Obligatory Exercise 2

TMA4275 Lifetime analysis Spring 2019

Candidate number: 10006

23 mars 2019

In this exercise we used the **R**-libraries:

```
# for ggplot and dataframe
library(tidyverse)
# for survival analysis functions
library(survival)
# to plot survival curves in ggplot
library(survminer)
# to create tables in rmd
library(kableExtra)
# to get more information about surival regression
library(SurvRegCensCov)
# to predict usable covariates
library(rms)
```

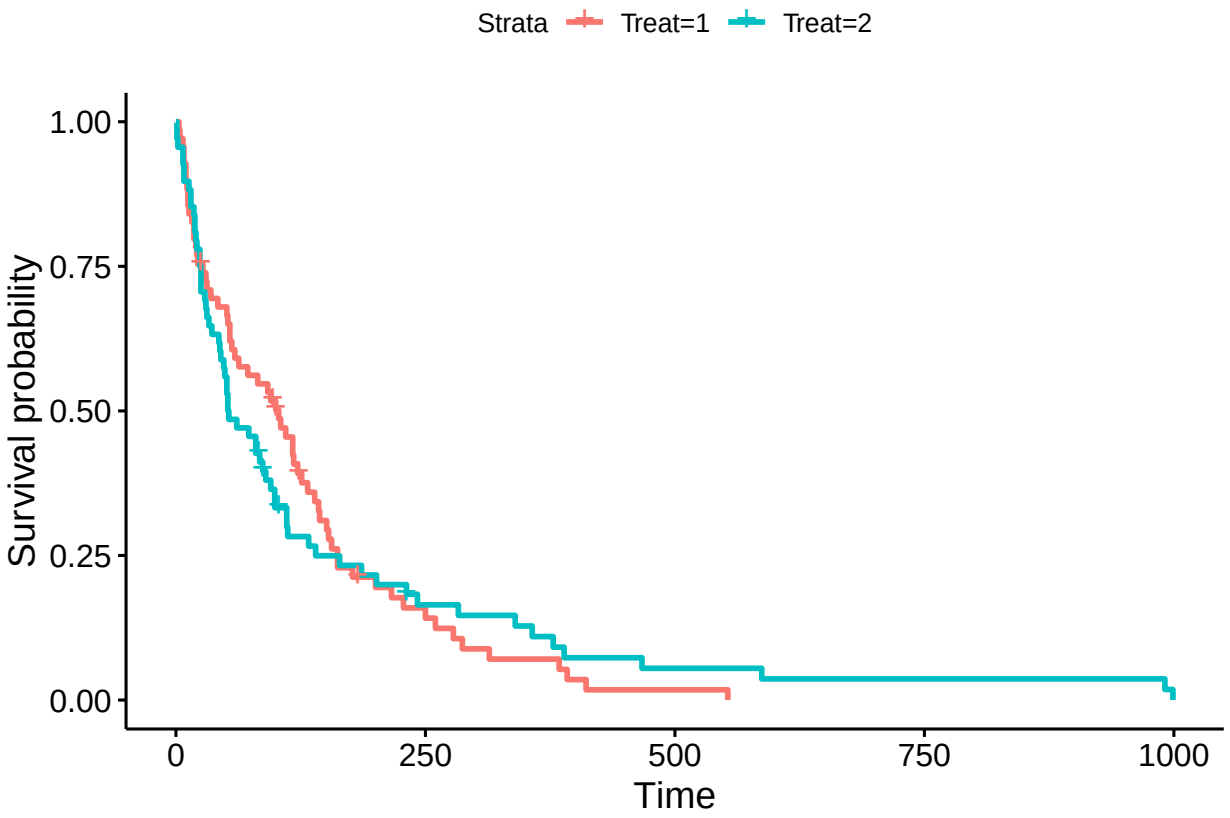
a)

We want to look at the lifetime distributions of the patients in the different treatment groups, *standard* and *test*.

```
lungcancer.df <- as.data.frame(
  read.table("../data/TMA4275VeteranLungCancer.txt",
             header = TRUE))
KM0 <- survfit(Surv(Y, C) ~ Treat,
               type="kaplan-meier",
               conf.type="log",
               data=lungcancer.df)
ggsurvplot(KM0, data = lungcancer.df)
```

Table 1: Estimated median and expected lifetime by treatment

treatment	median	expected
Treat=1	103.0	123.9282
Treat=2	52.5	134.0478

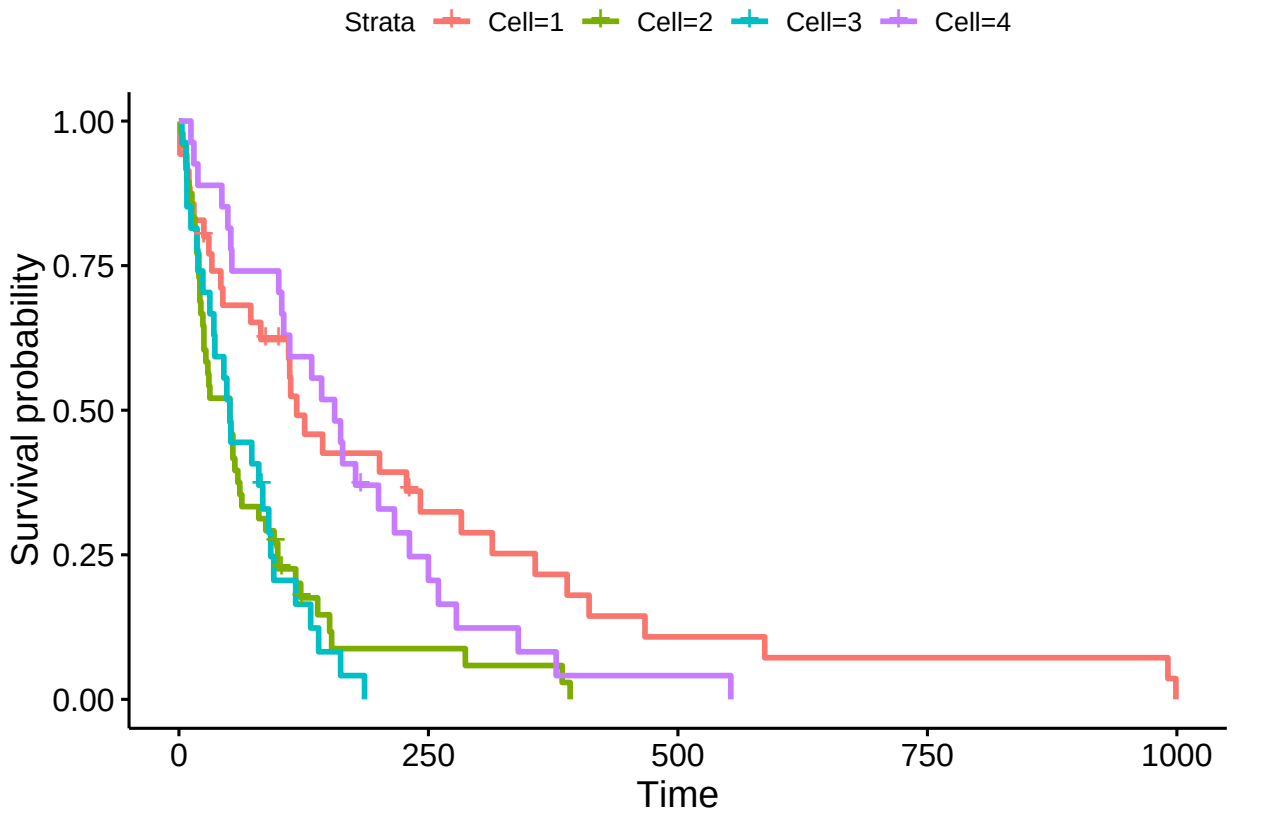


In the figure above the red curve is the survival distribution of the *standard* treatment and the blue curve is the survival distribution of the *test* treatment. From this plot we can see that the *test* treatment is lower for small lifetimes, but doesn't hit the survival probability of 0% before double the lifetime of the *standard* treatment. This tells us that *test* is worse for '*bad*' patients, but it is better for '*good*' patients. What a '*bad*' or '*good*' patient is, is hard to say anything about yet, but might be possible after furter survival analysis and the inclusion of more covariates. The median and expected lifetimes of patients in the two groups is calculated in the code below and the results of which is shown in 1. From this we can see that median of the survival probaility is lower for the *test* treatment, but the expected lifetime is higher than *standard* treatment.

```
treat.median <- summary(KM0)$table[, "median"]
treat.expected <- summary(KM0)$table[, "*rmean"]
table.treat <- data.frame(treatment = names(treat.median),
                          median = unname(treat.median),
                          expected = unname(treat.expected))
kable(table.treat,
      caption = "\\label{tab:treat}Estimated median and expected
      lifetime by treatment") %>%
  kable_styling(bootstrap_options = c("striped", "hover",
                                      "condensed","responsive"),
                full_width = F,position = "center")
```

Now we do a similar assesment of a grouping by celltypes, which consist of the four different celltypes *Squamous*(1), *small-cell*(2), *adeno*(3) and *large*(4).

```
KM1 <- survfit(Surv(Y, C) ~ Cell,
               type="kaplan-meier",
               conf.type="log",
               data=lungcancer.df)
ggsurvplot(KM1, data = lungcancer.df)
```



The figure above shows the survival distributions in the grouping by celltype. Overall $Cell = 2$ and $Cell = 3$ is the worst, and $Cell = 1$ looks like the best $Cell$ for lifetime. The code bellow calculates the median and expected lifetimes in the grouping by celltype, and the results are shown in the table 2.

```
cell.median <- summary(KM1)$table[, "median"]
cell.expected <- summary(KM1)$table[, "*rmean"]
table.cell <- data.frame(treatment = names(cell.median),
                         median = unname(cell.median),
                         expected = unname(cell.expected))
kable(table.cell,
      caption = "\\label{tab:cell}Estimated median
      and expected lifetime by cell type") %>%
  kable_styling(bootstrap_options = c("striped", "hover",
                                      "condensed","responsive"),
                full_width = F,position = "center")
```

From this result we see that the $Cell = 1$ has the highest expected lifetime, followed by $Cell = 4$. But the roles are reversed for the median. For $Cell = 2$ and $Cell = 3$ they are almost equal in both accounts, but has much lower lifetimes than the two others. The same conclusions as the one we drew from the figures. The reason the values of the median and expected lifetime is not being consistently heigher than other group factors, is because there are alot of data beeing either censored or dead at the lower lifetimes. But the survival distribution might have data points at high lifetimes, and thereby have a long right tail that will affect the expected lifetime alot.

Table 2: Estimated median and expected lifetime by cell type

treatment	median	expected
Cell=1	118	188.45594
Cell=2	51	78.98110
Cell=3	51	65.55556
Cell=4	156	167.19342

Table 3: Estimated median and expected lifetime by cell type

Levels.of.Cell	x2	x3	x4
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

b)

In this exercise we will perform a fit to the Weibull regression model, with *Treat*, *PS*, *Month*, *Age* and *Prior* as ordinary covariates, and *Cell* as a factor represented by three dummy variables, x_2, x_3 and x_4. The result of which is shown in the code below, using `survreg` from the R-package **survival**.

```
wei.lung<-survreg(Surv(Y,C) ~ Treat + PS + Month+  Age + Prior + factor(Cell), data = lungcancer.df, dis
wei.lung
```

```
## Call:
## survreg(formula = Surv(Y, C) ~ Treat + PS + Month + Age + Prior +
##     factor(Cell), data = lungcancer.df, dist = "weibull")
##
## Coefficients:
##   (Intercept)         Treat            PS         Month           Age
##   3.490536315  -0.228522672   0.030068303  -0.000468814   0.006099184
##         Prior factor(Cell)2 factor(Cell)3 factor(Cell)4
##  -0.004389765  -0.826184615  -1.132725093  -0.397680785
##
## Scale= 0.9281153
##
## Loglik(model)= -715.6   Loglik(intercept only)= -748.1
##  Chisq= 65.08 on 8 degrees of freedom, p= 4.65e-11
## n= 137
```

In the table 3 the defined indicators or dummy variables are shown. And we adapt this to our regression model.

```
table.b <- data.frame(Levels.of.Cell = c(1,2,3,4), x2 =c(0,1,0,0), x3 = c(0,0,1,0), x4 = c(0,0,0,1))
kable(table.b,align=rep('c', 4),
      caption = "\\label{tab:b}Estimated median
      and expected lifetime by cell type") %>%
  kable_styling(bootstrap_options = c("striped", "hover",
                                      "condensed","responsive"),
                full_width = F,position = "center")
```

Given $x = \{x_1(\text{Treat}), x_2(\text{Cell=2}), x_3(\text{Cell=3}), x_5(\text{Cell=4}), x_5(\text{PS}), x_6(\text{Month}), x_7(\text{Age}), x_8(\text{Prior})\}$, the values of the covariates given some person, where the "*dummy variables*" x_2, x_3, x_4 is set by table 3, the model

for log-lifetime T is given by the expression, with corresponding covariates $\mathbf{x}$,

$$ln(T) = \beta_0 + \sigma\epsilon + \sum_{i=1}^{8} \beta_i x_i.$$

Where $\epsilon \sim G(0, \sigma)$, a gumbel distribution with shape σ. This gives us the hazard rate function

$$\begin{aligned} z(t) &= \alpha \cdot exp\{\beta_0 + \textstyle\sum_{i=1}^{8} \beta_i x_i\}^{-\alpha} \cdot t^{\alpha-1} \\ &= \alpha e^{-\alpha\beta_0} t^{\alpha-1} e^{-\alpha \cdot \sum_{i=1}^{8} \beta_i x_i} \\ &= z_0(t) \cdot e^{-\alpha \cdot \sum_{i=1}^{8} \beta_i x_i} \end{aligned} .$$

Now we will calculate the estimated median lifetimes of patients with respectivly the same covariates as patient number 1 and 19. Which means we look at the factor of *Cell*, because this is the only covariate that would have discluded variables in the model based on its value. This means that we use for patient 1, Cell = 1 and patient 19 we use Cell = 2. In the function `lifeWeil` the median lifetimes of a set of data is calculated.

```
coef <- wei.lung$coefficients
lifeWeil <- function(data, coef){
  res = vector()
  for (i in seq(1,length(data[,1]))){
    x = c(1,data$Treat[i], data$PS[i], data$Month[i], data$Age[i], data$Prior[i],as.numeric(data$Cell[i
          as.numeric(data$Cell[i]==3),as.numeric(data$Cell[i]==4))
    res = c(res,exp(as.numeric(coef)%*%x))
  }
  return(median(sort(res)))
}
patient1 <- lungcancer.df[lungcancer.df$Cell == lungcancer.df$Cell[1],]
patient2 <- lungcancer.df[lungcancer.df$Cell == lungcancer.df$Cell[19],]
life1 <- lifeWeil(patient1,coef)
life2 <- lifeWeil(patient2,coef)
cat("Estimated median lifetime equal patient 1:", life1)
```

```
## Estimated median lifetime equal patient 1: 240.7295
```

```
cat("Estimated median lifetime equal patient 19:", life2)
```

```
## Estimated median lifetime equal patient 19: 83.78111
```

c)

To look at the p-values to determine significant effect of the covariates in our model, we can look at the output from the function `summary.survreg`.

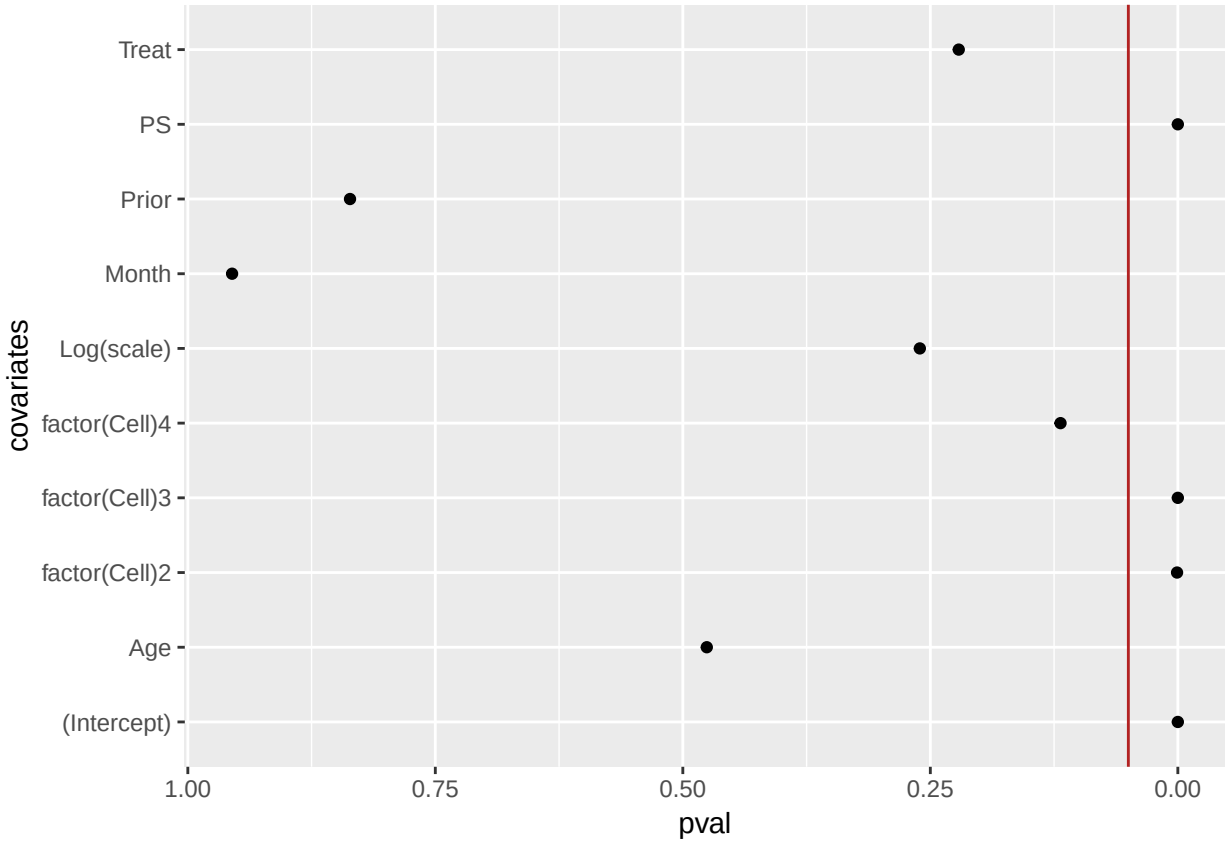
```
summary(wei.lung)
```

```
##
## Call:
## survreg(formula = Surv(Y, C) ~ Treat + PS + Month + Age + Prior +
##     factor(Cell), data = lungcancer.df, dist = "weibull")
##                   Value Std. Error     z       p
## (Intercept)    3.490536   0.691171  5.05 4.4e-07
## Treat         -0.228523   0.186844 -1.22  0.2213
## PS             0.030068   0.004828  6.23 4.7e-10
## Month         -0.000469   0.008361 -0.06  0.9553
## Age            0.006099   0.008553  0.71  0.4758
```

```
## Prior          -0.004390   0.021228 -0.21  0.8362
## factor(Cell)2 -0.826185   0.246312 -3.35  0.0008
## factor(Cell)3 -1.132725   0.257598 -4.40 1.1e-05
## factor(Cell)4 -0.397681   0.254749 -1.56  0.1185
## Log(scale)    -0.074599   0.066311 -1.12  0.2606
##
## Scale= 0.928
##
## Weibull distribution
## Loglik(model)= -715.6   Loglik(intercept only)= -748.1
##  Chisq= 65.08 on 8 degrees of freedom, p= 4.7e-11
## Number of Newton-Raphson Iterations: 6
## n= 137
```

From the p-values we can see that PS and $Cell$ has significant effect. We can also look at the figure below to to get a more visual representation of the p-values.

```
temp<-data.frame(covariates = names(summary(wei.lung)$table[,4]) ,pval = summary(wei.lung)$table[,4])
ggplot(temp) +
  geom_point(aes(x = pval, y = covariates)) +
  geom_vline(xintercept = 0.05, color = "firebrick") +
  scale_x_reverse()
```



The red line is the significance level of 0.05. We observe that only PS and two factors of $Cell$ is within acceptable p-values.

To look at the significant difference between the two treatments we use the function `ConvertWeibull` from the library **SurvRegCensCov**. Using this function we get the HR(Hazard rate) or Risk of Death. We derive this by hand later in this exercise.

```
ConvertWeibull(wei.lung)$HR
```

```
##                       HR        LB        UB
## Treat           1.279184 0.8598048 1.9031196
## PS              0.968122 0.9578947 0.9784585
## Month           1.000505 0.9829924 1.0183301
## Age             0.993450 0.9756476 1.0115771
## Prior           1.004741 0.9607056 1.0507948
## factor(Cell)2 2.435555 1.4262021 4.1592473
## factor(Cell)3 3.388737 1.9026899 6.0354235
## factor(Cell)4 1.534926 0.8901011 2.6468874
```

Looking at the output and HR value for $Treat$, we can conclude that doing the *test* treatment will increase the risk of death by approximately 28%. The scale parameter of our Weibull distribution is α =0.9281153, which is pretty close to an exponential disitribution of $\alpha = 1$. This means that an exponential disitribution might be a better model than our Weibull model. To determine this we need a better way to compare the models. One way of doing this is to calculate the AIC(Akaike's Information Criterion)-values of the two, and then choose the model with the smallest value. We will do this later in this exercise, but we recognize that the scale parameter is close to that of a exponential distribution. Now we will take out the non-significant covariates in our model. This can be done by looking at the p-values found previously. Instead of just choosing, we can use the function `fastbw` from the library **rms** to determine which covariates to use in our model based on p-values. If we choose a significance level of 5% and then force our model to include $Treat$ as a covariate, we get the results shown in the code below, under '*Factors in Final model*'.

```
psm.lung <- psm(Surv(Y,C)~Treat + PS + Month+  Age + Prior + factor(Cell), data = lungcancer.df,dist="w
fastbw(psm.lung,rule = "p",sls = 0.05, force=c(1),type = "individual")
```

```
##
## Parameters forced into all models:
##  Treat
##
##  Deleted Chi-Sq d.f. P      Residual d.f. P      AIC
##  Month   0.00   1    0.9553 0.00     1    0.9553 -2.00
##  Prior   0.07   1    0.7957 0.07     2    0.9655 -3.93
##  Age     0.55   1    0.4577 0.62     3    0.8914 -5.38
##
## Approximate Estimates after Deleting Factors
##
##                 Coef    S.E. Wald Z         P
## (Intercept)  3.84429 0.46356  8.293 1.110e-16
## Treat       -0.21046 0.18089 -1.163 2.446e-01
## PS           0.02906 0.00456  6.373 1.853e-10
## Cell=2      -0.80044 0.23854 -3.356 7.918e-04
## Cell=3      -1.10055 0.25107 -4.383 1.168e-05
## Cell=4      -0.38904 0.25390 -1.532 1.255e-01
##
## Factors in Final Model
##
## [1] Treat PS    Cell
```

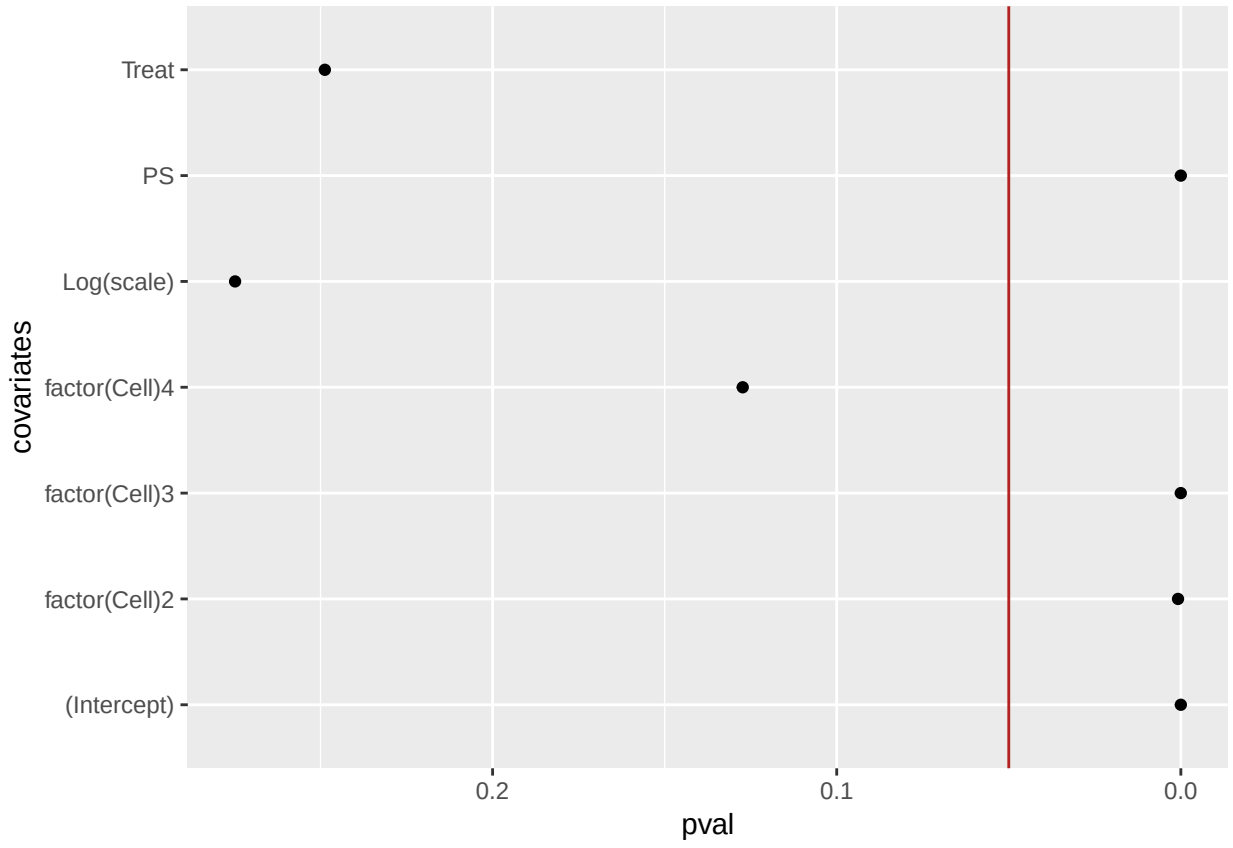
This means that the reduced model includes the covariates $Treat$, PS and the factors of $Cell$. Which is also what we concluded earlier.

```
wei.lung.new <- survreg(Surv(Y,C)~Treat + PS + factor(Cell), data = lungcancer.df, dist = "weibull")
summary(wei.lung.new)
```

```
##
## Call:
## survreg(formula = Surv(Y, C) ~ Treat + PS + factor(Cell), data = lungcancer.df,
##     dist = "weibull")
##                  Value Std. Error     z       p
## (Intercept)    3.84267    0.46262  8.31 < 2e-16
## Treat         -0.20942    0.18157 -1.15 0.24875
## PS             0.02907    0.00456  6.38 1.8e-10
## factor(Cell)2 -0.79950    0.23891 -3.35 0.00082
## factor(Cell)3 -1.09977    0.25198 -4.36 1.3e-05
## factor(Cell)4 -0.38762    0.25422 -1.52 0.12732
## Log(scale)    -0.07236    0.06626 -1.09 0.27483
##
## Scale= 0.93
##
## Weibull distribution
## Loglik(model)= -715.9   Loglik(intercept only)= -748.1
##  Chisq= 64.47 on 5 degrees of freedom, p= 1.4e-12
## Number of Newton-Raphson Iterations: 5
## n= 137
```

Above we have done a Weibull fitting of the new reduced model. Looking at the scale α =0.9301971, it is closer to that of a exponential distribution. If we look at the p-value plots below, we can see that only $Treat$ and $Cell = 4$ fall outside of the significance level of 0.05.

```
temp2<-data.frame(covariates = names(summary(wei.lung.new)$table[,4]) ,pval = summary(wei.lung.new)$tabl
ggplot(temp2) +
  geom_point(aes(x = pval, y = covariates)) +
  geom_vline(xintercept = 0.05, color = "firebrick") +
  scale_x_reverse()
```



If we again look at the risk of death(HR) values from this model fit, the significant difference between the two treatments have dropped alittle, but the risk of death for the *test* treatment is 25% higher than for the *standard* treatment.

```
ConvertWeibull(wei.lung.new,conf.level = 0.95)$HR
```

```
##                       HR        LB        UB
## Treat          1.2524947 0.8523129 1.8405715
## PS             0.9692303 0.9596117 0.9789453
## factor(Cell)2 2.3619730 1.4069946 3.9651302
## factor(Cell)3 3.2618603 1.8584899 5.7249343
## factor(Cell)4 1.5169594 0.8822649 2.6082481
```

d)

The Cox-model for the reduced model we found in c) is given by the equation

$$z(t;\mathbf{x}) = z_0(t)e^{\beta_1 x_1+\beta_2 x_1+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5},$$

where $z_0(t)$ is any non-parametric hazard rate function and is unkown in the cox model. The reduced Weibull model for hazard rate is given by the equation

$$z(t;\mathbf{x}) = \alpha e^{-\alpha\beta_0}\cdot t^{\alpha-1}\cdot e^{-\alpha(\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5)},$$

and the $z_0(t)$ function is known. Now we perform a cox regression on our data.

```
fit.coxph <- coxph(Surv(Y,C) ~ Treat + PS+ factor(Cell), data = lungcancer.df)
fit.coxph
```

```
## Call:
## coxph(formula = Surv(Y, C) ~ Treat + PS + factor(Cell), data = lungcancer.df)
##
##                     coef exp(coef)  se(coef)      z        p
## Treat           0.261744  1.299194  0.200923  1.303  0.19267
## PS             -0.031271  0.969213  0.005165 -6.054 1.41e-09
## factor(Cell)2   0.824980  2.281836  0.268911  3.068  0.00216
## factor(Cell)3   1.153994  3.170833  0.295038  3.911 9.18e-05
## factor(Cell)4   0.394625  1.483828  0.282243  1.398  0.16206
##
## Likelihood ratio test=61.07  on 5 df, p=7.307e-12
## n= 137, number of events= 128
```

If we look at the relative risk of this Cox-model. The relative risk is found by increasing the value of one covariate by 1 and keeping all the other covariates constant. Then calculating the quotient of the new and previous hazard rate function, $z(t, x_i + 1)$, $z(t, x_i)$, one gets the relative risk RR. Given by the equation

$$RR = e^{\beta_i},$$

i being the covariate of interest, where in our case $i \in \{1, 2..., 5\}$. This means that from the output of our Cox-regression above we look at the values of $exp(coef)$, which is the relative risk factor. The highest relative risk is from Cell = 3. The risk of death is increased with 30% if one uses the *test* treatment istead of the *standard* treatment. If one increases the *Performance status* by 1 the change is not that significant. But since PS has a large range, $x_5 \in [10, 90]$, the change is more significant than what the relative risk factor is suggesting.

e)

In this exercise we will look at the interaction between two covariates, $Treat$ and $Cell$, while keeping $Cell$ as a factor. We are still using table 3 for the values of x_2,x_3 and x_4. With the interaction of the covariates, we introduce three new variables $z_2 = x_1 \cdot x_2$, $z_3 = x_1 \cdot x_3$ and $z_4 = x_1 \cdot x_3$. We will use the same covariates as in the reduced model from c), $Treat$, PS, and $Cell$, and we use Weibull regression to determine the β_i- values.

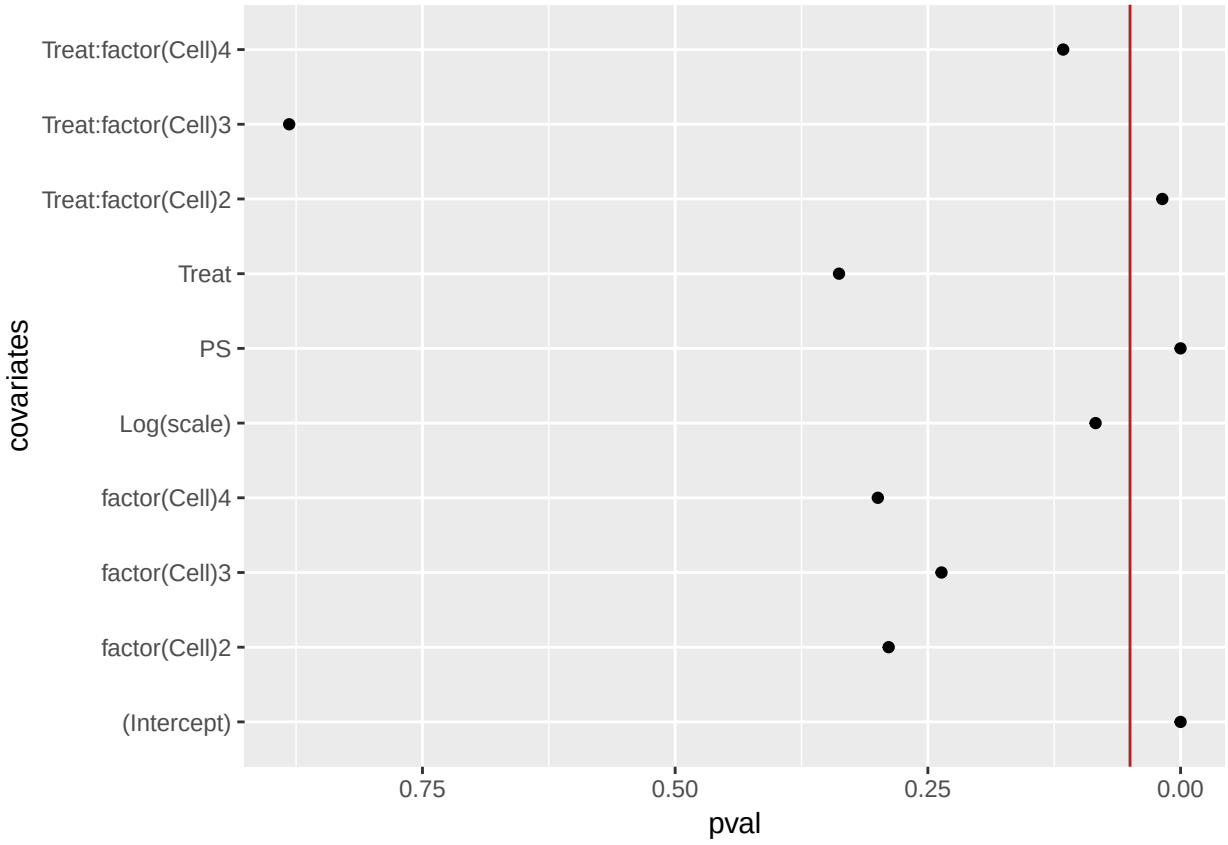
```
wei.lung.int<- survreg(Surv(Y,C) ~ Treat + PS+ factor(Cell) + Treat*factor(Cell), data = lungcancer.df,
summary(wei.lung.int)
```

```
##
## Call:
## survreg(formula = Surv(Y, C) ~ Treat + PS + factor(Cell) + Treat *
##     factor(Cell), data = lungcancer.df, dist = "weibull")
##                          Value Std. Error     z       p
## (Intercept)           3.03586    0.56348  5.39 7.1e-08
## Treat                 0.31709    0.33087  0.96   0.338
## PS                    0.02842    0.00454  6.26 3.8e-10
## factor(Cell)2         0.71458    0.67362  1.06   0.289
## factor(Cell)3        -1.01522    0.85762 -1.18   0.237
## factor(Cell)4         0.80009    0.77114  1.04   0.299
## Treat:factor(Cell)2 -1.02077    0.43169 -2.36   0.018
## Treat:factor(Cell)3 -0.07629    0.51329 -0.15   0.882
## Treat:factor(Cell)4 -0.75657    0.48151 -1.57   0.116
## Log(scale)          -0.11825    0.06845 -1.73   0.084
##
## Scale= 0.888
##
## Weibull distribution
```

```
## Loglik(model)= -712.4   Loglik(intercept only)= -748.1
##  Chisq= 71.45 on 8 degrees of freedom, p= 2.5e-12
## Number of Newton-Raphson Iterations: 5
## n= 137
```

As hinted in the exercise paper the only interaction which has $p-$values lower than a significance level of 0.05 is z_2. Which is also shown in the figure below. Also all of the factors of $Cell$ has now fallen outside of the significance level. PS is still as significant as before.

```
temp3<-data.frame(covariates = names(summary(wei.lung.int)$table[,4]) ,pval = summary(wei.lung.int)$tabl
ggplot(temp3) +
  geom_point(aes(x = pval, y = covariates)) +
  geom_vline(xintercept = 0.05, color = "firebrick") +
  scale_x_reverse()
```



We will now to calculate the estimated relative risk of patients with $Cell = 2$ in the *standard* treatment, compared to the patient with the *test* treatment. First we find a general formula for the Relative Risk(RR) given the $Cell$. Let $\mathbf{x}_1 = \{x_1, x_2, x_3, x_4, x_5, x_1 \cdot x_2, x_1 \cdot x_3, x_1 \cdot x_4\}$ and $\mathbf{x}_2 = \{x_1 + 1, x_2, x_3, x_4, x_5, (x_1 + 1) \cdot x_2, (x_1 + 1) \cdot x_3, (x_1 + 1) \cdot x_4\}$, then the relative risk is given by the equation.

$$\begin{aligned} RR_{\text{weib}}(x_1|Cell = j; j \in \{2,3,4\}) &= \frac{z(t;\mathbf{x}_2|Cell=j)}{z(t;\mathbf{x}_1|Cell=j)} \\ &= \frac{z_0(t)\cdot\exp\{-\alpha\cdot(\beta_1(x_1+1)+\beta_j x_j+\beta_5 x_5+\beta_{j+4}(x_1+1)x_j)\}}{z_0(t)\cdot\exp\{-\alpha\cdot(\beta_1 x_1+\beta_j x_j+\beta_5 x_5+\beta_{j+4}x_1 x_j)\}} \\ &= \exp\{-\alpha\cdot(\beta_1+\beta_{j+4})\} \end{aligned}.$$

If the $Cell = 1$, the expression simply becomes

$$RR(x_1|Cell = 1) = \exp\{-\alpha \cdot \beta_1\}.$$

Table 4: Estimated Relative Risk of *test* treatment compared to *standard* treatment in the different *Cell*-types.

Cell	RR
1	0.7544832
2	1.8686171
3	0.8073987
4	1.4776759

From this knowledge we can create a function in **R** that calculates the relative risk of all the different *Cell* types respectivly. This is done in the function `relativeRisk` and the result of which is shown in table 4.

```
relativeRisk <- function(coef,Cell,scale){
  if (Cell > 1){
    return(exp(-scale*(coef[2] + coef[Cell+5])))
  }
  return(exp(-scale*coef[2]))
}
coef.int <- wei.lung.int$coefficients
scale.int <- wei.lung.int$scale
table.e <- data.frame(Cell = c(1,2,3,4),
                      RR = c(relativeRisk(coef.int,Cell = 1,scale.int),
                             relativeRisk(coef.int,Cell = 2,scale.int),
                             relativeRisk(coef.int,Cell = 3,scale.int),
                             relativeRisk(coef.int,Cell = 4,scale.int)))
kable(table.e,align=rep('c', 4),
      caption = "\\label{tab:e}Estimated Relative Risk of
      \\textit{test} treatment\n compared to \\textit{standard}
      treatment in the different \\textit{Cell}-types.") %>%
  kable_styling(bootstrap_options = c("striped", "hover",
                                      "condensed","responsive"),
                full_width = F,position = "center")
```

From the results in table 4 we can see that for patients with *Cell*-type 1 it is approximately 25% lower risk of death using the *test* treatment. For patients with *Cell*-type 2 it is approximately 87% higher risk of death using the *test* treatment. Further on for patients with *Cell*-type 3, it is approximately 19% lower risk of death using the *test* treatment. And last for patients with *Cell*-type 4, it is approximately 48% higher risk of death using the *test* treatment. Thereby we can conclude that patients with *Cell*-type 1 should take *test* treatment, patients with *Cell*-type 2 should take the *standard* treatment, patients with *Cell*-type 3 should take *test* treatment and patients with *Cell*-type 4 should take the *standard* treatment.

f)

In the final model in exercise e), we used a Weibull regression with a reduced number of covariates and a interaction covariate, $Treat$, $Cell$, PS and $Treat \times Cell$ with $Cell$ as factor. We will now fit different distributions to this model with out data. We tested the exponential, lognormal and log-logistic distributions to the final model in exercise e). We are not taking into account that the different distributions might choose different covariates as better for the model. The number of covariates in our model is $p = 4 (x_1,\ x_2$ or x_3 or $x_4,\ x_5,\ z_2$ or z_3 or $z_4)$. Fit of exponential distribution:

```
exp.lung<- survreg(Surv(Y,C)~Treat + PS + factor(Cell)+ factor(Cell)*Treat, data = lungcancer.df, dist =
summary(exp.lung)
```

```
##
## Call:
```

```
## survreg(formula = Surv(Y, C) ~ Treat + PS + factor(Cell) + factor(Cell) *
##     Treat, data = lungcancer.df, dist = "exponential")
##                       Value Std. Error     z       p
## (Intercept)          2.9784     0.6333  4.70 2.6e-06
## Treat                0.2908     0.3713  0.78   0.433
## PS                   0.0293     0.0050  5.86 4.7e-09
## factor(Cell)2        0.6062     0.7547  0.80   0.422
## factor(Cell)3       -1.0476     0.9598 -1.09   0.275
## factor(Cell)4        0.7759     0.8682  0.89   0.371
## Treat:factor(Cell)2 -0.9428     0.4829 -1.95   0.051
## Treat:factor(Cell)3 -0.0405     0.5733 -0.07   0.944
## Treat:factor(Cell)4 -0.7283     0.5420 -1.34   0.179
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -713.8   Loglik(intercept only)= -751.2
##  Chisq= 74.89 on 8 degrees of freedom, p= 5.2e-13
## Number of Newton-Raphson Iterations: 5
## n= 137
```

Fit of lognormal distribution:

```
lognorm.lung<- survreg(Surv(Y,C)~Treat + PS + factor(Cell)+ factor(Cell)*Treat, data = lungcancer.df, di
summary(lognorm.lung)
```

```
##
## Call:
## survreg(formula = Surv(Y, C) ~ Treat + PS + factor(Cell) + factor(Cell) *
##     Treat, data = lungcancer.df, dist = "lognormal")
##                       Value Std. Error     z       p
## (Intercept)          2.37802    0.66720  3.56 0.00036
## Treat               -0.05468    0.37890 -0.14 0.88525
## PS                   0.03691    0.00484  7.63 2.4e-14
## factor(Cell)2       -0.20063    0.78272 -0.26 0.79770
## factor(Cell)3       -1.09485    0.98365 -1.11 0.26569
## factor(Cell)4        0.42049    0.90826  0.46 0.64339
## Treat:factor(Cell)2 -0.25271    0.49801 -0.51 0.61184
## Treat:factor(Cell)3  0.26728    0.58019  0.46 0.64503
## Treat:factor(Cell)4 -0.21506    0.56873 -0.38 0.70533
## Log(scale)           0.06496    0.06257  1.04 0.29920
##
## Scale= 1.07
##
## Log Normal distribution
## Loglik(model)= -715.4   Loglik(intercept only)= -749.5
##  Chisq= 68.2 on 8 degrees of freedom, p= 1.1e-11
## Number of Newton-Raphson Iterations: 4
## n= 137
```

Fit of log-logistic distribution:

```
loglog.lung<- survreg(Surv(Y,C)~Treat + PS + factor(Cell)+ factor(Cell)*Treat, data = lungcancer.df, dis
summary(loglog.lung)
```

```
##
```

```
## Call:
## survreg(formula = Surv(Y, C) ~ Treat + PS + factor(Cell) + factor(Cell) *
##     Treat, data = lungcancer.df, dist = "loglogistic")
##                          Value Std. Error     z       p
## (Intercept)           2.3864     0.6670  3.58 0.00035
## Treat                 0.0884     0.3848  0.23 0.81825
## PS                    0.0353     0.0045  7.86 3.9e-15
## factor(Cell)2        -0.4157     0.7736 -0.54 0.59106
## factor(Cell)3        -0.8561     0.9160 -0.93 0.35001
## factor(Cell)4         0.5365     0.8665  0.62 0.53581
## Treat:factor(Cell)2 -0.1892     0.4894 -0.39 0.69905
## Treat:factor(Cell)3  0.0431     0.5464  0.08 0.93714
## Treat:factor(Cell)4 -0.3741     0.5377 -0.70 0.48651
## Log(scale)           -0.5450     0.0741 -7.35 2.0e-13
##
## Scale= 0.58
##
## Log logistic distribution
## Loglik(model)= -712.2   Loglik(intercept only)= -750.3
##  Chisq= 76.22 on 8 degrees of freedom, p= 2.8e-13
## Number of Newton-Raphson Iterations: 4
## n= 137
```

From the loglikelihood, the number of parameters p and the distribution factor k we can calculate the AIC of the distributions by the equation

$$AIC = -2 \cdot logL + 2(k + 2 + p).$$

This is the AIC expression suggested by Collett(1994). The AIC values is calculated by this expression in the code under and the results is shown in the table 5.

```
# using p = 4
# k = 1 for weibull
AIC.wei <- -2*wei.lung.int$loglik[1] + 2*(1 + 4 + 2)
# k = 0 for exponential
AIC.exp <- -2*exp.lung$loglik[1] + 2*(0 + 4 + 2)
# k = 1 for lognormal model
AIC.lognorm <- -2*lognorm.lung$loglik[1] + 2*(1 + 4 + 2)
# k = 1 for log-logistic model
AIC.loglog <- -2*loglog.lung$loglik[1]+2*(1 + 4 + 2)

table.f <- data.frame(Distribution = c("weibull",
                                       "exponential",
                                       "lognormal",
                                       "loglogistic"),
                      AIC = c(AIC.wei,AIC.exp,
                              AIC.lognorm,AIC.loglog))

kable(table.f,align=rep('c', 4),
      caption = "\\label{tab:f}AIC values of four different
      distribution fits on the reduced model of the lung cancer
      data set.") %>%
  kable_styling(bootstrap_options = c("striped", "hover",
                                      "condensed","responsive"),
                full_width = F,position = "center")
```

Table 5: AIC values of four different distribution fits on the reduced model of the lung cancer data set.

Distribution	AIC
weibull	1510.182
exponential	1514.442
lognormal	1512.948
loglogistic	1514.532

In the table 5 we can see that the Weibull model gives us the lowest AIC value, and thereby we can conclude that this is the best distribution to fit our model to. It also seems like the lognormal distribution is better than the special case of Weibull, the exponential distribution. We can also see that the scale value in our Weibull model has diverged further away from that of the exponential distribution in the model suggested in e), $\alpha =$`0.8884712`.