Assignment 4

Luis Nicolas Luarte Rodriguez

```
#load libs
if(!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(survival, knitr, ggplot2, knitr, kableExtra, tidyverse, My.stepwise)
```

```
# load the dataset
setwd('/home/nicoluarte/uni/PHD/stat_course')
dataSet <- as_tibble(read.csv('survival65.txt', sep="\t"))
head(dataSet)
```

```
## # A tibble: 6 x 11
##    TIME STATUS LOGBUN   HGB PLATELET   AGE LOGWBC FRACTURE LOGPBM PROTEIN
##   <dbl>  <int>  <dbl> <dbl>    <int> <int>  <dbl>    <int>  <dbl>   <int>
## 1  1.25      1   2.22   9.4        1    67   3.66        1   1.95      12
## 2  1.25      1   1.94  12          1    38   3.99        1   1.95      20
## 3  2         1   1.52   9.8        1    81   3.88        1   2          2
## 4  2         1   1.75  11.3        0    75   3.81        1   1.26       0
## 5  2         1   1.30   5.1        0    57   3.72        1   2          3
## 6  3         1   1.54   6.7        1    46   4.48        0   1.93      12
## # ... with 1 more variable: CALCIUM <int>
```

Run a Cox regression model using all possible predictors. Report and interpret the model

```
coxModel <- coxph(Surv(TIME, STATUS) ~ ., data = dataSet)
summary(coxModel)
```

```
## Call:
## coxph(formula = Surv(TIME, STATUS) ~ ., data = dataSet)
##
##   n= 65, number of events= 48
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## LOGBUN    1.85557   6.39536  0.65628  2.827  0.00469 **
## HGB      -0.12629   0.88136  0.07212 -1.751  0.07994 .
## PLATELET -0.25488   0.77501  0.51194 -0.498  0.61858
## AGE      -0.01306   0.98702  0.01957 -0.668  0.50439
## LOGWBC    0.35389   1.42460  0.71576  0.494  0.62101
## FRACTURE  0.34232   1.40821  0.40725  0.841  0.40059
## LOGPBM    0.38165   1.46470  0.48743  0.783  0.43364
## PROTEIN   0.01302   1.01311  0.02612  0.498  0.61817
## CALCIUM   0.12976   1.13856  0.10502  1.236  0.21659
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## LOGBUN       6.3954     0.1564    1.7670    23.147
## HGB          0.8814     1.1346    0.7652     1.015
## PLATELET     0.7750     1.2903    0.2841     2.114
## AGE          0.9870     1.0131    0.9499     1.026
## LOGWBC       1.4246     0.7020    0.3503     5.794
## FRACTURE     1.4082     0.7101    0.6339     3.128
## LOGPBM       1.4647     0.6827    0.5634     3.808
## PROTEIN      1.0131     0.9871    0.9625     1.066
## CALCIUM      1.1386     0.8783    0.9268     1.399
##
## Concordance= 0.675  (se = 0.051 )
## Likelihood ratio test= 17.62  on 9 df,   p=0.04
## Wald test            = 17.93  on 9 df,   p=0.04
## Score (logrank) test = 18.97  on 9 df,   p=0.03
```

When considering all possible predictors, only one of theses reaches statistical significance 'log blood urea nitrogen at diagnosis' (LOGBUN), 'hemoglobin at diagnosis' (HGB) is significant at a trend level. Considering this, we can interpret, that adjusting for all included variables, at a given point in time, the expected hazard of dying from multiple myeloma relative to a 1 point increase in LOGBUN, given treatment with alkylating agents, is 6.3 times higher. However, the confidence intervals are quite large (1.7 - 23.1) probably due to small sample size or natural variation in the population. Considering the whole model, the survival predictions are modest with a concordance of 0.675, and compared to a 'null' (intercept only model) the complete model is more informative (likelihood ratio test p=0.04).

Select the most relevant predictor using forward selection. Report and interpret the model

```
# Create the null model
coxNull <- coxph(Surv(TIME, STATUS) ~ 1, data = dataSet)
coxForward <- step(coxNull, scope = ~ LOGBUN +
           HGB +
           PLATELET +
           AGE +
           LOGWBC +
           FRACTURE +
           LOGPBM +
           PROTEIN +
           CALCIUM,
       direction = "forward", test = "Chisq")
```

```
## Start:  AIC=308.39
## Surv(TIME, STATUS) ~ 1
##
##             Df    AIC    LRT Pr(>Chi)
## + LOGBUN     1 302.37 8.0153 0.004638 **
## + HGB        1 305.42 4.9678 0.025823 *
## + PLATELET   1 307.69 2.6946 0.100691
## <none>         308.39
## + CALCIUM    1 309.32 1.0703 0.300878
## + FRACTURE   1 309.38 1.0077 0.315465
## + LOGPBM     1 309.76 0.6294 0.427568
```

```
## + LOGWBC    1 309.85 0.5411 0.461992
## + PROTEIN   1 310.26 0.1254 0.723219
## + AGE       1 310.38 0.0147 0.903504
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=302.37
## Surv(TIME, STATUS) ~ LOGBUN
##
##            Df    AIC    LRT Pr(>Chi)
## + HGB       1 300.12 4.2527  0.03919 *
## <none>        302.37
## + PLATELET  1 302.55 1.8213  0.17716
## + CALCIUM   1 303.00 1.3692  0.24196
## + FRACTURE  1 303.23 1.1465  0.28429
## + LOGPBM    1 303.24 1.1376  0.28617
## + AGE       1 303.64 0.7326  0.39205
## + PROTEIN   1 303.89 0.4840  0.48660
## + LOGWBC    1 304.31 0.0662  0.79690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=300.12
## Surv(TIME, STATUS) ~ LOGBUN + HGB
##
##            Df    AIC     LRT Pr(>Chi)
## <none>        300.12
## + CALCIUM   1 300.40 1.72499   0.1891
## + AGE       1 300.76 1.36384   0.2429
## + FRACTURE  1 300.96 1.15840   0.2818
## + LOGPBM    1 301.38 0.74622   0.3877
## + PROTEIN   1 301.51 0.60760   0.4357
## + LOGWBC    1 301.75 0.36666   0.5448
## + PLATELET  1 301.90 0.22377   0.6362
```

```
summary(coxForward)
```

```
## Call:
## coxph(formula = Surv(TIME, STATUS) ~ LOGBUN + HGB, data = dataSet)
##
##   n= 65, number of events= 48
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## LOGBUN  1.71597   5.56209  0.61855  2.774  0.00553 **
## HGB    -0.11966   0.88722  0.05742 -2.084  0.03717 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## LOGBUN    5.5621     0.1798    1.6547   18.6961
## HGB       0.8872     1.1271    0.7928    0.9929
##
## Concordance= 0.675  (se = 0.043 )
## Likelihood ratio test= 12.27  on 2 df,   p=0.002
## Wald test            = 12.51  on 2 df,   p=0.002
```

```
## Score (logrank) test = 13.07  on 2 df,   p=0.001
```

For the forward model selection I used the Akaike information criterion as it provides an estimate of the goodness of fit, while considering the number of parameters, thus penalizing overfitting. The model selection algorithm first selected LOGBUN, THEN HGB, and tested if increasing the number of variables improved the model. Under AIC or likelihood ratio test, adding more variables increased the amount of information lost, and was not significantly different from the null model, respectively. LOGBUN, in this model, estimated an expected hazard of dying of 5.5 times higher (the full model gave 6.3), whereas HGB (hemoglobin at diagnosis), which got only to trend level in the full model, now estimates that a single point increase in hemoglobin reduces the hazard by a factor of 0.88. Thus, higher hemoglobin levels and lower blood urea nitrogen might be good indicator for myeloma.

```
# Create the full model
coxFull <- coxph(Surv(TIME, STATUS) ~ ., data = dataSet)
coxBackward <- step(coxFull, direction = "backward", test = "Chisq")
```

```
## Start:  AIC=308.77
## Surv(TIME, STATUS) ~ LOGBUN + HGB + PLATELET + AGE + LOGWBC +
##     FRACTURE + LOGPBM + PROTEIN + CALCIUM
##
##            Df    AIC    LRT Pr(>Chi)
## - LOGWBC    1 307.01 0.2401 0.624163
## - PROTEIN   1 307.02 0.2422 0.622597
## - PLATELET  1 307.02 0.2431 0.621947
## - AGE       1 307.22 0.4432 0.505562
## - LOGPBM    1 307.40 0.6291 0.427681
## - FRACTURE  1 307.52 0.7505 0.386322
## - CALCIUM   1 308.21 1.4312 0.231574
## <none>        308.77
## - HGB       1 309.84 3.0619 0.080148 .
## - LOGBUN    1 314.32 7.5426 0.006026 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=307.01
## Surv(TIME, STATUS) ~ LOGBUN + HGB + PLATELET + AGE + FRACTURE +
##     LOGPBM + PROTEIN + CALCIUM
##
##            Df    AIC    LRT Pr(>Chi)
## - PROTEIN   1 305.17 0.1517 0.696888
## - PLATELET  1 305.17 0.1566 0.692276
## - AGE       1 305.69 0.6788 0.409998
## - LOGPBM    1 305.73 0.7145 0.397955
## - FRACTURE  1 305.75 0.7401 0.389640
## - CALCIUM   1 306.49 1.4738 0.224748
## <none>        307.01
## - HGB       1 308.05 3.0395 0.081259 .
## - LOGBUN    1 313.75 8.7402 0.003113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=305.17
## Surv(TIME, STATUS) ~ LOGBUN + HGB + PLATELET + AGE + FRACTURE +
##     LOGPBM + CALCIUM
##
```

```
##            Df    AIC    LRT Pr(>Chi)
## - PLATELET  1 303.36 0.1948 0.658944
## - FRACTURE  1 303.79 0.6254 0.429060
## - LOGPBM    1 303.89 0.7264 0.394064
## - AGE       1 304.44 1.2683 0.260080
## - CALCIUM   1 304.77 1.6067 0.204963
## <none>        305.17
## - HGB       1 306.12 2.9493 0.085917 .
## - LOGBUN    1 311.83 8.6605 0.003252 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=303.36
## Surv(TIME, STATUS) ~ LOGBUN + HGB + AGE + FRACTURE + LOGPBM +
##     CALCIUM
##
##            Df    AIC    LRT Pr(>Chi)
## - LOGPBM    1 301.94 0.5797 0.446415
## - FRACTURE  1 302.03 0.6665 0.414262
## - AGE       1 302.51 1.1513 0.283287
## - CALCIUM   1 303.34 1.9761 0.159803
## <none>        303.36
## - HGB       1 306.35 4.9929 0.025452 *
## - LOGBUN    1 309.91 8.5456 0.003464 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=301.94
## Surv(TIME, STATUS) ~ LOGBUN + HGB + AGE + FRACTURE + CALCIUM
##
##            Df    AIC    LRT Pr(>Chi)
## - FRACTURE  1 300.61 0.6731 0.411987
## - AGE       1 301.48 1.5412 0.214442
## - CALCIUM   1 301.78 1.8352 0.175514
## <none>        301.94
## - HGB       1 305.24 5.2969 0.021364 *
## - LOGBUN    1 308.23 8.2881 0.003991 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=300.61
## Surv(TIME, STATUS) ~ LOGBUN + HGB + AGE + CALCIUM
##
##           Df    AIC    LRT Pr(>Chi)
## - AGE      1 300.40 1.7826 0.181832
## <none>       300.61
## - CALCIUM  1 300.76 2.1437 0.143153
## - HGB      1 304.12 5.5029 0.018985 *
## - LOGBUN   1 306.63 8.0127 0.004645 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=300.4
## Surv(TIME, STATUS) ~ LOGBUN + HGB + CALCIUM
```

```
##
##           Df    AIC    LRT Pr(>Chi)
## - CALCIUM  1 300.12 1.7250 0.189052
## <none>       300.40
## - HGB      1 303.00 4.6086 0.031813 *
## - LOGBUN   1 305.18 6.7848 0.009194 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=300.12
## Surv(TIME, STATUS) ~ LOGBUN + HGB
##
##          Df    AIC    LRT Pr(>Chi)
## <none>      300.12
## - HGB     1 302.37 4.2527 0.039188 *
## - LOGBUN  1 305.42 7.3002 0.006895 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(coxBackward)
```

```
## Call:
## coxph(formula = Surv(TIME, STATUS) ~ LOGBUN + HGB, data = dataSet)
##
##   n= 65, number of events= 48
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## LOGBUN  1.71597   5.56209  0.61855  2.774  0.00553 **
## HGB    -0.11966   0.88722  0.05742 -2.084  0.03717 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## LOGBUN    5.5621     0.1798    1.6547   18.6961
## HGB       0.8872     1.1271    0.7928    0.9929
##
## Concordance= 0.675  (se = 0.043 )
## Likelihood ratio test= 12.27  on 2 df,   p=0.002
## Wald test            = 12.51  on 2 df,   p=0.002
## Score (logrank) test = 13.07  on 2 df,   p=0.001
```

The backward selection algorithm converged into the same model, so interpretations are identical. Additionally, the concordance is similar to the full model, but with less standard error 0.043 vs 0.051.