Survival Analysis: Application in R

```
library(survival)
library(survminer)
library(ggplot2)
library(KMsurv)
```

The remission times of 42 patients with acute leukemia were reported in a clinical trial undertaken to assess the ability of 6-mercaptopurine (6-MP) to maintain remission (i.e., to remain disease free). Each patient was randomized to receive 6-MP or a placebo. Patients were observed until they had a recurrence of the disease, they were censored, or until the study was terminated after one year. The variables in the dataset are:

`time` - the remission times in weeks;

`status` - 1 if patient had a recurrence of leukemia; 0 if censored

`group` - 1 if patient is in the 6-MP group; 0 if in the placebo group

```
leuk <- read.table("leukemia.dat", header = TRUE)
```

Using the `survival` package, there is a function called `Surv` that creates a survival object, which contains the follow-up time and the event indicator together. This object is used as the outcome.

A **Cox proportional hazards model** can be fitted using the `coxph()` function from the `survival` package.

```
m1 <- coxph(Surv(time, status) ~ group, data = leuk)
summary(m1)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ group, data = leuk)
##
##   n= 42, number of events= 30
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## group -1.5721    0.2076   0.4124 -3.812 0.000138 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## group    0.2076      4.817   0.09251    0.4659
##
## Concordance= 0.69  (se = 0.041 )
## Likelihood ratio test= 16.35  on 1 df,   p=5e-05
## Wald test            = 14.53  on 1 df,   p=1e-04
## Score (logrank) test = 17.25  on 1 df,   p=3e-05
```

The focus of a Cox regression is the relative hazard over and above the baseline associated with the predictors. By exponentiating β_j, we can evaluate the hazard ratio associated with a 1-point increase in the predictor, X_j.

A β_j greater than 0 or equivalently a hazard ratio, e^{β_j}, greater than 1, indicates that as the value of the jth predictor variable increases, the hazard of the event increases and thus the length of survival decreases.

HR = 1: No effect
HR > 1: Increase in hazard
HR < 1: Reduction in the hazard

Results from the Cox model suggest a significant effect of group such that being in the 6-MP group reduces the hazard by a factor of 0.21. Specifically, the 6-MP group has an 80% lower rate of re-occurance than the placebo group, and this effect is statistically significant, $z = -3.81, p < .001$. The 95% confidence interval is 0.093 - 0.466, which does not include 1.0 (again indicating statistical significance at $p < .05$).

The output gives p-values for three alternative tests for overall significance of the set of predictors included in the fitted model: the Likelihood-ratio test, Wald test, and score logrank statistics. These three methods are asymptotically equivalent. For large enough N, they will give similar results. For small N, they may differ somewhat. The Likelihood ratio test has better behavior for small sample sizes, so it is generally preferred. In this simplistic example with a single group predictor, the score test is the same as the log rank test.

Additional functions (e.g., `ggcoxzph()`) for exploring departure from the proportional hazards assumption and for other diagnostics (e.g., `ggcoxdiagnostics()`) are available in the `survminer` package. Several types of residuals have been proposed for survival models including Martingale, deviance, score, and Schoenfeld residuals. Scaled Schoenfeld residuals are used to check the proportional hazards assumption. Deviance residuals are used to find outliers, in this case individuals who lived unusually short or long times, after taking into account their observed characteristics.

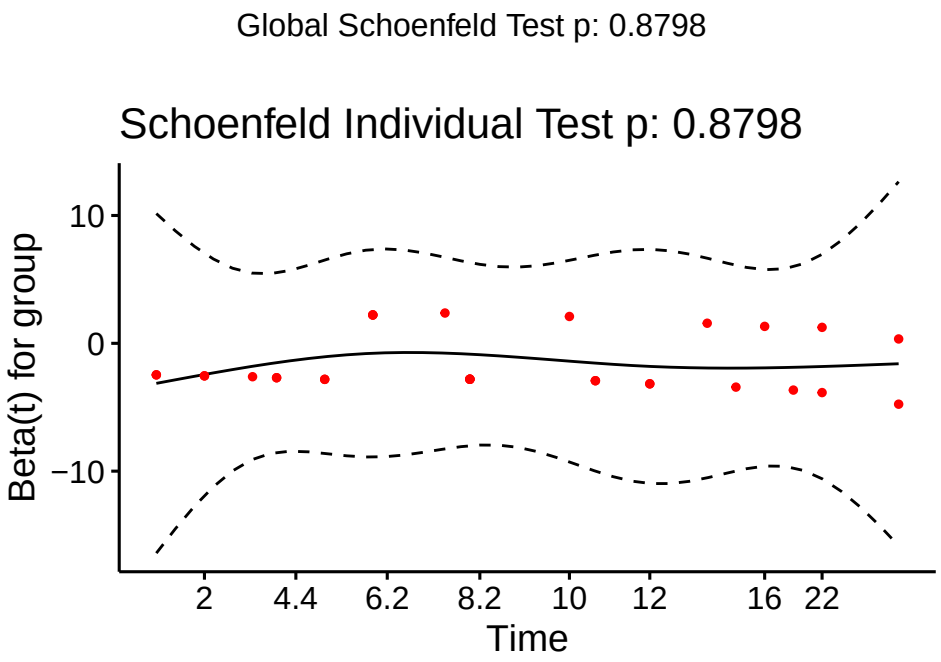
We can test whether the data are sufficiently consistent with the proportional hazards assumption with respect to each of the variables separately as well as globally, using the `cox.zph()` function. For each covariate, the function `cox.zph()` correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole.

```
cox.zph.m1 <- cox.zph(m1)
cox.zph.m1
```

```
##         chisq df    p
## group  0.0229  1 0.88
## GLOBAL 0.0229  1 0.88
```

The proportional hazard assumption is supported by a non-significant relationship between residuals and time, and refuted by a significant relationship. Thus, in this case, the proportional hazard assumption is supported. The scaled Schoenfeld residuals are independent of time. A plot that shows a random pattern against time is evidence that the proportional hazards assumption holds.

```
ggcoxzph(cox.zph.m1)
```



A plot that shows a non-random pattern against time is evidence that the proportional hazards assumption is violated. Violations of proportional hazards assumption can be resolved by:

- Adding a covariate*time interaction
- Stratification, which is useful for "nuisance" confounders, where you do not want to estimate the effect. That is, you cannot examine the effects of the stratification variable.

Example: Data were gathered from 3,470 annual personal interviews conducted for the National Longitudinal Survey of Youth (NLSY, 1995) from 1979 through 1986 to study hospitalized pneumonia. Overall, 73 (2.10%) of the children were reported to be hospitalized for pneumonia within the first year of life. We want to examine the association between the time to hospitalized pneumonia and some child- and/or mother-specific characteristics.

The study included the following potential risk factors:

`mthage` - Age of mother (Years)

`smoke` - Cigarette use by mother during pregnancy (Yes: 44%)

`alcohol` - Alcohol use by mother during pregnancy (Yes: 66%)

`sib` - Presence of siblings of the child (Yes: 48%)

`bweight` - Normal birthweight of child ($\geq$ 5.5 lb, Yes: 36%)

`urban` - Urban environment for mother (Yes: 76%)

`poverty` - Mother at poverty level (Yes: 92%)

`region` - Region of the country (1=northeast, 2=north central, 3=south, 4=west)

The data are contained in the `KMsurv` package and are called `pneumon`. The variable `hospital` indicates whether or not hospitalization, the event, occurred (1 = yes, 0 = no) and `chldage` is the age (in months) that the child had pneumonia.

```
data(pneumon)
pneumon$sib <- (pneumon$nsibs > 0)
cox.pneumon <- coxph(Surv(chldage, hospital) ~ mthage + smoke + sib +
```

```
                      factor(region), data=pneumon)
summary(cox.pneumon)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ mthage + smoke + sib +
##     factor(region), data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## mthage          -0.15295   0.85817  0.04863 -3.145  0.00166 **
## smoke            0.36746   1.44406  0.15313  2.400  0.01641 *
## sibTRUE          0.83730   2.31011  0.25792  3.246  0.00117 **
## factor(region)2  0.06388   1.06597  0.34234  0.187  0.85197
## factor(region)3 -0.35711   0.69970  0.34078 -1.048  0.29468
## factor(region)4 -0.62506   0.53523  0.43537 -1.436  0.15109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## mthage             0.8582     1.1653    0.7802     0.944
## smoke              1.4441     0.6925    1.0696     1.950
## sibTRUE            2.3101     0.4329    1.3935     3.830
## factor(region)2    1.0660     0.9381    0.5449     2.085
## factor(region)3    0.6997     1.4292    0.3588     1.365
## factor(region)4    0.5352     1.8684    0.2280     1.256
##
## Concordance= 0.682  (se = 0.03 )
## Likelihood ratio test= 29.16  on 6 df,   p=6e-05
## Wald test            = 28.92  on 6 df,   p=6e-05
## Score (logrank) test = 29.61  on 6 df,   p=5e-05
```

As before, given two nested models, twice the difference in partial log-likelihoods is distributed as a χ^2 statistic with degrees of freedom equal to the difference in the number of parameters.

For example, to test the significance of region effect, we can fit a reduced model:

```
cox.reduced <- coxph(Surv(chldage, hospital) ~ mthage + smoke + sib,
                     data=pneumon)
summary(cox.reduced)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ mthage + smoke + sib,
##     data = pneumon)
##
##   n= 3470, number of events= 73
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## mthage   -0.15007   0.86065  0.04854 -3.092  0.00199 **
## smoke     0.41259   1.51072  0.15140  2.725  0.00643 **
## sibTRUE   0.83112   2.29589  0.25724  3.231  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
```

```
## mthage     0.8607     1.1619    0.7825    0.9466
## smoke      1.5107     0.6619    1.1228    2.0326
## sibTRUE    2.2959     0.4356    1.3867    3.8012
##
## Concordance= 0.666  (se = 0.029 )
## Likelihood ratio test= 24.62  on 3 df,   p=2e-05
## Wald test            = 24.51  on 3 df,   p=2e-05
## Score (logrank) test = 25  on 3 df,   p=2e-05
```

And compare the log likelihood for the full and the reduced models:

```
anova(cox.pneumon, cox.reduced)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(chldage, hospital)
##  Model 1: ~ mthage + smoke + sib + factor(region)
##  Model 2: ~ mthage + smoke + sib
##    loglik Chisq Df Pr(>|Chi|)
## 1 -572.48
## 2 -574.75 4.538  3     0.2089
```

Thus, we can conclude that `region` does not significantly improve the fit of the model, $\chi^2(3) = 4.54$, $p = 0.209$.

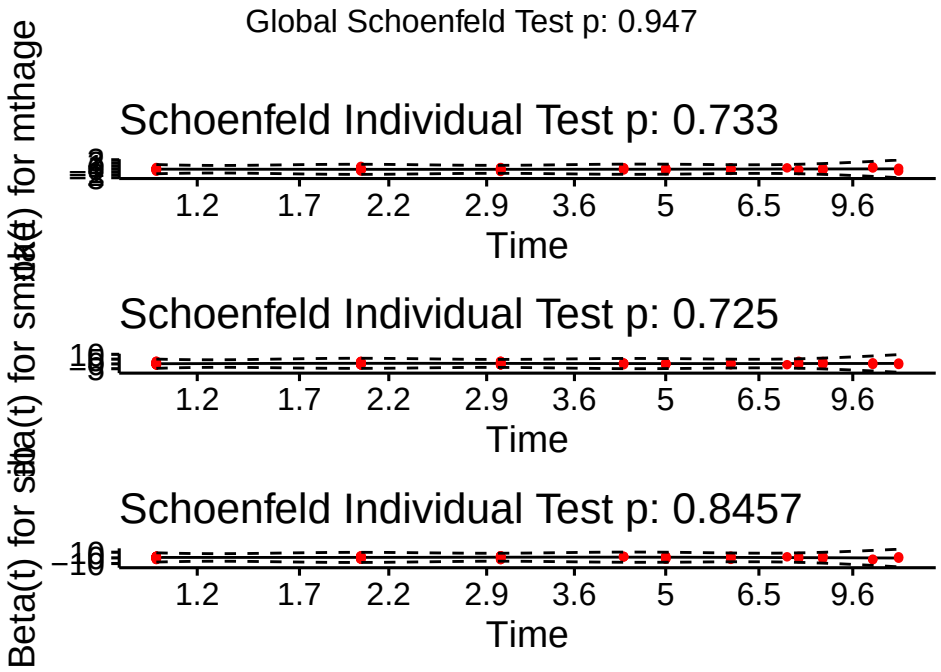
Holding constant smoking and having siblings, a year increase in mother's age reduces the child's hazard of being hospitalized for pneumonia in the first year of life by a factor of .86 (or by 14%; 95% CI: 0.783, 0.947) and this HR is statistically significant, $z = -3.09, p = .002$. Holding constant mother's age and having siblings, smoking increases the child's hazard of being hospitalized for pneumonia in the first year of life by a factor 1.51 (or by 51%; 95% CI: 1.12, 2.03) and the HR is statistically significant, $z = 2.725, p = .006$. Holding constant mom's age and smoking, having siblings also increases the child's hazard of being hospitalized for pneumonia in the first year of life by a factor of 2.30 (95% CI: 1.387, 3.801), $z = 3.231, p = .001$.

Next, we evaluate the proportional hazards assumption:

```
cox.zph.reduced <- cox.zph(cox.reduced)
cox.zph.reduced
```

```
##         chisq df    p
## mthage 0.1164  1 0.73
## smoke  0.1237  1 0.73
## sib    0.0378  1 0.85
## GLOBAL 0.3667  3 0.95
```

```
ggcoxzph(cox.zph.reduced)
```

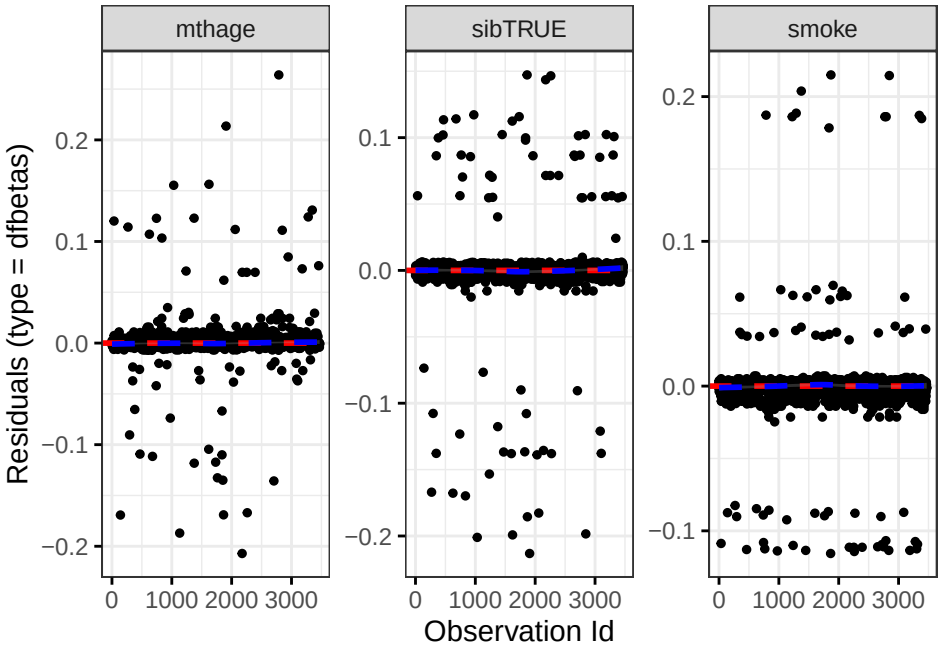


We do not find any evidence that the proportional hazards assumption does not hold.

Next, check for influential observations using DFBETAS.

```
ggcoxdiagnostics(cox.reduced, type = "dfbetas", linear.predictions = FALSE)
```

```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## i Please use `gather()` instead.
## i The deprecated feature was likely used in the survminer package.
##   Please report the issue at <https://github.com/kassambara/survminer/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



It is 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 "experienced the event too soon" compared to expected survival times. Negative values correspond to individuals that "lived too long". Very large or small values are outliers, which are poorly predicted by the model.

Risk scores for an individual can be computed using that individual's predictor values in conjunction with the β estimates to produce a hazard ratio estimate for that person relative to the baseline.