Cox model: introduction

Cox proportional hazards regression models are used in survival analysis to examine the effect of several variables on the time a specified event takes to happen. The key feature of these models is the assumption that the hazard ratios between individuals are constant over time, which is known as the proportional hazards assumption. The Cox model does not require the baseline hazard function to be specified, allowing for more flexibility. The Cox proportional hazards model is:

$$h(t \mid X) = h_0(t) exp(\beta_1 X_1 1 + \beta_2 X_2 + \ldots + \beta_p X_p)$$

where $h(t \mid X)$ is the hazard function at time t given covariates X, $h_0(t)$ is the baseline hazard function at time t.

'Veteran' dataset

veteran: dataset of 137 observations x 8 variables form a two-treatment randomized trial for lung cancer.

trt: 1=standard 2=test
celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
time: survival time
status: censoring status
karno: Karnofsky performance score (100=good)
diagtime: months from diagnosis to randomisation
ageA: in years
prior: prior therapy 0=no, 10=yes

```
> head(veteran)
  trt celltype time status karno diagtime age prior
1   1 squamous   72      1    60        7  69     0
2   1 squamous  411      1    70        5  64    10
3   1 squamous  228      1    60        3  38     0
4   1 squamous  126      1    60        9  63    10
5   1 squamous  118      1    70       11  65    10
6   1 squamous   10      1    20        5  49     0
```

Summary

Here below the minimal code to run a Cox regression in R.

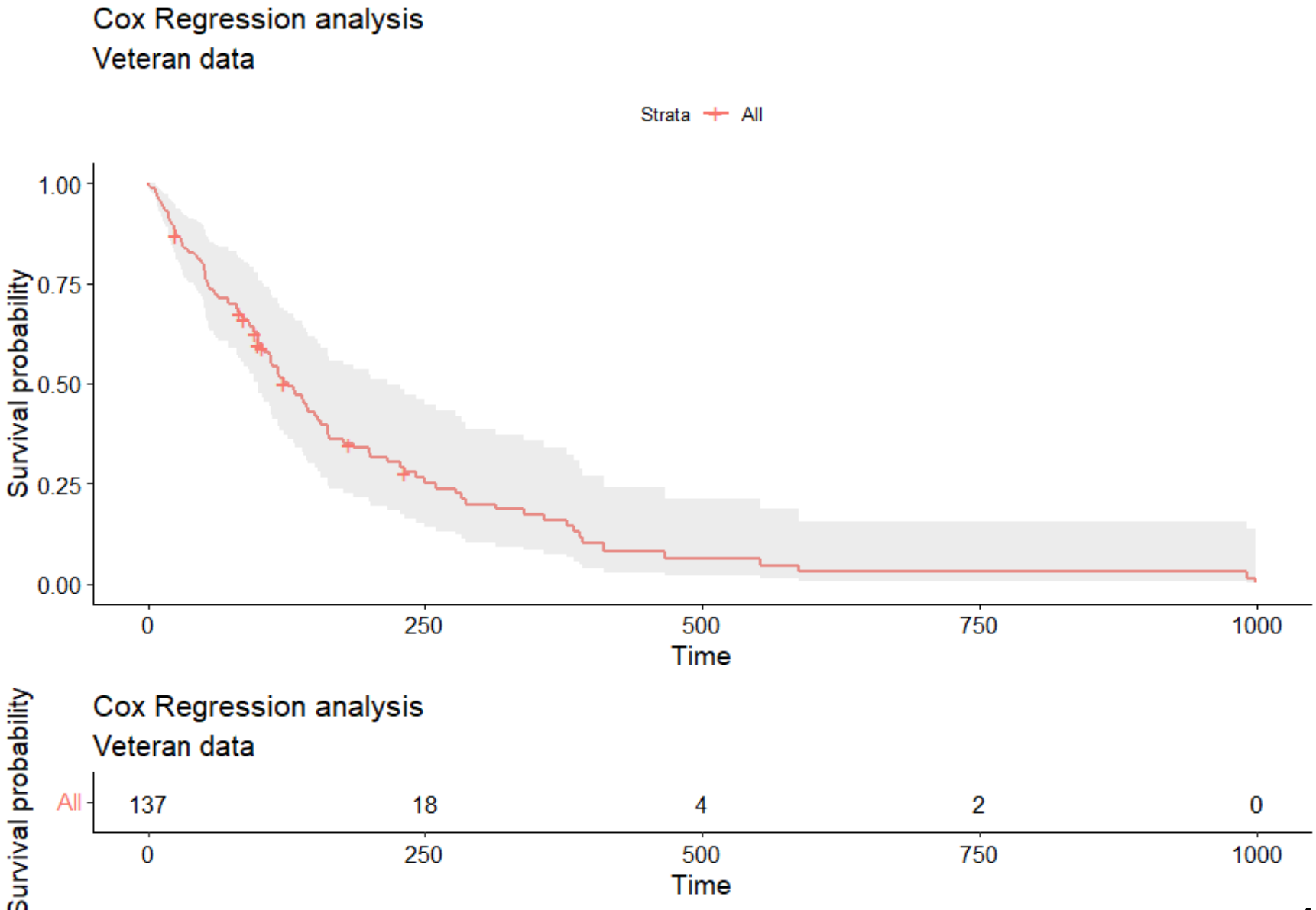
```
> # fit the Cox proportional hazards model
> cox_model = coxph(Surv(time, status) ~ age + celltype + trt + karno, data=
      veteran)
> summary(cox_model)
Call:
coxph(formula = Surv(time, status) ~ age + celltype + trt + karno,
    data = veteran)

  n= 137, number of events= 128

                       coef exp(coef)  se(coef)      z Pr(>|z|)
age                -0.008903  0.991136  0.009224 -0.965   0.3345
celltypesmallcell   0.856340  2.354528  0.271322  3.156   0.0016 **
celltypeadeno       1.178807  3.250494  0.296440  3.977 6.99e-05 ***
celltypelarge       0.402332  1.495308  0.282544  1.424   0.1545
trt                 0.303048  1.353980  0.205656  1.474   0.1406
karno              -0.032685  0.967843  0.005409 -6.043 1.51e-09 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
       1

Concordance= 0.738  (se = 0.021 )
Likelihood ratio test= 61.98  on 6 df,   p=2e-11
Wald test            = 62.35  on 6 df,   p=1e-11
Score (logrank) test = 66.62  on 6 df,   p=2e-12
```

Plot of the survival function



Main observations

- The coefficient for age is positive and statistically significant, indicating that older age is associated with an increased hazard (higher risk of death).
- Different cell types (squamous, small cell, adeno, large) significantly affect the hazard rates, with each cell type showing a varying impact on survival times.
- The treatment variable (trt) has a significant coefficient, suggesting that the type of treatment administered impacts the hazard, with one treatment possibly leading to better survival outcomes compared to the other.
- The Karnofsky performance score (karno), which measures the patient's general well-being and ability to perform daily activities, is a significant predictor, with higher scores associated with lower hazard rates (better survival).

References

Applied Survival Analysis Using R, Dirk F. Moore, 2016, Springer, ISBN 978-3-319-31245-3 (e-book)

The R Project for Statistical Computing:
https://www.r-project.org/