schoenfeld residuals)**: If PH is true then the plot of the residuals should be horizontal and close to 0.

- ✓ Should not show a clear trend over time (*i.e.* not drastically increasing or decreasing).
- ✓ It should also be centered close to 0.

# Example dataset

- Time to event data for two groups: **Group – categorical variable** (*Group A* and *Group B*) coded 1 and 2, respectively.
- Time in months until event or until end of follow-up: **Time – continuous variable.**
- Whether the individual has had the event of interest: **Status – categorical variable** (*No event* and *event*) coded 0 and 1, respectively.
- The age of the individual at the start of the study: **Age – continuous variable.**

# Example dataset

| Group | Time | Status | Age |
|---|---|---|---|
| A | 9 | Event | 65 |
| A | 12 | No event | 61 |
| A | 14 | Event | 57 |
| A | 14 | Event | 55 |
| A | 16 | No event | 50 |
| A | 18 | Event | 52 |
| A | 24 | Event | 51 |
| A | 30 | No event | 50 |

| Group | Time | Status | Age |
|---|---|---|---|
| B | 3 | Event | 70 |
| B | 7 | Event | 64 |
| B | 9 | No event | 64 |
| B | 11 | Event | 61 |
| B | 12 | Event | 53 |
| B | 15 | Event | 51 |
| B | 19 | Event | 50 |
| B | 21 | Event | 48 |

# Cox regression

## Unadjusted Cox regression

```
> cox1<-coxph(Surv(data$Time,data$Event)~ 1+as.factor(data$Group))
> summary(cox1)
Call:
coxph(formula = Surv(data$Time, data$Event) ~ 1 + as.factor(data$Group))

  n= 16, number of events= 12

                         coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(data$Group)2 0.9224    2.5154   0.6307 1.462    0.144

                       exp(coef) exp(-coef) lower .95 upper .95
as.factor(data$Group)2     2.515     0.3976    0.7307      8.66

Concordance= 0.629  (se = 0.08 )
Likelihood ratio test= 2.25  on 1 df,   p=0.1
Wald test            = 2.14  on 1 df,   p=0.1
Score (logrank) test = 2.29  on 1 df,   p=0.1
```

# Cox regression

## Unadjusted Cox regression

```
> cox2<-coxph(Surv(data$Time,data$Event)~ 1+data$Age)
> summary(cox2)
Call:
coxph(formula = Surv(data$Time, data$Event) ~ 1 + data$Age)

  n= 16, number of events= 12

           coef exp(coef) se(coef)     z Pr(>|z|)
data$Age 0.4229    1.5264   0.1388 3.047  0.00231 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
data$Age     1.526     0.6551     1.163     2.004

Concordance= 0.892  (se = 0.034 )
Likelihood ratio test= 17.57  on 1 df,   p=3e-05
Wald test            = 9.28  on 1 df,   p=0.002
Score (logrank) test = 17.03  on 1 df,   p=4e-05
```

# Cox regression

Adjusted Cox regression, including all covariates

```
> cox3<-coxph(Surv(data$Time,data$Event)~ 1+as.factor(data$Group)+data$Age)
> summary(cox3)
Call:
coxph(formula = Surv(data$Time, data$Event) ~ 1 + as.factor(data$Group) +
    data$Age)

  n= 16, number of events= 12

                         coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(data$Group)2 2.1612    8.6812   0.9583 2.255  0.02412 *
data$Age               0.5999    1.8220   0.2002 2.997  0.00273 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                       exp(coef) exp(-coef) lower .95 upper .95
as.factor(data$Group)2     8.681     0.1152     1.327    56.793
data$Age                   1.822     0.5489     1.231     2.697

Concordance= 0.952  (se = 0.03 )
Likelihood ratio test= 24.32  on 2 df,   p=5e-06
Wald test            = 9.42  on 2 df,   p=0.009
Score (logrank) test = 21.8  on 2 df,   p=2e-05
```

# Cox regression

## Adjusted Cox regression, including all covariates

```
> cox3<-coxph(Surv(data$Time,data$Event)~ 1+as.factor(data$Group)+data$Age)
> summary(cox3)
Call:
coxph(formula = Surv(data$Time, data$Event) ~ 1 + as.factor(data$Group) +
    data$Age)

  n= 16, number of events= 12

                         coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(data$Group)2 2.1612    8.6812   0.9583 2.255  0.02412 *
data$Age               0.5999    1.8220   0.2002 2.997  0.00273 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                       exp(coef) exp(-coef) lower .95 upper .95
as.factor(data$Group)2     8.681     0.1152     1.327    56.793
data$Age                   1.822     0.5489     1.231     2.697

Concordance= 0.952  (se = 0.03 )
Likelihood ratio test= 24.32  on 2 df,   p=5e-06
Wald test            = 9.42  on 2 df,   p=0.009
Score (logrank) test = 21.8  on 2 df,   p=2e-05
```

Hazard ratio (1.822) for each unit increase in age with CI and p-value (p = 0.003).

# Cox regression

## Adjusted Cox regression, including all covariates

```
> cox3<-coxph(Surv(data$Time,data$Event)~ 1+as.factor(data$Group)+data$Age)
> summary(cox3)
Call:
coxph(formula = Surv(data$Time, data$Event) ~ 1 + as.factor(data$Group) +
    data$Age)

  n= 16, number of events= 12

                         coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(data$Group)2 2.1612    8.6812   0.9583 2.255  0.02412 *
data$Age               0.5999    1.8220   0.2002 2.997  0.00273 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                       exp(coef) exp(-coef) lower .95 upper .95
as.factor(data$Group)2     8.681     0.1152     1.327    56.793
data$Age                   1.822     0.5489     1.231     2.697

Concordance= 0.952  (se = 0.03 )
Likelihood ratio test= 24.32  on 2 df,   p=5e-06
Wald test            = 9.42  on 2 df,   p=0.009
Score (logrank) test = 21.8  on 2 df,   p=2e-05
```

Hazard ratio (8.681) for being in Group B, relative to Group A (reference) with CI and p-value (p = 0.024).

# Cox regression

| | Hazard ratio (95% CI) | p-value |
|---|---|---|
| Age | 1.822 (1.231, 2.697) | 0.003 |
| Group B | 8.681 (1.327, 56.793) | 0.024 |

Here you can see that the hazard is 82% higher for each additional year of age and this effect is highly significant (p = 0.003).
Having adjusted for age however there appears to be a very clear difference between the groups with a hazard ratio for Group B relative to Group A of 8.681 (95% CI: 1.327 to 56.793; p = 0.024).
Notice that this confidence interval is very wide and that the lower limit suggests that the true hazard ratio may be as low as 1.327.

# Cox regression

| | Hazard ratio (95% CI) | p-value |
|---|---|---|
| Group B | 2.515 (0.731, 8.66) | 0.144 |

If we take Age out of the model then the effect of the groups is reduced with Group B having an increased hazard ratio relative to Group A of 2.515 (95% CI: 0.731 to 8.66; p = 0.144), which is now not statistically significant at the 5% level.

**Model selection for Survival models** is as importante as it is for other modelling procedures and **needs to be thought about carefully**.

# Cox regression
# PH Assumption

```
> test.res <- cox.zph(cox3)
> test.res
                          chisq df    p
as.factor(data$Group) 0.6809  1 0.41
data$Age              0.0437  1 0.83
GLOBAL                1.2163  2 0.54
```

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.**

# Cox regression
# PH Assumption

```
> test.res <- cox.zph(cox3)
> test.res
                        chisq df    p
as.factor(data$Group) 0.6809  1 0.41
data$Age              0.0437  1 0.83
GLOBAL                1.2163  2 0.54
```

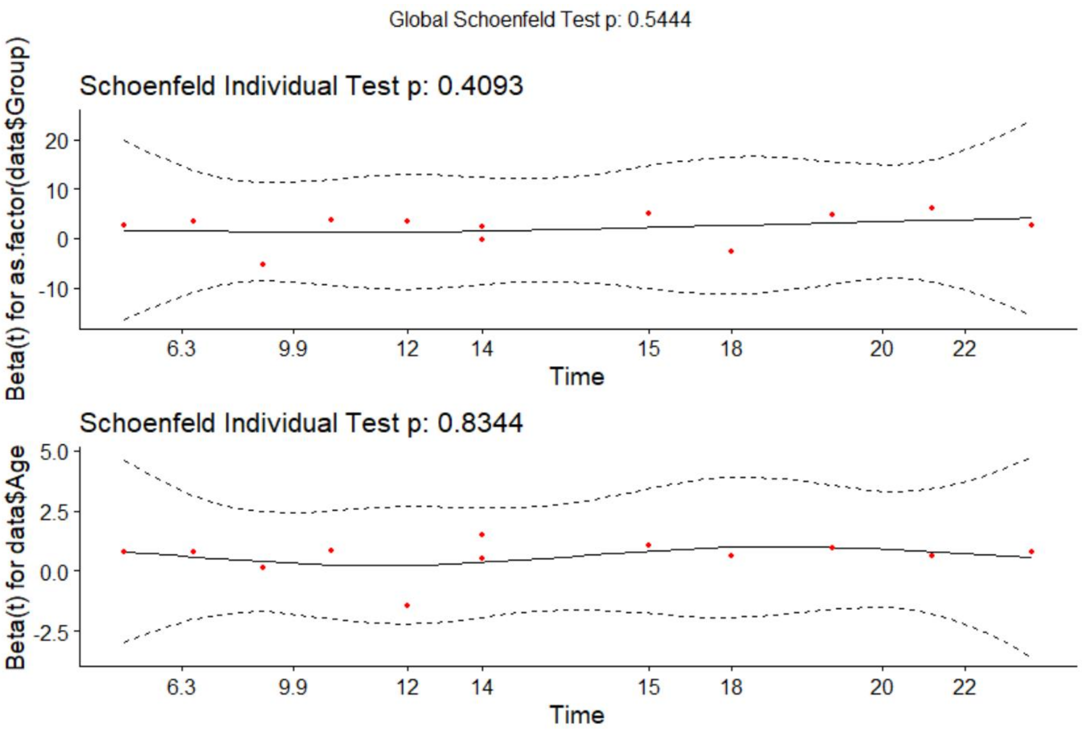
The test is not statistically significant for each of the covariates, and the global test is also not statistically significant. Therefore, we can assume the proportional hazards.

# Cox regression
# PH Assumption

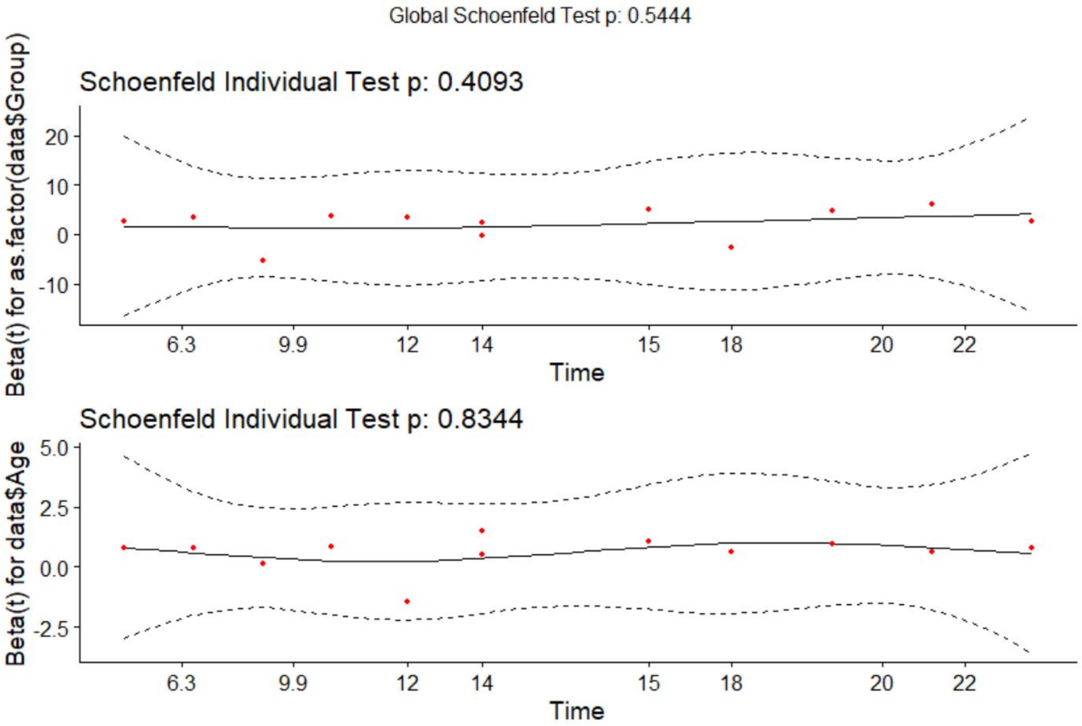
```
> ggcoxzph(test.res)
```

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.

# Cox regression
# PH Assumption



# Cox regression
# PH Assumption




These plots don´t seem to indicate any obvious trend and are generally centered close to zero, but we are dealing with a very small example dataset here.