Preliminary solutions Inger Persson
Analysis of survival data 2025-01-17

(6) **Task 1**

(3) **A** i) Censoring and/or truncation affect the likelihood of the event of interest
ii) Different statistical methods are appropriate for different types of censoring and/or truncation

(3) **B** Parametric methods require that you know the distribution of time-to-event. If you have a large sample you can estimate the distribution based on the events, but your conclusions will rely on the correct choice of distribution. Since we rarely know the distribution nonparametric/semiparametric methods are a robust choice.

(14) **Task 2**

A (2) Users who installed and stopped using the app before January 1, 2019, are not included in the dataset because they are unobservable. For users who installed the app before January 1, we condition on them still using the app on January 1. This means that **left truncation** is represented in this study.

B (8)

i) The user was still active as of December 2024, the event thus hasn’t been observed yet. This is a case of **right censoring**.

ii) The event (user inactivity) was observed after 17 months, which means that there is **no censoring**.

iii) **Right censoring** again, since we haven’t observed the event at the end of the study. (This observation is also from a truncated distribution, which will be adjusted for in the likelihood below)

iv) We know that the event happened sometime between August and October, **interval censoring** is thus present here.

C (4) Users who installed the app before before January 1, 2019 should be considered being sampled from a truncated distribution.

i) We have a right censored observation at 21 months, from a non-truncated distribution.

ii) This is an event at 16 months, from a non-truncated distribution.

iii) A right censored observation at 74 months (6 years plus 2 months), from a truncated distribution. We have to condition on using the app at least 3 months (from October to January).

iv) An interval censored observation at somewhere between 13 and 15 months, from a non-truncated distribution.

L = S (21) · f(16) · S (74)/S(3) · {S(13) – S(15)}

(26) **Task 3**

A (8) High performance:
25th percentile = 6 years. This means that 25% of the players who played in high performance teams have retired within 6 years from the start of their career.

Median = 10 years. This means that 50% of the players who played in high performance teams have retired within 10 years from the start of their career.

75th percentile = at least 18 years. The 75th percentile cannot be estimated since less than 75% of the players who played in high performance teams had not retired during the study. The 75th percentile must however be at least 18 years, which is the longest observed time in this group.
.

Low performance:
25th percentile = 4 years. This means that 25% of the players who played in high performance teams have retired within 4 years from the start of their career.

Median = 7 years. This means that 50% of the players who played in high performance teams have retired within 7 years from the start of their career.

75th percentile= 13 years. This means that 75% of the patients who received the new treatment have had a remission within 13 years from the start of their career.

These measures all have larger values for players in high performance teams, which means that the career longevity is higher for players in high performance teams.

B (3) Approximate 10-year probability of still playing can be seen from the table (high performance group) or the Kaplan-Meier curve (low performance group).
For players in high performance teams: **47.78%**
For players in high performance teams: **approx. 30%** (not shown in table)

The probability of still playing 10 years after career start is a lot higher for players in high performance teams, almost 20 percentage points higher.

C (2) The cumulative hazard for players in high performance teams is approximately 0.75, which means that a player in these teams is expected to retire 0.75 times in the first 10 years of the career (if events were repeatable).

D (13) $H_0: \ h_{high}(t) = \ h_{medium}(t) \ = h_{low}(t)$ for all $t \leq \tau$

H_a: $h\,(t)$ differ between some of the groups for some $t \leq \tau$

Where τ = largest time at which both groups have at least one subject at risk.

Test: Log-rank or Gehan's/Wilcoxon test. Motivation needed.

Assumptions:

- Random sample – OK (specified in the task)
- Independent samples – reasonable to assume that career longevity in the three compared groups are independent
- Non-informative/random censoring (i.e. that the censoring times are not related to the later, unknown, retirement times) – not specified in the task. Should be discussed with field experts, but at least the censoring plot doesn't contradict this since the censoring pattern is similar in all three groups.
- Right censored data – OK (information provided in task)
- Survival probabilities are the same for subjects recruited early and late in the study - not specified in the task. Should be discussed with a field expert, but reasonable to assume.
- Large samples (both tests are based on large-sample approximations to the distribution of the chi-square statistics) – OK, 356 events in the smallest group.

Choice of significance level:

Wrongly rejecting the null hypothesis here would mean that we claim that there is a difference in time to retirement between the three team performance groups, when in fact there is no difference. There are hardly any serious consequences from this, 5% is a reasonable choice.

Result:

P-value = <0.0001 (for both Log-rank and Gehan/Wilcoxon).

The *P*-value is smaller than the chosen α, thus H_0 is rejected.

Conclusion:

The test suggests that there is a significant difference in career longevity between high/medium/low performance teams in the investigated population of basketball players, where players in high performance team in general retire later than the others.

(30) **Task 4**

A (3) The variables *team_performance, nationality,* and *position* have to be recoded to 0/1 variables, or denoted as “class” variables in proc phreg.

The Martingale plot suggests that *age* can be used as a continuous covariate (and *age_group* is thus not needed).

B (17) Any chosen model would have to fulfill the **assumptions** for the Cox model. Many of them are already discussed in Task 2 (OK to refer to that):
• Random sample (for inference to be correct) - OK.
• Non-informative censoring. Not contradicted.
• Right-censored or left truncated data. Right censoring can be seen from task.

New for this task:
• Large sample (common rule of thumb: ≥ 10 events per covariate). We have a total of 2500 observations, and 1659 events. This is sufficient for any model you’ve chosen.
• Proportional hazards (to be checked when building the model)

Check of **proportionality (PH) assumption**:

1) include time-dependent covariate in model
To test the assumption of proportional hazards you can include the time-dependent covariate ln(t)*covariate in the model (if significant, the PH assumption is rejected)

H_0: The hazards for different values of covariate i are proportional
(all *i*=1 to *p* covariates are to be examined)
H_a: The hazards are not proportional

Significance level α = 5% fine to use (no serious consequences if we claim that the hazards are not proportional when they in fact are)

The test above is rejected for the *age* and *team performance* covariates, but fine (non-significant) for *nationality* and *position*.

2) Arjas plots
You should always make use of a graphical method in addition to the test above. Arjas plots have been provided, which show that the PH assumption is violated for *team performance* and *age_group* (the curve for one of the age groups crosses the line), but looks okay for *nationality* and *position*.

This concludes that the PH assumption doesn’t hold for *age* and *team performance*.

Comparison of AIC values
Model 6 has the lowest AIC value, which suggests that this model is to be used.

Test of equality of strata for Model 6:
For Model 6 to be valid, we need to check that it is reasonable to assume that the regression coefficients are the same in each stratum.

H_0 : All β's are the same for all s strata
H_a : At least one of the β 's is/are different

This can be tested, using the Likelihood ratio test.

Significance level α=5% fine to use (no serious consequences if we claim that the covariates are not the same when they in fact are, then we estimate separate models instead)

Test statistic:

$$-2\left[LL(\mathbf{b})-\sum_{j=1}^{s}LL_j(\mathbf{b}_j)\right] \sim \chi^2_{(s-1)p}$$

where s= no. of strata and p=no. of covariates

s = 9, p = 3

$-2LL(\mathbf{b})$ = 10255.056
$LL(\mathbf{b})$ = -5127.528

$-2LL(\mathbf{b}_{18\text{-}23,\ \text{high}})$ = 939.276
$-2LL(\mathbf{b}_{18\text{-}23,\ \text{low}})$ = 634.057
$-2LL(\mathbf{b}_{18\text{-}23,\ \text{medium}})$ = 1459.442
$-2LL(\mathbf{b}_{24\text{-}30,\ \text{high}})$ = 1064.305
$-2LL(\mathbf{b}_{24\text{-}30,\ \text{low}})$ = 824.604
$-2LL(\mathbf{b}_{24\text{-}30,\ \text{medium}})$ = 2006.674
$-2LL(\mathbf{b}_{31\text{-}35,\ \text{high}})$ = 889.155
$-2LL(\mathbf{b}_{31\text{-}35,\ \text{low}})$ = 626.143
$-2LL(\mathbf{b}_{31\text{-}35,\ \text{medium}})$ = 1779.885

Sum $-2LL_j(\mathbf{b}_j)$ = 10223.541
Sum $2LLj(\mathbf{b}j)$ = -5111.7705

Test statistic = 10255.056-10223.541 =
= -2(-5127.528+-5111.7705) =
= 31.515
df = (9-1)*3 = 24 df

According to Table c.2 the corresponding p-value is larger than 0.05, which means that the null hypothesis is not rejected (reject if χ^2_{test} larger than χ^2_{crit} = 36.41503).

Conclusion:
It is okay to assume that the regression coefficients are the same in each of the two strata and the stratified model can be used.

Cox-Snell plots
All models provide Cox-snell plots that suggest that the model fits the data.

Choice of model:
All of the above suggests that **Model 6** is a good choice.

C (8) The marginal effects of the covariates are presented below, i.e. the effect of each covariate holding the other covariates constant.

International players have a 0.9% lower risk of retirement on average, compared to domestic players (hazard ratio 0.991, 95% confidence interval 0.892 to 1.100). This estimate is however not significant.

Compared to players in the Center position, Forwards have a 0.5% lower risk of retirement on average, (hazard ratio 0.995, 95% confidence interval 0.855 to 1.164), and Guards have a 0.5% higher risk of retirement on average (hazard ratio 1.005, 95% confidence interval 0.869 to 1.167). This covariate is also not significant (as seen by the Type 3 Tests).

Age_group and team*_performance* cannot be interpreted in terms of hazard ratios, since the model is stratified on these variables, but the estimated survival plot shows that the risk of retirement is lower (the "survival" is higher) for younger players than for older players, and for players in high performance teams compared to medium or low performance teams.

D (2) Generalized $R^2 = 1 - e^{-(LRT/n)}$

where $LRT = -2\log L(0)-[-2\log L(p)] = 10255.114 – 10255.056 = 0.058$

$R^2 = 1 – \exp(-0.058/2500) = 1 - 0.99998 = 0.0000232$

According to the generalized R^2 the model shows practically no association between the covariates and time to retirement.

(4) **Task 5**

As always, it is a good thing to explain to your employer that you are following ethics codes for statisticians.

As long as you have given a few examples of part(s) of one or more of the code of ethics document(s) that are applicable in this situation, and told which document you are referring to, you'll get points for this task.