Johanna Weiss
Arvin Rastegar

Lifetime Data Analysis, Course 2021/22

Exercises Topics 5 and 6

September 21, 2022

Exercise 1

(a) From Figure 1 it can be seen that both of the survival functions drop relatively quickly, with the survival curve of the aneuploid DNA profile having a higher survival value for any value of t. The diploid profile has a median of 42 weeks whereas the aneuploid has a median survival time of 93 weeks. The maximum observed survival time in the diploid profile is 231 weeks, whereas the maximum observed survival time in the aneuploid profile is 400 weeks. The survival function of the aneuploid function reaches a value of 22.89% by the end of the study period and the diploid profile reaches a value of 8.33% at the end of the study.

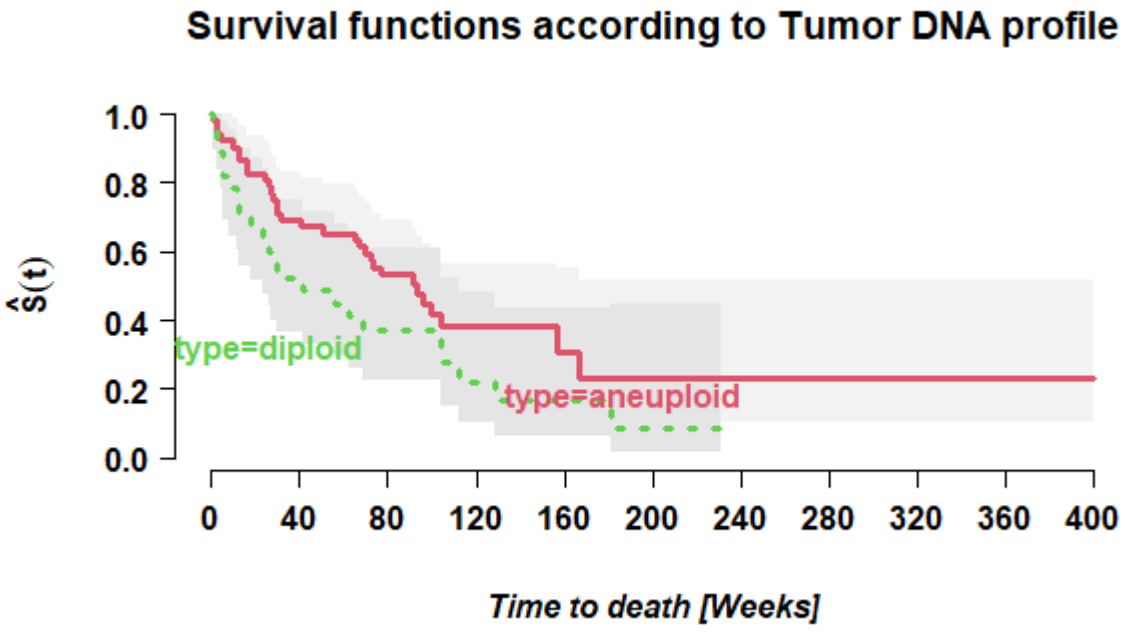

Figure 1: Survival functions with respect to Tumor DNA profile

(b) To assess whether the survival is related to the type of tumor, the logrank test can be used. It is using the following hypotheses using a significance level of 5%:

H_0: There is no observable difference in the two survival functions of the different tumor types, so $S_1(t) = S_2(t)$. This comprises the compound hypotheses $p_{i1} = p_{i2}$ for all $i = 1, ..., D$
H_1:There is a observable difference between the two survival functions, so $S_1(t) != S_2(t)$. This comprises the compound hypotheses $p_{i1} != p_{i2}$ for some $i = 1, ..., D$

Performing the test results in p-value of 0.09 with a Chi-squared of 2.8 and 1 degree of freedom. Therefore, H_O cannot be rejected, so it is not impossible for the survival functions to be the same. With regard to Figure 1, this seems to be a surprising result, since the two survival functions seem to be quite different. However, the test state that one cannot exclude the option that the survival functions are the same.

(c) The data can be fitted to a log-logistic regression model to check the same hypothesis as above. Since the only covariate variable is chosen to be the tumor type, the hypotheses are the following:

H_0: the parameter corresponding to the tumor type in the regression model is 0, so $\lambda = 0$, which means that the different values for tumors do not influence to the model significantly
H_1: the parameter corresponding to the tumor type in the regression model is not 0, so $\lambda! = 0$, which means that the different values for tumors do influence to the model.

Performing the log-logistic test on the fitted model results in a p-value of 0.051 with a Chi-squared of 3.79 and 1 degree of freedom. This means, we cannot reject H_0, so we conclude that the different type of tumors do not have a significant influence on the survival function.

(d) The acceleration factor for the tumor type diploid of the model is 2.2047. The acceleration factor indicates to which degree a covariate accelerates or decelerates the life span. An acceleration factor¿ 1 is an indication that the covariate is reducing the survival time, hence, it can be said that having tumor type diploid reduces the survival time of a patient. In comparison, this means that the median survival time for diploid tumour type is less than the median survival time of tumor type aneuploid.

The odds ratio of tumor type is 2.2794. This indicates that it is approximately 2.3 more likely to die with tumor type diploid compared to aneuploid. This is consistent with the acceleration factor and its conclusion that diploidtumor reduces the survival time.

(e) To assess the goodness of fit of the log-logistic regression, the residuals are looked at. The residuals of the log-logistic model are compared with the theoretical standard normal distribution, as can be seen in Figure 2

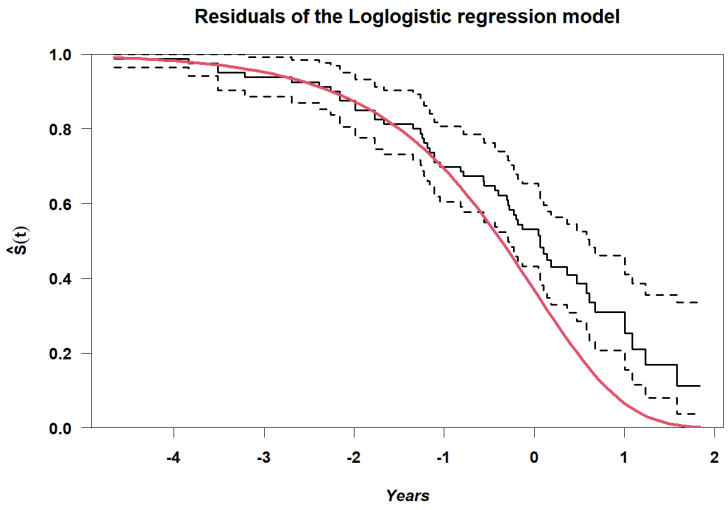

Figure 2: Distribution of the residuals vs. theoretical distribution

in the first half of the residual plot, the data seems to match the theoretical distribution very well. At a value of around -1 the residual graph and the theoretical curve start to behave differently. The theoretical curve is below the residual function, from -0.5 onwards outside of the confidence interval of the residual function. Since the two functions differ strongly in the second half of the plot, one can conclude that the log-logistic regression model does not seem to be a good model to fit the data.

Exercise 2

(a) See appendix

(b) See appendix

(c) The proportion of right-censored survival times is 0.6567.

(d) From plot 4 it can be seen that both models give reasonable probability plots. Since the purpose of the probability plots is to discard models that are clearly not valid, we cannot conclude which model is better. Since none of the model are clearly wrong and they are graphically close, they are both possible models.

Johanna Weiss
Arvin Rastegar

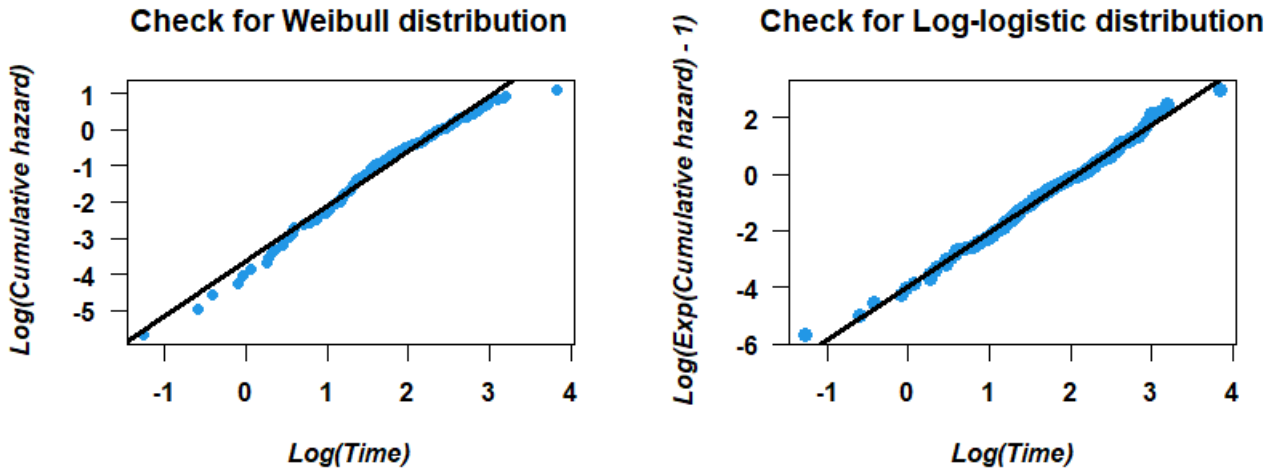

Figure 3: Survival functions with respect to Tumor DNA profile

Exercise 3

In this exercise we wanna show that the following condition holds for cox-proportional hazard model.

$$log(1 - \lambda_j(z)) = log(1 - \lambda_j(0)) \exp(\beta' z) \tag{1}$$

j = 1,...,g.

So when we are talking about the cox-proportional hazard model we know that this theoretical model we have:

$$\hat{\lambda}(t|Z_i) = \hat{\lambda}_0(t)(exp\hat{\beta}Z_i) \tag{2}$$

$$\hat{S}(t|Z_i) = [\hat{S}_0(t)]^{exp\hat{\beta}Z_i} \tag{3}$$

If we change formula 3 and replace the survival function with hazard conditional PDs or hazard rates h(t), $(S = 1 - \lambda(t))$ we can get to this formula:

$$1 - \lambda(t; Z) = [1 - \lambda_0(t)]^{\exp \hat{\beta}Z} \tag{4}$$

Now considering that the grouped the times together creating time intervals of $[0 = a_0, a_1), ..., [a_{g-1}, a_g)$ this makes the hazards into conditional probabilities of $\lambda_j(z) = P(T < a_j | T \geq a_{j-1}; Z)$ $j = 1, ..., g$ therefor for each interval we have :

$$1 - \lambda_j(z) = (1 - \lambda_j(0))^{\exp \hat{\beta}Z} \tag{5}$$

Now we apply the logarithm:

$$\log(1 - \lambda_j(z)) = \log(1 - \lambda_j(0))\exp \hat{\beta} Z \tag{6}$$

Exercise 4

(a) See appendix for r code.

(b) From figure 4, we can see that the survival functions of graft type in the case of allogenic are lower than autologous until near the 19 months which then the survival functions cross and the autologous has a lower survival rate than the allogenic, and the autologous survival function ends around 45 months when we have the last event for this graft type is seen.

Furthermore, in similar behavior, the non-Hodgkin lymphoma has a higher rate than Hodgkins until month 19 which the survival functions cross and it goes lower than the Hodgkins. The Hodgkins survival function ends at around 48 months, where the last recorded event is seen.

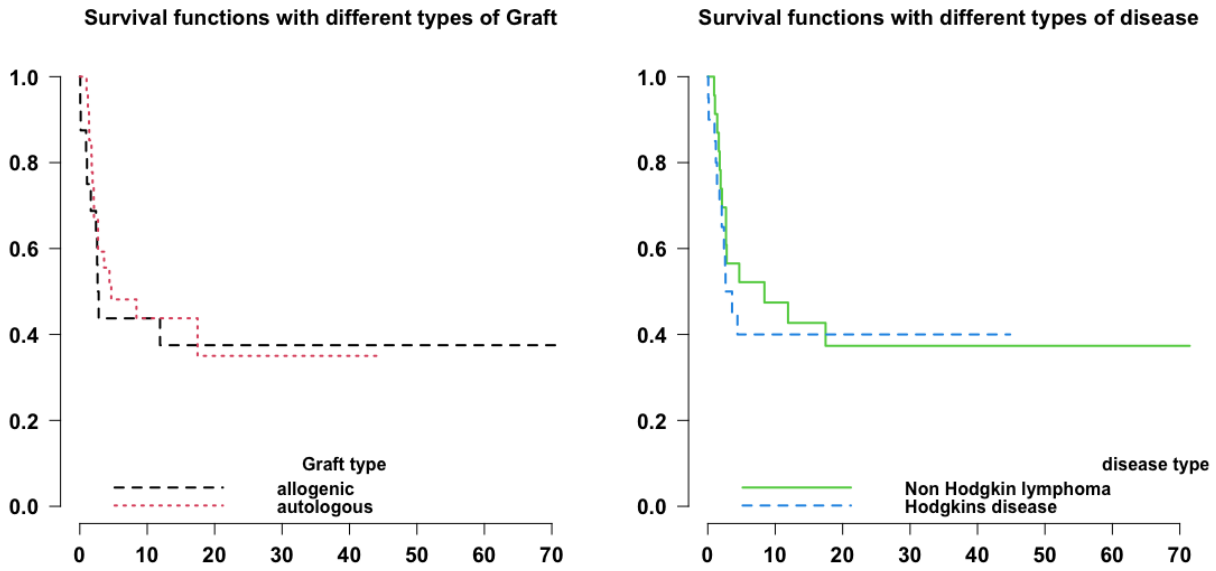

Figure 4: survival functions corresponding to the four combinations of graft and disease types in Months

```
> coxh <- update(coxh,  ~ . + gtype:dtype + score)
> summary(coxh)
Call:
coxph(formula = shodg ~ gtype + dtype + score + gtype:dtype,
    data = hodg)

  n= 43, number of events= 26

                                       coef exp(coef) se(coef)      z Pr(>|z|)
gtypeautologous                     0.53270   1.70353  0.58231  0.915   0.3603
dtypeHodgkins disease               1.68314   5.38244  0.69463  2.423   0.0154 *
score                              -0.05471   0.94676  0.01226 -4.463  8.1e-06 ***
gtypeautologous:dtypeHodgkins disease -1.65262   0.19155  0.91614 -1.804   0.0712 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                    exp(coef) exp(-coef) lower .95 upper .95
gtypeautologous                        1.7035     0.5870    0.5441    5.3335
dtypeHodgkins disease                  5.3824     0.1858    1.3795   21.0015
score                                  0.9468     1.0562    0.9243    0.9698
gtypeautologous:dtypeHodgkins disease  0.1915     5.2207    0.0318    1.1537

Concordance= 0.753  (se = 0.053 )
Likelihood ratio test= 28.87  on 4 df,   p=8e-06
Wald test            = 25.82  on 4 df,   p=3e-05
Score (logrank) test = 34.66  on 4 df,   p=5e-07
```

Figure 5: Cox-proportional hazards model

(c) We know the higher the concordance, we have a better fit in our model with a value of over 0.75, close to 1(the perfect fit), we have a good fit. We observe the two disease types and the Karnofsky score variables to be statistically significant in our model, unlike variable graft type and the interaction between the two variables, which are not statistically significant in our model. Finally, we also see all three tests have very low p values, which proves the global statistical significance of the model.

	logHR	HR	Lower 95%	Upper 95%
HR$\|\|$*allogenic*	1.065	2.902	0.296	28.447
HR$\|\|$*autologous*	0.533	1.704	0.544	5.334

(d) The allogenic hazard ratio is higher than the autologous. This proves, as seen in figure 4, in patients with graft type of autologous, we have a higher survival rate.

(e) We check the proportional hazards assumption.

	chisq	df	p
gtype	0.358	1	0.550
dtype	2.428	1	0.119
score	3.691	1	0.055
gtype:dtype	4.630	1	0.031
GLOBAL	12.623	4	0.013

We test the proportional hazard assumption using the Schoenfeld test. The null hypothesis is that the condition holds, and the other hypothesis is that it does not. All of the p-values are over 5%, which leads us to believe that we cannot reject the null hypothesis, but we also check the plots of this test to make things clearer.

The solid lines in the middle of the bands show if we allow the coefficient (the hazard ratio) to change over time, how much it will change. The dash lines are the confidence intervals around that line. No change means the band moves along the zero line, so we add a red line to see how often a change is seen over time.

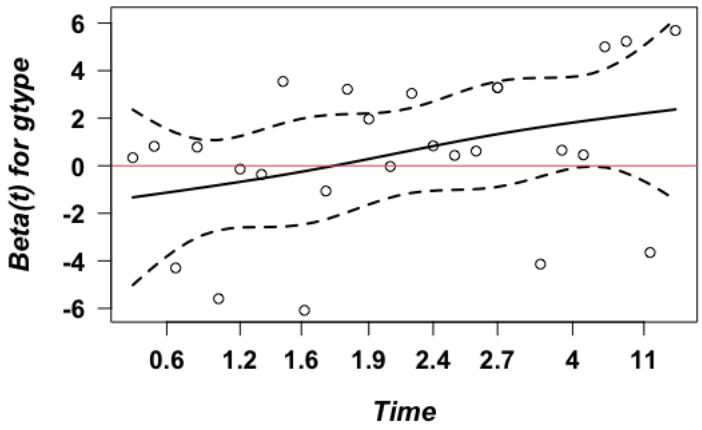

Figure 6: Graft type test.

It can be seen that in graft time the red line is in the bands all of the time. This makes us more confident in our assumptions.

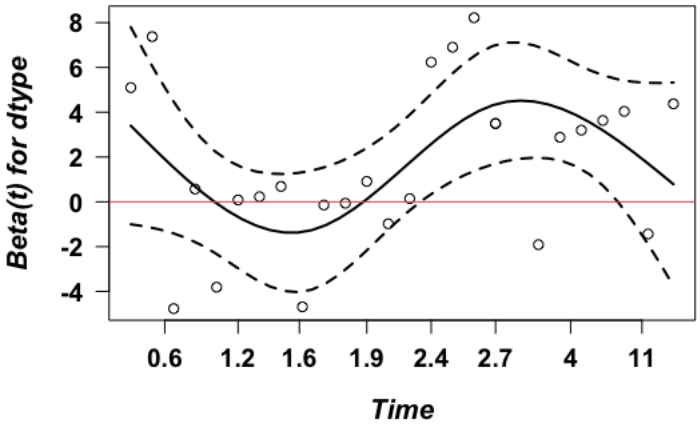

Figure 7: Disease type test.

In the disease type variable, the change over zero is happening more but it is still most of the time in the confidence bands.

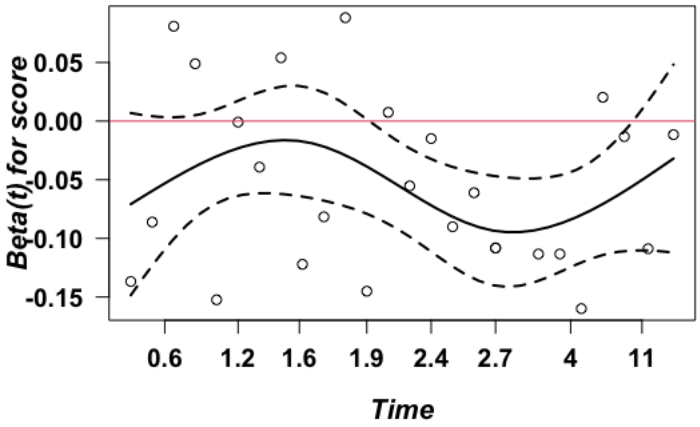

Figure 8: Karnofsky score test.

We can see that this behavior is very different the hazard ratio is changing most of the time. This is against our initial assumption.

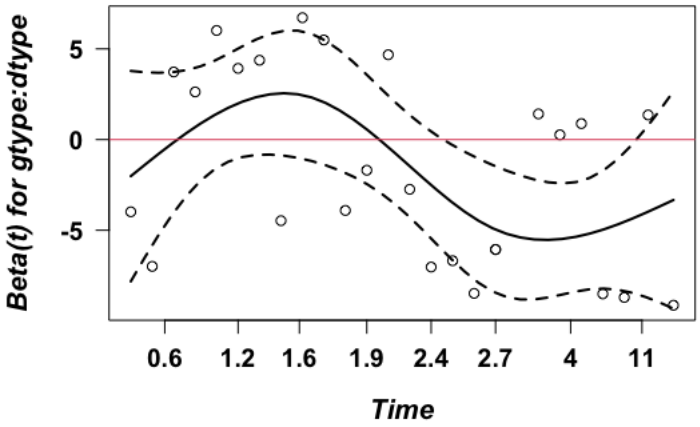

Figure 9: Interaction of graft type and disease type test.

In the interaction of graft type and disease type variable, the change of coefficient over time is happening, but it is still most of the time in the confidence bands.

(f) Influence of each individual in the global fit. We use Residuals based on the scores.

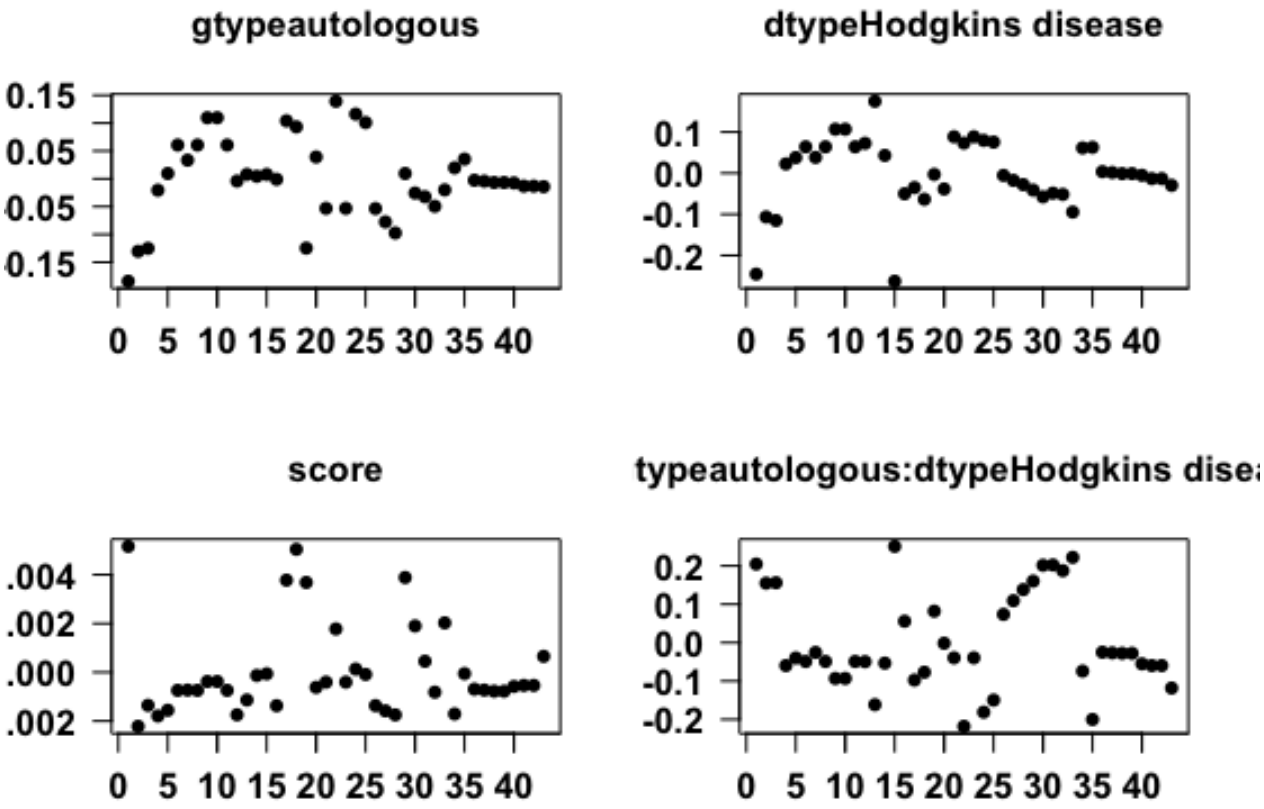

Figure 10: Residuals based on the scores

The above plots show that comparing the magnitudes of the largest dfbeta values to the regression coefficients suggests that some of the observations are influential individually, even though some of the dfbeta values for score are small compared with the others.

R Code

```
##### EXERCISE 1 #####
data(tongue)
tongue$type=factor(tongue$type, labels=c("aneuploid","diploid"))
summary(tongue)
attach(tongue)
```

```
### 1a
ston <- with(tongue, Surv(time, delta))
svf.type=npsurv(ston~type, tongue)
par(font = 2, font.axis = 2, font.lab = 4, las = 1)
survplot(svf.type, ylab = expression(bold(hat(S)(t))), col = 2:3,
   lwd = 3, xlab="Time to death [Weeks]")
title("Survival functions according to Tumor DNA profile")

### 1b
ston.type<- with(tongue, Surv(time, delta) ~ type)
survdiff(ston.type)
#H0: no diff
#H1: yes diff
#p=0.09 -> not reject

### 1c
svf.loglo=survreg(ston ~ type, tongue, dist = "loglo")
summary(svf.loglo)
#H0: type is not impacting the surv time param=0 for type
#H1: type is incluencing surv time
#p=0.051 -> not reject

### 1d
#acceleration factor
exp(-svf.loglo$coefficient[2])
#odds ratio
exp(-svf.loglo$coefficient[2] / svf.loglo$scale)


### 1e
lnopred <- predict(lnomod3, type = "linear")
residsLN <- (log(larynx$time) - lnopred) / lnomod3$scale
residsLN

loglo.pred <- predict(svf.loglo, type = "linear")
summary(loglo.pred)
resids <- (log(tongue$time) - loglo.pred) /svf.loglo$scale
resids
par(font = 2, font.lab = 4, font.axis = 2, las = 1, oma = c(0, 0,
   1, 0),
    mar = c(5, 5, 4, 2))
```

```
plot(survfit(Surv(resids, tongue$delta) ~ 1), xlab = "Years", lwd
   = 2,
     ylab = expression(bold(hat(S)(t))), yaxs = "i")
title("Residuals of the Loglogistic regression model")
survgumb <- function(x) {
  return(exp(-exp(x)))
}
curve(survgumb(x), from = min(resids), to = max(resids), col = 2,
   lwd = 3,
      add = TRUE)


#### EXERCISE 2 ####
### 2a
survlog=rlnorm(300,2,1)
### 2b
lam=1/20
censexp=rexp(300,lam)

### 2c
Y=pmin(survlog,censexp)
del=as.numeric(survlog<=censexp)
#proportion of censoring
table(del)
(sum(del)/length(del))
head(cbind(survlog, censexp, Y, del))

### 2d
#stype=2 - cum.hazard
#ctype=1 - Nelson-Aalen Formula
svf <- survfit(Surv(Y, del) ~ 1, stype = 2, ctype = 1)
svf
summary(svf)
## Uncensored survival times
times <- summary(svf)$time

## Nelson-Aalen estimate of the cumulative hazard function
chaz <- -log(summary(svf)$surv)

## The probability plots
par(mfrow = c(1, 2), las = 1, font.lab = 4, font.axis = 2, pch =
   16)
```

```
plot(log(chaz) ~ log(times), xlab = "Log(Time)",
     ylab = "Log(Cumulative hazard)", col = 4)
abline(lm(log(chaz) ~ log(times)), lwd = 3)
title("Check for Weibull distribution")

plot(log(exp(chaz) - 1) ~ log(times), xlab = "Log(Time)",
     ylab = "Log(Exp(Cumulative hazard) - 1)", cex = 1.3, col = 4)
abline(lm(log(exp(chaz) - 1) ~ log(times)), lwd = 3)
title("Check for Log-logistic distribution")


#### EXERCISE 4 ####

data(hodg)
### 4a
hodg$gtype=factor(hodg$gtype, labels=c("allogenic", "autologous"))
hodg$dtype=factor(hodg$dtype, labels=c("Non Hodgkin lymphoma", "
   Hodgkins disease"))
summary(hodg)
### 4b
?hodg
hodg$months <- hodg$time / 30
shodg <- with(hodg, Surv(months, delta))
summary(shodg)
plot(shodg)
# (b) Draw the survival functions corresponding to the four
   combinations of graft and disease types
# measuring time until relapse or death in years. Comment on the
   graph.
par(mfrow = c(1, 2), font = 2, font.lab = 4, font.axis = 2, las =
   1, mar = c(3, 3, 4, 2))

plot(survfit(shodg ~ gtype, hodg), col = 1:2, xlab = "Months", lwd
    = 2, lty = 2:3,
     ylab = expression(bolditalic(hat(S)(t))), bty = "n")
legend("bottomleft", levels(hodg$gtype), col = 1:2, lwd = 2, bty =
    "n", lty = 2:3,
       title = "Graft type")
title("Survival functions with different types of Graft")

plot(survfit(shodg ~ dtype, hodg), col = 3:4, xlab = "Months", lwd
    = 2, lty = 1:2,
```

```
      ylab = expression(bolditalic(hat(S)(t))), bty = "n")
legend("bottomleft", levels(hodg$dtype), col = 3:4, lwd = 2, bty =
    "n", lty = 1:2,
       title = "disease type")
title("Survival functions with different types of disease")

# (c) Fit the proportional hazards model that includes graft type,
    disease type, the interaction of both,
# and the Karnofsky index. Interpret the model fit.
## The fit of a Cox model
## ======================
(coxh <- coxph(shodg ~ gtype + dtype, hodg))


## Including the interaction between both variables and the
   Karnofsky index
##
   ------------------------------------------------------------------------


coxh <- update(coxh, ~ . + gtype:dtype + score)
summary(coxh)
coxh$var

# (d) Estimate the hazard ratios associated to the graft type (
   comparing autologous to allogenic
# transplantations) and interpret both values.
library(Epi)
ci.lin(coxh)
round(ci.lin(coxh, Exp = TRUE), 3)
(ctmat <- matrix(c(2, 0, 0, 0, 1, 0, 0, 0), byrow = TRUE, nr = 2))
round(ci.lin(coxh, ctr.mat = ctmat, Exp = TRUE), 3)

# A somewhat nicer presentation
HRmat <- round(ci.lin(coxh, ctr.mat = ctmat, Exp = TRUE), 3)[, c
   (1, 5:7)]
rownames(HRmat) <- c("HR| allogenic", "HR| autologous")
colnames(HRmat) <- c("logHR", "HR", "Lower 95%", "Upper 95%")
HRmat

# (e) Check the proportional hazards assumption.
## Checking proportional hazard assumption
## ---------------------------------------
```

```
residuals(coxh, "schoenfeld")
## (i) Use of function cox.zph
chpa =  cox.zph(coxh)
## (ii) Use of function plot.cox.zph
#windows(width = 12, height = 7)
par(mfrow = c(1, 1), font = 2, font.lab = 4, font.axis = 2, las =
    1,
     cex.lab = 1.3, cex.axis = 1.2)
plot(chpa[1], lwd = 2)
abline(h=0, col= 2)
plot(chpa[2], lwd = 2)
abline(h=0, col= 2)
plot(chpa[3], lwd = 2)
abline(h=0, col= 2)
plot(chpa[4], lwd = 2)
abline(h=0, col= 2)
# (f ) Concerning the estimation of the four model parameters, are
    there any in
# influential observations?
dfbet <- residuals(coxh, type = "dfbeta")
dim(dfbet)
par(mfrow = c(2, 2), font = 2, font.lab = 4, font.axis = 2, las =
    1,
     cex.lab = 1.3, cex.axis = 1.2)
for (i in 1:5) {
  plot(dfbet[, i], pch = 16, ylab = "")
  title(names(coef(coxh))[i])
  axis(1, at = seq(5, 45, 5))
}
```