Survival Analysis of VA Lung Cancer Dataset

Ali Saadat Varnosfaderanii

Introduction

The goal of the project is to perform survival analysis on the data from 137 advanced lung cancer patients. Patients were randomized according to one of two chemotherapeutic agents (standard vs test). Of particular interest is the possible differential effects of therapy on tumor cell type. Tumors are classified into one of four broad groups (squamous, small, adeno, and large). Covariates are:

1. Karnofsky performance which is an indicator of patients' general well-being; it ranges from 0-100 and larger numbers show better medical status.
2. Time from diagnosis to the start of study (in month).
3. Age in years.
4. Previous therapy.

Explanatory Data Analysis (EDA)

Table 1 represents the summary statistics of continuous variables.

Table 1: Summary statistics for the numeric variables

	N	Mean	SD	Min	Q1	Median	Q3	Max
time	137	121.63	157.82	1	25	80	144	999
karno	137	58.57	20.04	10	40	60	75	99
diagtime	137	8.77	10.61	1	3	5	11	87
age	137	58.31	10.54	34	51	62	66	81

All EDA plots (univariate and bivariate) is summarized in figure 1, which is colored based on treatment group. We can draw the following conclusions from figure 1:

1. Two treatment groups have approximately equal number of subjects.
2. There is no clear difference between the distribution of variables in case vs control.
3. The *time* variable, which represents the event time, is right-skewed.

4. There is a moderate correlation (≈ 0.4) between *karno* and *time* variables.
5. 30% of the subjects had a prior treatment.
6. Age ranges from 34 to 81 years old.

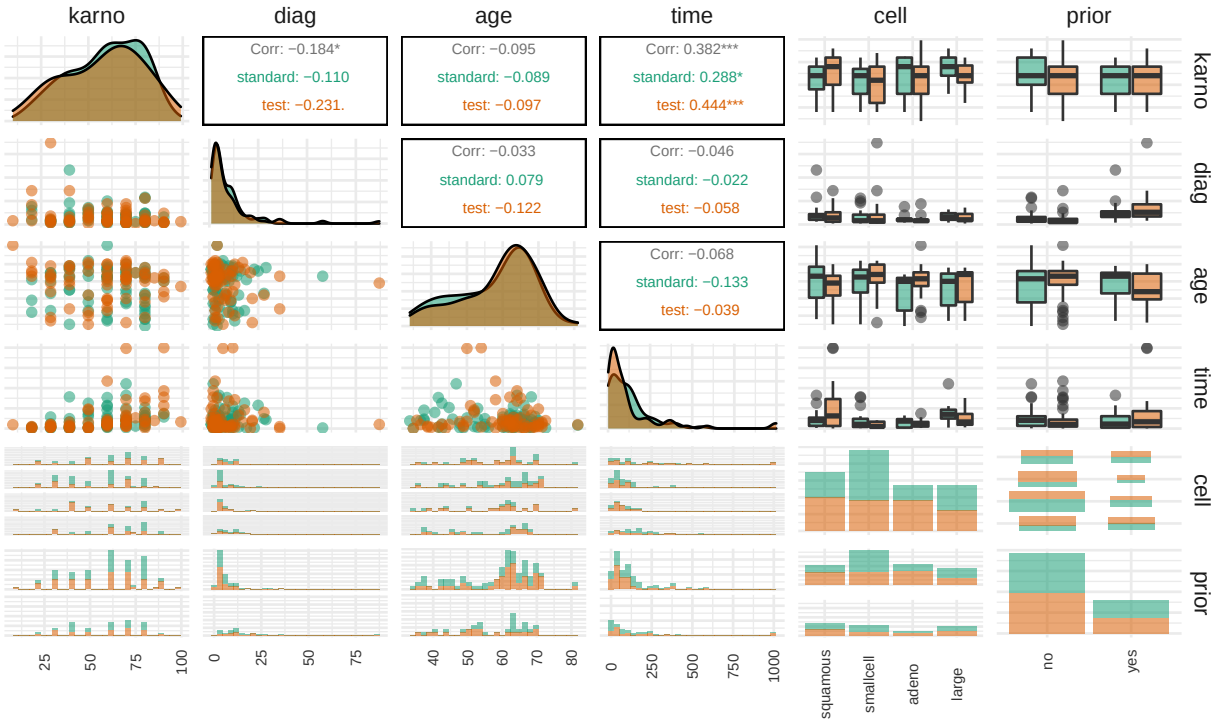

Figure 1: Pairs plot for all the variables

Model Fitting: Kaplan-Meier (KM) Estimator and Log-rank Test

Kaplan-Meier (KM) Estimator, also known as the product limit estimator, is a non-parametric method used to estimate the survival function. It is defined as the probability of surviving in a given length of time while considering time in many small intervals. KM estimator works as follow:

1. compute the probabilities of occurrence of event at a certain point of time
2. multiply this probability by all the probabilities in the earlier intervals

Mathematically, we can write KM estimator as

$$\hat{S}(t) = \prod_{j:t(j)\leq t} (1 - \frac{d_j}{r_j})$$

where r_j is the number of individuals at risk just before t(j) (including censored individuals at t(j)), and d_j is the number of individuals experiencing the event at time t(j).

Logrank test compares survival curves (estimated with KM) of the two groups. It computes expected number of death for each unique death time in the data, assuming that the chance of dying for subjects at risk is the same for each group. The test compares the observed number of deaths in each group to the expected number using χ^2 test with the statistic: $\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$. The null hypothesis is that there is no difference between the groups in the probability of an event at any time point; in other words $H_{NULL} : S_{control}(t) = S_{case}(t)$ and $H_{ALT} : S_{control}(t) \neq S_{case}(t)$

To explore the impact of treatment on survival, we estimate the the survival probability of patients stratified by their treatments using KM estimator(figure 2 left). To compare the survival curves, logrank test is performed. $Pvalue = 0.93$ ($Chisq = 0$ on 1 degrees of freedom) in the left corner is very high, and we cannot reject the null hypothesis. Thus, we cannot conclude that treatment has a significant effect on survival.

Another variable to explore is the cell type. Figure 2 right shows how survival probability can differ based on the cell type. We observe that *squamous* and *large* cell types have a higher survival probability compared to *small* and *adenocarcioma* cells.

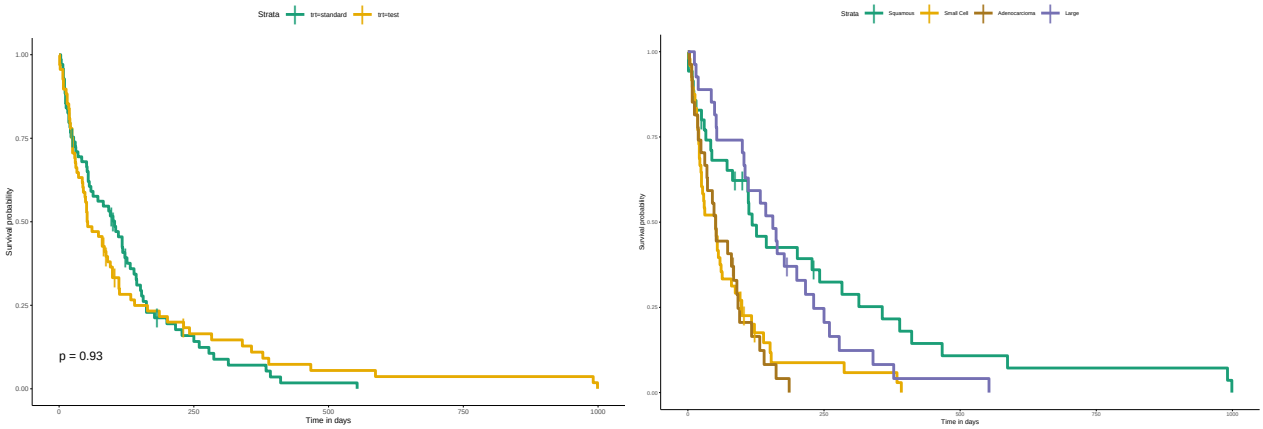

Figure 2: left: Survival stratified by treatment, right: Survival stratified by cell-type

We perform pairwise comparison between cell types; after Benjamini-Hochberg adjustment, there are significance differences between *small-large*, *small-squamous*, *adeno-large*, and *adeno-squamous* (table 2).

Table 2: Pairwise log-rank test between celltypes; each value represents adjusted pvalue

	squamous	smallcell	adeno
smallcell	0.001	-	-
adeno	0.001	0.756	-
large	0.437	0.003	0

Model Fitting: Cox Propotional Hazard (Cox-PH)

Cox-ph model is fitted using maximum partial likelihood. To find the best model, we to perform an exhaustive search (without interaction) so we can select a simple and informative model by choosing a model with minimum AIC. To begin with, we confirm that there is no multicollinearity between variables ($VIF < 2.2$ for all the variable). Then, we fit a cox-ph model with all the variables included. We can see that PH assumption is violated by *celltype* and *karno* variables (table 3):

Table 3: Schoenfeld's test for model with all variables

	chisq	df	p
celltype	15.23	3.00	0.00
trt	0.26	1.00	0.61
karno	12.94	1.00	0.00
prior	2.17	1.00	0.14
age	1.83	1.00	0.18
diagtime	0.01	1.00	0.91
GLOBAL	34.55	8.00	0.00

To solve the PH-violation:

1. we stratify the baseline hazard according to the *celltype* (figure 3).
2. we split the time into smaller intervals (split points are 90 and 180 days) and calculate *karno* coefficient in each interval.

With these actions, Schoenfeld's global test is no longer significant (table 4):

Table 4: Schoenfeld's test after solving PH violations

	chisq	df	p
trt	0.38	1.00	0.54
prior	2.32	1.00	0.13
age	4.21	1.00	0.04
diagtime	0.49	1.00	0.48
karno:strata(tgroup)	2.27	3.00	0.52
GLOBAL	12.19	7.00	0.09

After handling the PH-violation, we calculate AIC (using *extractAIC* function) for all the possible models (combination of variables without interaction). We observe that the lowest AIC (=628.4) belongs to a cox-model with *celltype* and *karno* as variables. The maximum AIC (=1015.9) belongs to the model with *treatment*, *prior_treatment*, *age*, and *diagnos_time*. Therefore, we conclude that only *celltype* and *karno* are the important variables.

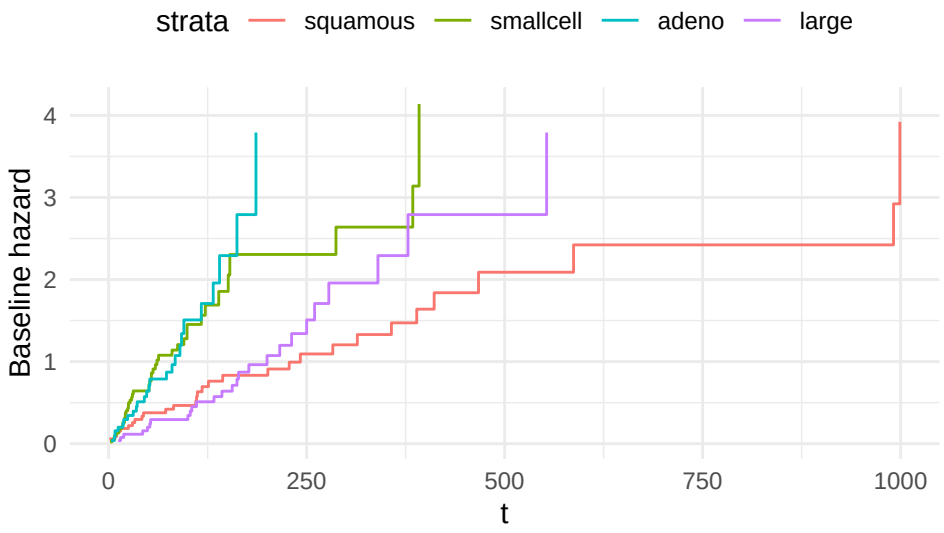

Figure 3: Baseline hazard stratification by cell-types

Cox-PH Final Model

The final model is $\hat{\mathbf{h}}(\mathbf{t}) = \mathbf{h_0}(\mathbf{t}, \mathbf{celltype})\mathbf{exp}(\beta_{\mathbf{karno}}(\mathbf{t}) * \mathbf{x_{karno}})$ where:

- $h_o(t, celltype)$ being the baseline hazard specific to a cell type from figure 3. We observe that for *squamous* and *large* cell types, the baseline hazard increases slowly, but for *small* and *adeno* cell types it increases sharply.
- $\beta_{karno}(t)$ depends on the time because we split the time into three intervals to solve PH-violations:

$$\beta_{karno}(t) = \begin{cases} -0.05, & t \in [0, 90) \\ 0.01, & t \in [90, 180) \\ -0.02, & t \geq 180 \end{cases}$$

To get a deeper insight into $\beta_{karno}(t)$, we plot the hazard-ratio ($HR = exp(\beta)$) in figure 4.

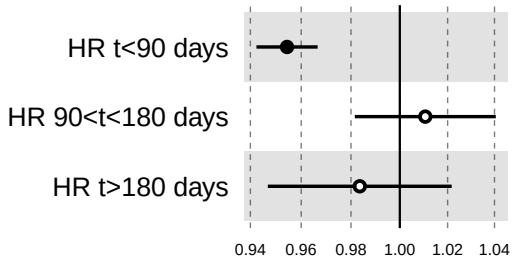

Figure 4: Hazard ratio for karno variable with 95% CI

We observe that *karno* score in the first interval ($t < 90$ days) is significantly associated with decreased risk of death. But we don not see such an association for interval 2 and 3 ($t >= 90$ days). This suggests that the impact of *karno* is limited to the first three months. That being said, it must be noted that in acute illnesses, any measure that is over six months is usually no longer relevant. Furthermore, many of the patients with low *karno* score have been lost; at the beginning, 28% of the patients have $karno < 40$, but after six month, only 4% of them remained in the study.

Model Assessment

There are three important assumptions for cox PH that must be assessed:

1. **Proportional hazard assumption**, which means that the relative hazard remains constant over time with different predictors. We check this assumption using **Schoenfeld** residuals. Schenofel residual is the difference between the observed covariate and the expected. In principle, Schoenfeld residuals are independent of time, so a non-random pattern is indication of PH-violation. We do not observe such a pattern in figure 5 top-left, so PH assumption is valid. (Note: PH assumption was violated in our first model which included all the variables, and we solved this issue by stratifying the baseline hazard according to the *celltype* and splitting the time into smaller intervals. There are other methods to solve the violation of PH-assumption. For example, adding time-interaction terms might help to handle the PH-violation; however it is often not straightforward to interpret the interaction terms. Another method is to fit a time-varying coefficient model which is more complex and beyond the scope of this project)

2. **There is no influential observation or outlier.** To detect outliers, **Deviance** residuals is utilized. The Deviance residual is the measure of deviance contributed from each observation. The idea behind it is to examine the difference between the log-likelihood for subject i under a given model and the maximum possible log-likelihood for that subject. We visualize deviance residuals (figure 5 top-right). It is clear that deviance residuals are fairly symmetric around zero, meaning that there is no outlier.

3. **All continuous covariates in the model must have a linear form.** To validate this assumption, we use **Martingale** residuals, which are the discrepancy between the observed value of a subject's failure indicator and its expected value integrated over the time for which that patient was at risk. Martingle residual ranges from $-\infty$ to 1; negative values are assigned to subjects that lived longer than expected and positive values belong to subjects that died sooner than expected by model. We plot the Martingale residuals against continuous covariates (here *karno*) to detect nonlinearity. For a given continuous covariate, patterns in the plot may suggest that the variable is not properly fit. In figure 5 bottom-left no pattern is observed, so this assumption is valid.

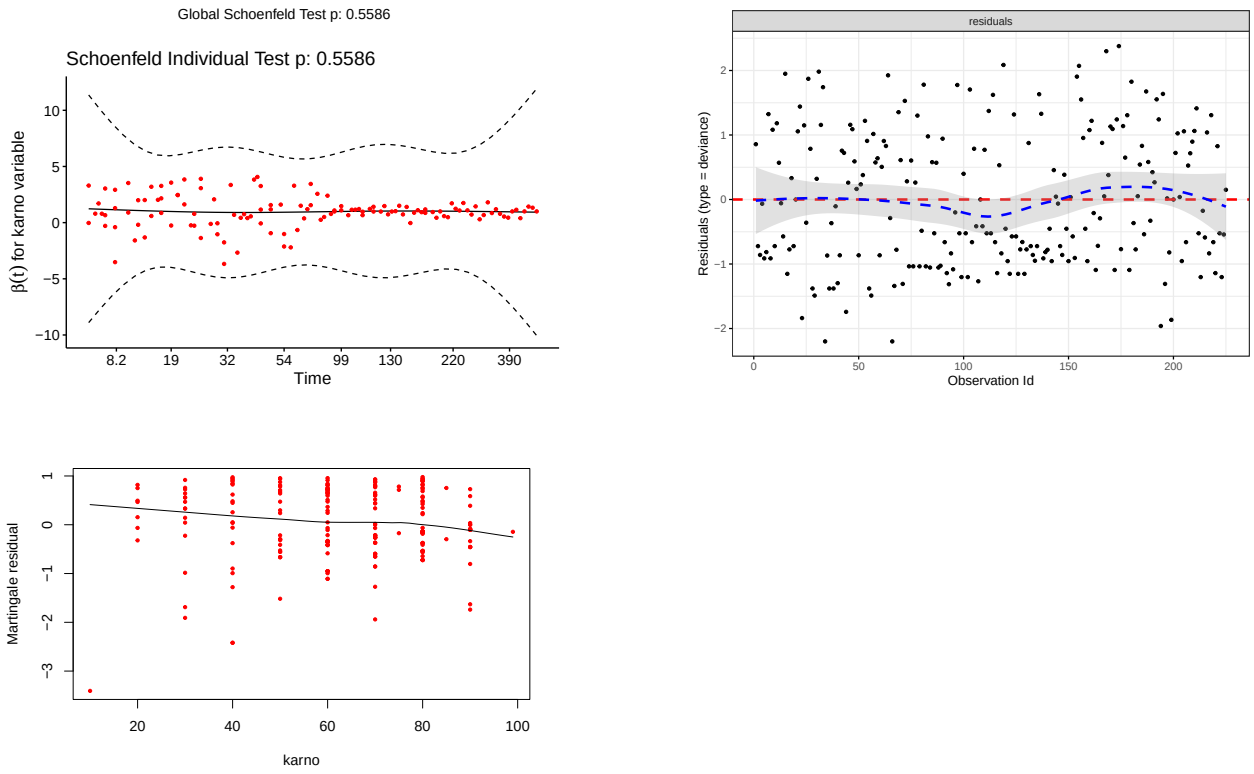

Figure 5: Residuals (top-left: Schenofel, top-right: Deviance, bottom-left: Martingale)

Conclusion

In this project, the survival analysis of VA lung cancer dataset was performed. In summary:

- the survival probability was calculated using Kaplan-Meier estimator
- the survival curves were compared using log-rank test
- an exhaustive search was performed to find the best cox-ph model
- cox-ph model was fitted using maximum partial likelihood
- the cox-ph assumptions was verified by plotting various residuals

The results suggest that treatment with new chemotherapeutic agent has no significant effect on the survival probability of patients. One of the most critical variables that impacts the patients outcome is tumor cell type, as we observed that survival curves differs significantly based on cell type. It was shown that patients with *squamous* and *large* cell types have a higher survival probability compared to patients with *small* and *adenocarcioma* cells. Furthermore, Karnofsky performance is strongly associated with the survival outcome, but its effect is limited to the first three months, so this variable should not be used as a predictor after three months of diagnosis. From a transnational perspective, outcomes of this project could include novel disease bio-markers, better prediction models, or innovative targets for diagnostic or therapeutic development; however, the sample size is small and further investigation is required to make a conclusive statement.