Obligatory Exercise 2

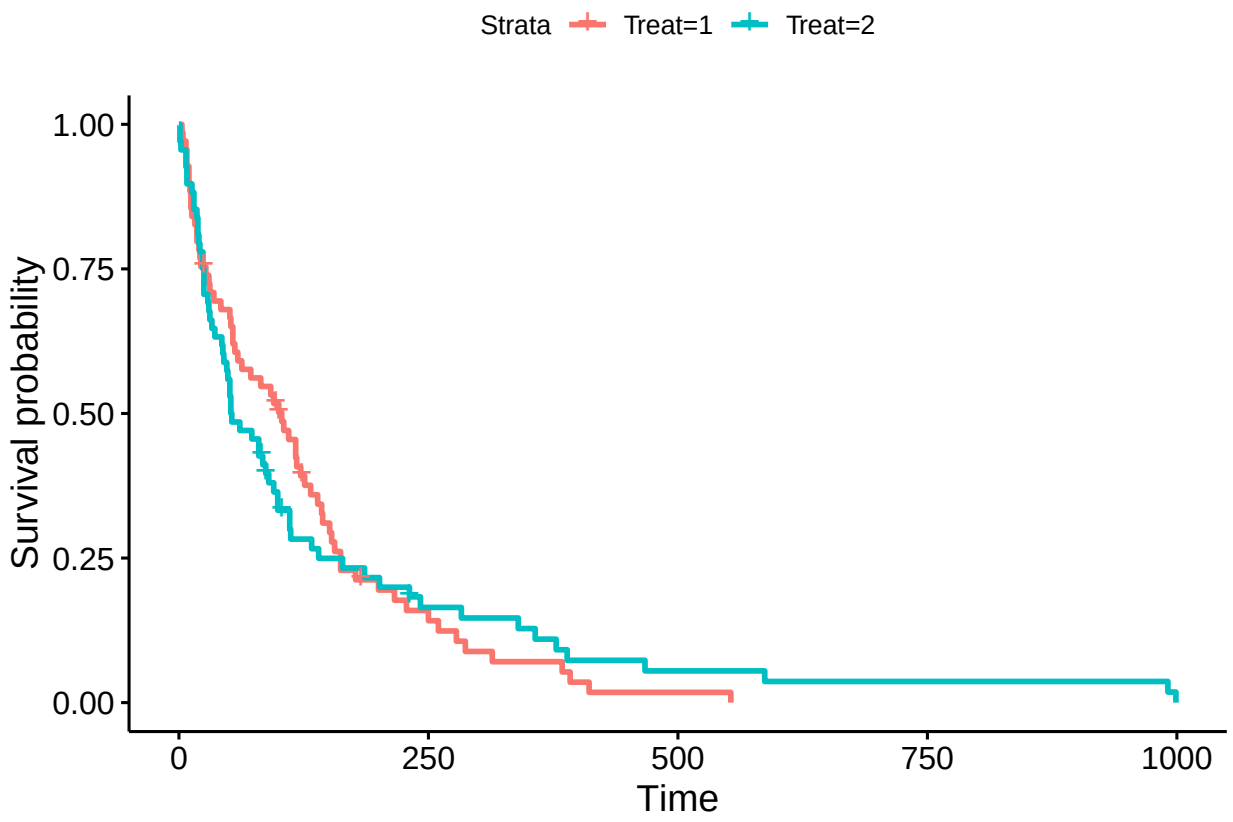
TMA4275 Lifetime analysis Spring 2019

Candidate number: 10006

23 mars 2019

a)

```
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)
```



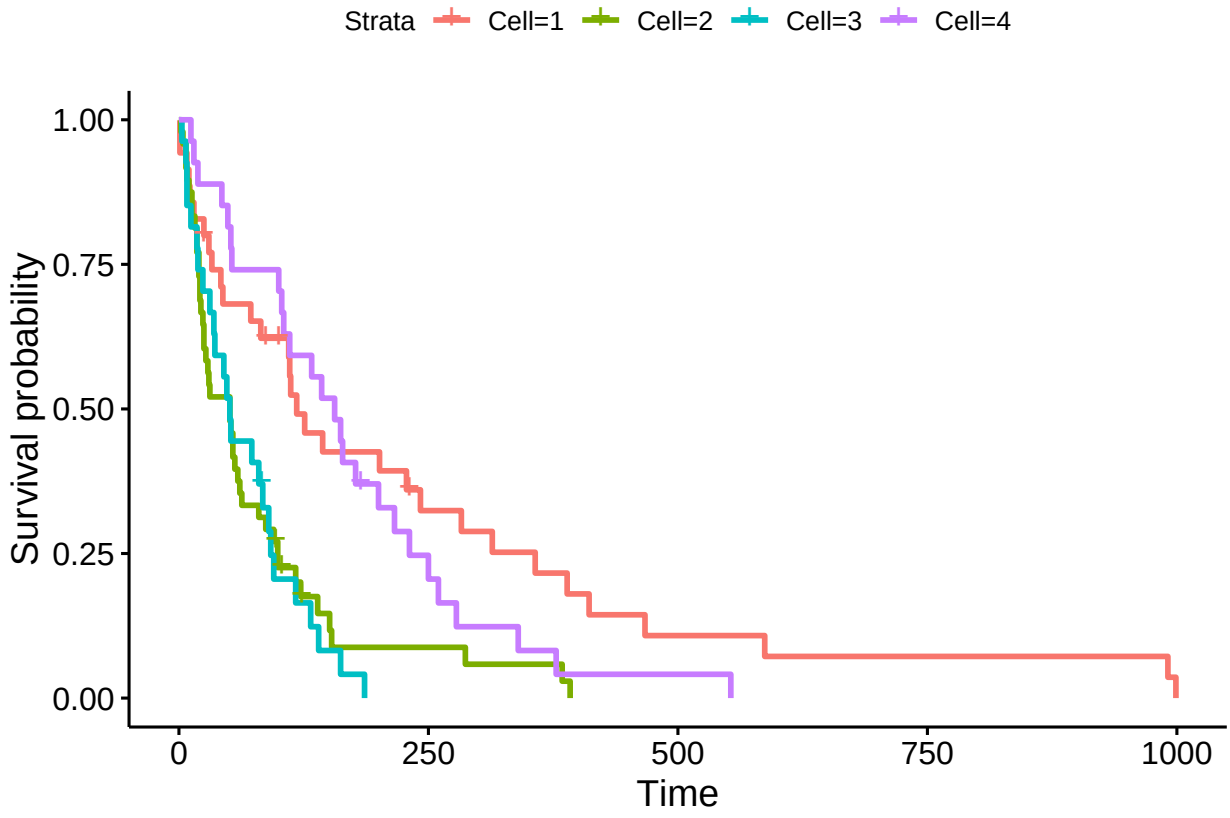

```
treat.median <- summary(KM0)$table[, "median"]
treat.expected <- summary(KM0)$table[, "*rmean"]
table.treat <- data.frame(treatment = names(treat.median),
                          median = unname(treat.median),
                          expected = unname(treat.expected))
kable(table.treat,
      caption = "\\label{tab:treat}Estimated median and expected
      lifetime by treatment") %>%
```

Table 1: Estimated median and expected lifetime by treatment

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

```
  kable_styling(bootstrap_options = c("striped", "hover",
                                      "condensed","responsive"),
                full_width = F,position = "center")
```

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




```
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")
```

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 model parameter, and Cell as a factor. The result of which is shown in the code bellow, 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 is 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, were the "*dummy variables*" x_2, x_3, x_4 is set by table 3. From this

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. This means that for patient 1 which has Cell = 1 and we find a patient with the same, in this case we chose patient 3. For patient 19 which has Cell = 2, we chose patient 23. In the function `lifeWeil` the median lifetime from the our Weibull model 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 bellow to to get a more visual representation of the p-values.

```
psm.lung <- psm(Surv(Y,C)~Treat + PS + Month+  Age + Prior + factor(Cell), data = lungcancer.df,dist="w
#plot(anova(psm.lung),margin=c("chisq","df","P"))
```

To look at the significant difference between the two treatments we can use the `ConvertWeibull` to get the Event Time Ratio[ETR] of our model.

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

```
##                     ETR        LB        UB
## Treat         0.7957083 0.5517117 1.1476131
## PS            1.0305249 1.0208196 1.0403225
## Month         0.9995313 0.9832843 1.0160467
## Age           1.0061178 0.9893916 1.0231268
## Prior         0.9956199 0.9550462 1.0379173
## factor(Cell)2 0.4377162 0.2701043 0.7093387
## factor(Cell)3 0.3221542 0.1944446 0.5337421
## factor(Cell)4 0.6718765 0.4077991 1.1069617
```

Looking at the output and ETR value for Treat, we can conclude that doing the *test* treatment will significantly decrease the survival time by approximately 20%. The scale parameter of our Weibul disitribution is α =0.9281153, this is pretty close to a exponential disitribution which has $\alpha = 1$. This means that a exponential disitribution might be better than our Weibull distribution. But to determine this the best way is to compare the AIC(Akaike's Information Criterion)-values of the two models, and then choose the one with the smallest. Now we will take out the non-significant covariates in our model. This can be done by the p-values found previously, or we can use the function `fastbw` 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 Treatment we get the results shown in the code bellow in '*Factors in Final model*'.

```
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 we use 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")
ConvertWeibull(wei.lung.new,conf.level = 0.95)
```

```
## $vars
##                   Estimate          SE
## lambda          0.01606644 0.009730814
## gamma           1.07504097 0.071234443
## Treat           0.22513731 0.196400967
## PS             -0.03125301 0.005088627
## factor(Cell)2   0.85949730 0.264311692
## factor(Cell)3   1.18229768 0.287012105
## factor(Cell)4   0.41670792 0.276520819
##
## $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
##
## $ETR
##                      ETR        LB        UB
## Treat          0.8110528 0.5681942 1.1577145
## PS             1.0294982 1.0203440 1.0387345
## factor(Cell)2 0.4495528 0.2814600 0.7180336
## factor(Cell)3 0.3329477 0.2031849 0.5455828
## factor(Cell)4 0.6786698 0.4123532 1.1169859
```

Above we have done a new model fitting of a Weibull model using the covariates determined above. If we look at the HR(Hazard ratio)-values which is the risk of death, keeping all covariates constant except the one at interest. From this we find the significance of the covariates on risk of death, which means that $Cell = 3$ is the biggest factor to risk of death. The least significant covariate is PS, HR-value closest to 1 and thereby affects risk the least.

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 a covariate by 1 and see calculate the difference in hazard rate $z(t,\mathbf{x})$. Doing the calculations with the hazard rate function one ends up with equation, with RR being relative risk,

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

with i being the covariate of interest, where in out 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. From this we can see that the highest relative risk is from Cell $= 3$. The risk of death is increased with 26% 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 has a large range, $x_5 \in [10, 90]$, the change is more significant than the 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 $p-$value lower than a significance level of 0.05 is z_2.

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 $\mathbf{x} = \{x_1, x_2, x_3, x_4, x_5, x_1 \cdot x_2, x_1 \cdot x_3, x_1 \cdot x_4\}$ and $\mathbf{x} = \{x_1 + 1, x_2, x_3, x_4, x_5, (x_1 + 1) \cdot x_2, (x_1 + 1) \cdot x_3, (x_1 + 1) \cdot x_4\}$, 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

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

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

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

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 of *Cell*-type 4 should take the *standard* treatment.

f)