Survival Analysis

The purpose of this TP is to practice using **R** for carrying out survival analysis for two data sets.

As usual, make sure that you read the **help** for any new functions and data that you use.

Glioma data

Load the **glioma** data and read the help for this data set so that you understand a little about the scientific background and question of interest as well as the variables included in the data. In this study, the main interest is to investigate whether patients receiving a novel radioimmunotherapy (RIT) have, on average, longer survival times than patients in the control group.

To look for differences in survival, we examine the Kaplan-Meiere estimates of the survival functions for the treatment and control groups. The Kaplan-Meier estimates for survival functions are computed by the **survfit** function from the **survival** package. To get this, you use a model formula of the form:

Surv(time, event) ~ group

where **time** is the survival times, **event** is a *logical* variable taking the value **TRUE** when the event of interest (*e.g.* death) is observed and **FALSE** in the case of censoring. The variable **group** is a grouping factor.

We are interested in testing differences between treated and control patients separately for Grade III and Grade IV (GBM) glioma patients (glioma is a type of brain tumor). Plot the Kaplan-Meier curves for treated and control patients within each grade (making square plots):

```
    data("glioma", package = "coin")
    library("survival")
    layout(matrix(1:2, ncol = 2))
    par(pty = "s")     # to make square plots

    g3 <- subset(glioma, histology == "Grade3")
    plot(survfit(Surv(time, event) ~ group, data = g3), main="Grade III Glioma", lty = c(2,1), ylab =
"Probability", xlab = "Survival Time (months)", xlim= c(0, max(glioma$time) * 1.05))
legend("bottomleft", c("Control", "Treated"), lty = c(2,1), bty = "n")

    g4 <- subset(glioma, histology == "GBM")
    plot(survfit(Surv(time, event) ~ group, data = g4), main="Grade IV Glioma", lty = c(2,1), ylab =
"Probability", xlab = "Survival Time (months)", xlim= c(0, max(glioma$time) * 1.05))
legend("topright", c("Control", "Treated"), lty = c(2,1), bty = "n")
```

Interpret these plots: does there appear to be a difference in survival between treated and control patients for Grade III tumors? Grade IV?

To test for differences, try using the **log-rank** test separately for each group:

```
    survdiff(Surv(time, event) ~ group, data = g3)
    survdiff(Surv(time, event) ~ group, data = g4)
```

Do these results indicate significant differences between treatment groups?

One problem with using the log-rank test here is that the test is an asymptotic one but the number of patients is rather small. In addition, there are tied results - equal survival times for some patients. In this case, we can use an alternative approach: condition on the observed data and compute the distribution of the test statistics using a permutation test to get an **exact** log-rank test *p*-value, using the **surv_test** in the **coin**:

```
    library("coin")
    logrank_test(Surv(time, event) ~ group, data = g3, distribution = "exact")
    logrank_test(Surv(time, event) ~ group, data = g4, distribution = "exact")
```

Do these results agree with the asymptotic results?

Another question that we can address is whether the new treatment is superior for both groups (Grade III **and** Grade IV tumors) simultaneously. To do this, we **stratify** or **block** by tumor grade. Here, we only approximate the exact conditional distribution, since the exact distribution is hard to compute.

```
    logrank_test(Surv(time, event) ~ group | histology, data = glioma, distribution =
approximate(B=10000))
```

Do these results agree with the asymptotic results? Briefly summarize your overall findings.

Breast cancer data

This study investigated the effects of treatment with Tamoxifen in women with node-positive breast cancer. There are seven prognostic variables measures for each of the 686 women. Here, we are interested in assessing the impact of the covariates on patient survival time. To carry this out, we will do Cox proportional hazards modeling (Cox regression). First though, look at the Kaplan-Meier estimates, stratified by whether or not the patient received hormonal therapy (Tamoxifen) or not.

Load the **GBSG2** data and read the help, then make the Kaplan-Meier curves:

```
    data("GBSG2", package = "TH.data")
    plot(survfit(Surv(time, cens) ~ horTh, data = GBSG2), lty = 1:2, mark.time = FALSE, ylab =
"Probability", xlab = "Survival Time (days)")
legend("bottomleft", legend = c("yes", "no"), lty = c(2,1), title = "Hormonal Therapy", bty = "n")
```

For fitting a Cox model, we use roughly the same rules as for linear models, except we use the `coxph` function and the response variable is a `Surv` object. The `.` operator refer in the formula refers to all other variables that haven't yet been included in the model, thus saving you from typing in all the variables.

Fit a Cox model with all variables and examine the results:

```
    GBSG2.coxph <- coxph(Surv(time, cens) ~ ., data = GBSG2)
    summary(GBSG2.coxph)
    ci <- confint(GBSG2.coxph)
    exp(cbind(coef(GBSG2.coxph), ci))["horThyes",]
```

Which variables are the most important predictors of survival?

Since the relative risk (exponentiated coefficient) for patients undergoing hormone therapy is of greatest interest, estimate this risk and obtain a confidence interval:

```
    ci <- confint(GBSG2.coxph)
    exp(cbind(coef(GBSG2.coxph), ci))["horThyes",]
```

Interpret this result in terms of relative risk of hormone therapy compared to the control group.

In general, model checking and selection for proportional hazards models is complicated. However, one way to check the proportional hazards assumption is by looking at the parameter estimates over time: if these don't vary much over time, then the proportional hazards assumption seems reasonable. To test for constant regression coefficients over time (2-sided test), both globally and for each variable separately, use the `cox.zph` function:

```
    GBSG2.zph <- cox.zph(GBSG2.coxph)
    GBSG2.zph
```

Is there evidence for some variables that the proportional hazards assumption is violated? Which variable(s)?

Let's say that there is a continuous variable in this data set `xx`. Plot the coefficients over time (but use the actual data set variable instead of `xx`):

```
    plot(GBSG2.zph, var = "xx")
```

The solid line in the center represents an average value for the given time point, while the dashed lines are 95% confidence bands. What conclusions do you draw from the numerical and graphical results regarding the possibility of time-varying effects?

The residuals computed in an ordinary linear regression do not apply in the survival context. Instead, martingale residuals can be used to check the model fit. When evaluated at the true coefficient value, the expected martingale residual is zero. We can therefore check for systematic deviations from the assumed model by inspecting scatterplots of the martingale residuals against covariate values. Make some martingale residual plots for different variables, using the following code as a template:

```
    res <- residuals(GBSG2.coxph)
    plot(res ~ age, data = GBSG2, ylim = c(-2.5, 2.5), pch=".", ylab = "Martingale residuals")
    abline(h=0, lty=2)
```

For this variable, we can see in the plot that the martingale residuals are scattered fairly evenly above and below 0, and in addition they do not seem to show any particular pattern. This indicates no systematic deviation from expectation.

Briefly summarize your results and conclusions. Do the women receiving Tamoxifen have longer survival times than women who do not? What evidence is there to justify your answer? Which variable is the most important predictor of survival? Explain.