M2 Data Science: Health, Insurance, Finance
Survival and Longitudinal Data Analysis
Final Project

Survival Analysis of Mortality of HIV Patients

Student: CHEAM Richard

Lecturer: Dr. PARK Juhyun

December, 2024

Université d'Évry | Paris-Saclay

Contents

1 Introduction

In the context of HIV (human immunodeficiency virus) research, quantity of T-lymphocytes, called CD4 cells, is a critical biomarker for monitoring the disease progression. A decrease in CD4 levels often signals a decline in health and can indicate the progression to AIDS (acquired immunodeficiency syndrome).

In a longitudinal study on HIV-infected patients who either failed or were intolerant to AZT (zidovudine) therapy, participants were randomly assigned to receive one of two alternative antiretroviral drugs: didanosine (ddI) or zalcitabine (ddC). The goal was to compare the efficacy and safety of these alternative treatments.

In this report, readers will gain insight into the survival analysis of mortality among HIV patients in the above study by using various statistical approaches to address research questions of interest such as the factors that influence patient survival, the difference in treatment effect, the link between CD4 cells and the risk of death, and so on.

2 The dataset

For this project, the dataset was provided by Dr. PARK Juhyun, containing information of 467 patients. There are 1,405 observations over time and 9 variables without any missing values, duplicate rows, and inconsistent data (entry errors). In the study, CD4 cells count were measured from the beginning then at 2, 6, 12, and 18 months thereafter respectively.

Numerical covariates are:

- `subject`: patients identifier
- `time`: time to death or censoring
- `cd4`: CD4 cells count
- `time_obs`: time points at which the CD4 cells count was recorded

Categorical covariates are:

- `death`: life status
- `treatment`: type of treatment
- `sex`: gender of the patient
- `prev_infection`: previous opportunistic infection of AIDS
- `azt`: reaction to AZT therapy

Information of patients by each group can be confirmed with Table 3. Each categorical variable has two levels (see the x-axis of boxplot in Figure 1).

3 Exploratory data analysis (EDA)

3.1 Numerical analysis

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
`time`	0.47	12.23	14.07	13.89	17.00	21.40
`cd4`	0.000	3.162	5.477	7.023	10.440	24.125

Table 1: Descriptive statistics of `time` and `cd4`

From Table 1, it can be seen that:

- Survival `time` ranges from a minimum of 0.47 to a maximum of 21.40 with a median and a mean close to each other, suggesting a relatively symmetric distribution without extreme outliers. Plus, the 1st quartile (25th percentile) is 12.23 and the 3rd quartile (75th percentile) is 17.00, indicating that half of the survival times fall between these values.

- CD4 cells count varies from 0.000 to 24.125 with a median of 5.477 and a mean of 7.023, suggesting positive skewness (see Figure 7). Lower quartile of 3.162 and the upper quartile of 10.440 with a greater maximum of 24.125, showing the appearance of outliers in which there is only few people who have significant amount of cd4 cells.

Moreover, there are not many subjects who reached 20 values of CD4 cells count in the study. At `time_obs` 0, the start of the study, there was not a single subject with cd4 over 20, but we observe higher CD4 cells count (above 20) later on, which potentially indicates a positive progress of the treatment (see Figure 8).

Figure 1 shows that:

- ddI treatment seems to have slightly higher mean of CD4 cells count than ddC treatment. Also, the whisker is higher with only one outlier with the highest CD4 cells count, approaching 25.
- Female group seems to have higher CD4 cells count than male group with slightly higher mean.
- People without previous opportunistic AIDS infection have much higher CD4 cells count, which is logical. But, notice that there are many points outside AIDS group (green box at the very left) whisker, which indicates a possible positive progression of the treatment over time.
- People who were intolerance to the AZT therapy have higher CD4 cells count than those who failed.

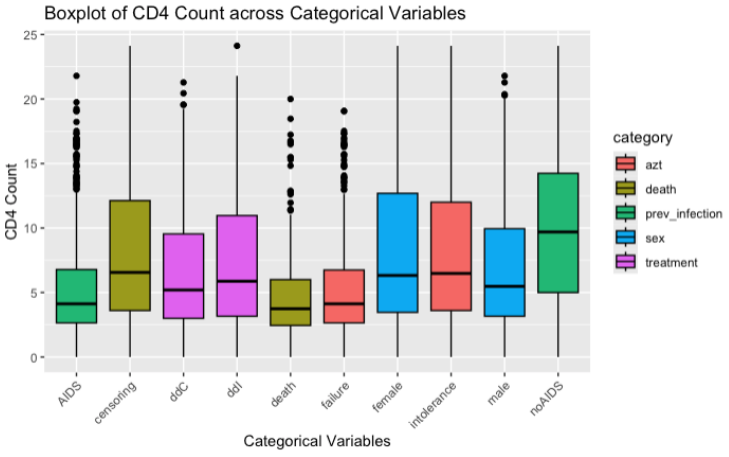

FIGURE 1: Boxplot of `cd4` against each categorical covariate

3.2 Categorical analysis

From Figure 2, x-axis represents the categorical variables, while y-axis measures the survival status.

- We can almost observe a horizontal line for `treatment` and `sex`, which suggests that there is no noticeable difference between them.

- The mosaic plot bars are not equal, for both `prev_infection` and `azt`, which tells that most of the data are censored (smaller amount of death) for people without previous AIDS infection and were intolerant to AZT therapy.

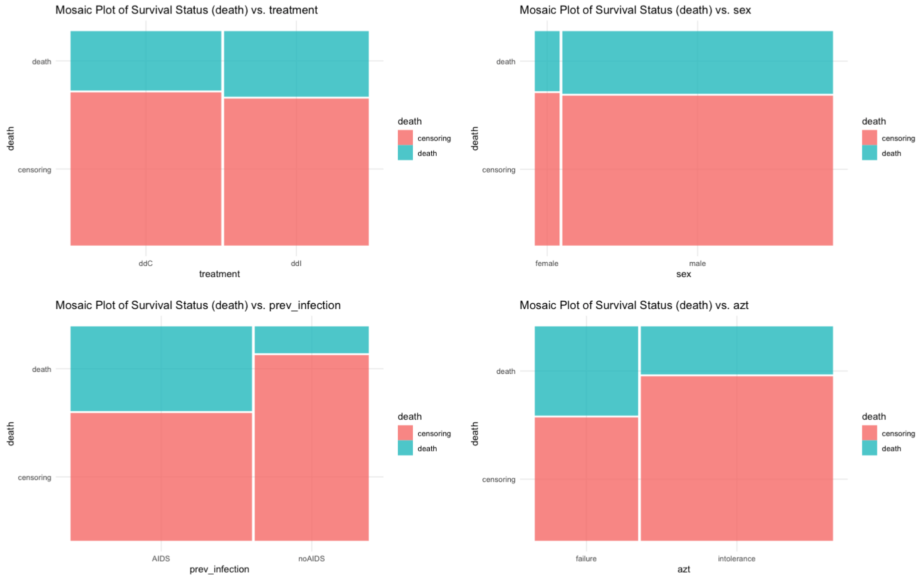

FIGURE 2: Mosaic plot of each categorical covariate by `death`

Since the aim of the study was to compare the efficacy and safety of `treatment`, each categorical variable will be shown conditionally against death proportion. According to Figure 3:

- There is no sight of significant effectiveness of the treatment though we can see that for female group, ddC treatment is slightly or barely better than ddI treatment. Same for male group.
- For prev_infection condition, there would be no difference between the treatment if the subject had AIDS infection before. Nevertheless, if the subject had no previous infection, ddC treatment would be better than ddI treatment as we can see on the right side of the plot that the censoring proportion is larger than death.
- Once again, ddC treatment shows a better survival proportion until the end of the study than ddI. But, what is interesting is that if they failed AZT therapy, it might potentially be more effective with ddI treatment.

It is worth noting that the mosaic plot only shows proportions, not statistical significance, so this conclusion is a preliminary observation. To confirm, statistical tests will be needed.

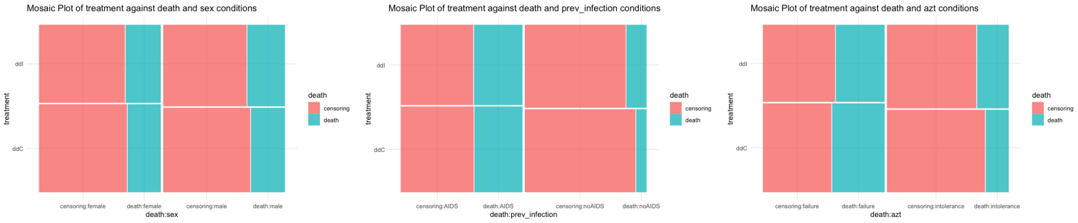

FIGURE 3: Conditional `treatment` mosaic plot of each categorical covariate by `death`

3.3 Relationship between covariates

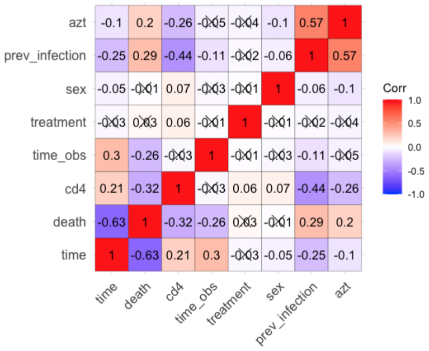

Figure 4: Correlation matrix between covariates

In Figure 4, the mark 'X' indicates insignificance coefficient (big p-value). The correlation matrix can convey various meaning, however, those interesting are that:

- The correlation coefficient of `treatment` and `cd4` is really low. This means that the treatments do not have noticeable effect on CD4 cells.
- `cd4` is being negatively correlated with `death` means that the lower CD4 cells count becomes, the higher likelihood of experience the event (death) is. Whereas, the positive correlation with `time` says that high `cd4` associates to longer survival times.
- Besides the logical negative relation between `death` and `time`, we can observe that `death` is also correlates with `prev_infection` and `azt`, though they are relatively weak.

Overall, we do not observe an extreme correlation between covariates. The highest are 0.57 and -0.44 of `azt/prev_infection` and `cd4/prev_infection` respectively (excluded `death` and `time`). Nonetheless, relationships that could influence survival of patients are `cd4, prev_infection` and `azt`.

3.4 Kaplan-Meier analysis

According to Figure 5, for curves that overlapping each other, group `treatment/sex`:

- ddI treatment seems to have slightly better survival probability as time progresses, leveled off at approximately 65% at time 18 vs. 58% of ddC treatment, though it is observable that ddC is more effective at the beginning. While for `sex`, male has better survival rate at the end.
- The overlapping confidence intervals suggest that the survival difference between may not be statistically significant. The noticeable wide interval for female group indicates extreme uncertainty in the survival estimation.

Whereas, for `prev_infection/azt`:

- Individuals with previous opportunistic AIDS infection have noticeable lower survival probability ever since the start of the study. People without infection are expected to survive with approximately 85% vs. 45% survival rate. Though it is not as significant as `prev_infection`, we still can clearly see that as time goes (beyong 10 months) intolerance patients to AZT therapy are highly expected to survive than the failed group (70% vs. 45%).

- The gap between the curves statistically suggests a significant difference, as the confidence intervals for the two groups rarely overlap. The confidence intervals for AIDS and failure start getting wider as well compare to noAIDS and intolerance respectively from time 13.

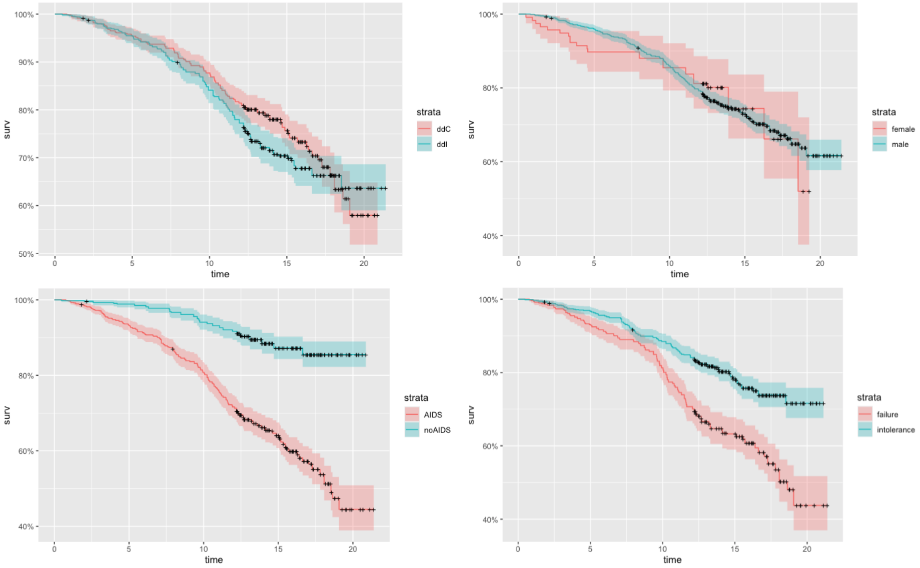

FIGURE 5: Survival function of each categorical covariate (see strata to identify the covariate)

Overall, survival estimates are more precise in the earlier time periods and become less reliable as time goes on. While `treatment` and `sex` might have a less impact on survival, `prev_infection` and `azt` have a notable impact on survival time. This assumption will be confirmed by statistical test below.

3.5 Time-varying effects of covariates

In Figure 9 where each panel corresponds to a covariate in Aalen's Additive Regression model. If the slope is negative while the covariate increases, it corresponds to a decreasing hazard rate and vice versa. Thus, we can observe that:

- The estimates of cumulative regression function for `sex`, `azt` and `treatment` appear to fluctuate near zero, suggesting a weak or negligible effect on the hazard.
- The positive slope over time of ddI treatment suggests that not only this treatment might initially have no effect but later on it instead increases the hazard. While for male patients, the survival might slightly improve over time compared to female.
- The coefficients of `cd4` and `prev_infection` are negative and becomes more negative over time, so there is nothing that contrary to the evidence above that higher amount CD4 cells improves survival and those without prior AIDS infection have higher survival rate than those with AIDS infection.

3.6 2-sample log-rank test

By performing 2-sample log-rank test, for `treatment` and `sex` respectively, observed and expected number of events (`death`) are relatively close, leading to small values of chi-square statistic, which in turn makes p-values large (above 0.05). Thus, under the null hypothesis that two survival functions are identical, there is no enough evidence to reject it. Conversely, for both `prev_infection` and `azt`, we have enough evidence to conclude that the survival experiences of their respective groups are not identical (statistical significance). Hence, the statistical inference from the test aligns with the assumption above.

4 Statistical modelling and analysis

In this section, different methodologies will be applied to find a model that best describe the survival projection of HIV patients.

4.1 Train-Test approach

To avoid having the same individual's measurements across both the training and testing sets, the splitting process was done by subjects identifier, by following a 80/20 partition of training and testing sets. Information on training and testing sets can be found in Table 4.

4.2 Data transformation

To properly address the time-dependent covariate `cd4` and attain the accurate modelling, the data was reformulated from long-format into start-stop format, where each observation represents a time interval during which the covariate `cd4` remains constant.

4.3 Variable selection

To ensure the predictive power of cox models, stepwise regression was performed by using bidirectional elimination; as a result, the model which yield the lowest AIC has 4 predictors: `cd4, prev_infection, sex, treatment`. Nevertheless, a list of different predictors with associated AIC/BIC is also provided in Table 5. Whereas, by virtue of random forest inherent feature selection which naturally selects relevant features during training, a full model including all predictors will be used.

4.4 Methodology

From Figure 11, there is no visible systematic pattern; on top of that, via Schoenfeld residuals, the global test p-value is above the level of 0.05, so there is no strong evidence to reject the proportional hazards assumption that covariates have constant coefficient. Thus, linear cox model is considered to be sufficient for this dataset. However, non-linear effect of cox model will also be introduced to see the difference. Thus, 4 statistical models were utilized to estimate the survival rate as listed below:

- Linear cox proportional hazards model (efron method)
- Non-linear cox proportional hazards model (efron method)
- Random forest (ntree = 500, mtry = 3, nodedepth = -1)
- Calibrated random forest (ntree = 100, mtry = 1, nodedepth = 7)

`ntree`: number of trees, `mtry`: number of variables that can be split each node, `nodedepth`: maximum depth of a tree

4.5 Performance

Metrics that were used to measure the performance of the aforementioned models are integrated brier score [1] and Harell's concordance index [2]. From Table 2, the optimized or calibrated random forest yield the lowest integrated brier score; but strangely, the model with the highest c-index is default random forest instead. Although the statistical test for nonlinearity is proved to be not significant, we can see that it has slightly better predictive performance than the linear model.

Model	Embedded brier score	Harrell's C-index
Linear cox	0.192	0.735
Non-linear cox	0.173	0.737
Random forest	0.153	0.785
Optimized random forest	0.152	0.748

Table 2: Performance table of different models

Visual representation of brier score evaluated on different times and survival curves for each model can be confirmed in Figure 6. The forests based model out-performed the cox proportional hazards model for this dataset with lower brier scores and higher survival curves.

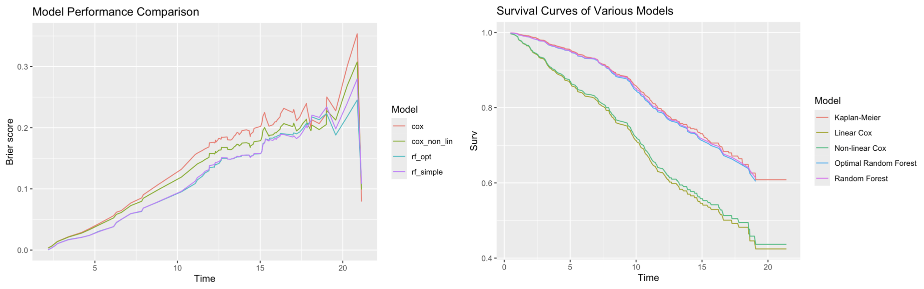

Figure 6: Brier score (left) and survival curve (right) for each model

4.6 Synthetic profile and scenario analysis

In this section, we will see the application of methodology by creating hypothetical scenarios to test the impact of specific covariates by fixing certain set of covariates and varying the others. The dummy data was made under two conditions: replace continuous variables by mean and discrete variables by mode. As a result, the information obtained is described below:

- ddC treatment is unconditionally better than ddI. It does not matter whether the subject is male or female, failed or was tolerate to AZT therapy, previously infected to AIDS or not.
- `cd4` and `prev_infection` are still influential for survival probability. Theoretically, if a subject were to have high cd4 cells count and previously got infected, he/she will likely to survive after 18 months. Although this theory is not common, but it shows that `cd4` is the most influential factor.
- With identical information of `cd4`, `treatment` and `prev_infection`, even though a male failed the AZT therapy, he will have higher survival probability than female that did not fail (intolerance).

Furthermore, consider a patient with AIDS infection who has been alive for 12 months since the beginning of the study. In theory of survival analysis, the survival function is a decreasing function of time. Thus, the longer the patient survives, the less likely they are to survive for an additional period. However, in practice, survival probability at times 18 ($P(T > 18)$) is numerically smaller than the one at times 12 $P(T > 12)$. Therefore, the formula below will inevitably yield an increased of conditional survival probability.

$$P(T > 18/T > 12) = \frac{P(T > 18, T > 12)}{P(T > 12)} = \frac{P(T > 18)}{P(T > 12)} = \frac{0.622}{0.738} \approx 0.843$$

So by knowing that a patient has been alive for 12 months, the survival probability at times 18 increases from 0.622 to 0.843 (note that the probability can be varied depending on the profile). This demonstrates how surviving certain period (12 months) positively influences survival projections beyond what the unconditional probability suggests.

5 Conclusion and remarks

In conclusion, two factors that most influence the survival of HIV patients of the longitudinal study of alternative treatment effects are notably `cd4` and `prev_infection`; nevertheless, it is worth noting that `sex (male)` and `treatment (ddI)` seem to have positive impact though it is really little.

Various aspects that should be investigated:

- From Figure 10, it is interesting that `azt` is seen as important covariate.
- For performance purposes, cox proportional hazards model with the lowest AIC model was chosen in this project, but the model can be simplified by just two influential predictors `cd4` and `prev_infection`. Plus, the model with these two predictos has the smallest BIC as well, which can be referred to accordingly.
- Optimized random forest's performance is indeed better than the default one with lower embedded brier score, however, it has lower c-index, which should be looked into.
- The Kaplan-Meier (KM) survival curve in Figure 6 is almost overlap with random forest ones. The KM curve shown is a baseline without any predictors, which depends solely on the observed data. This might suggest powerless predictors, which researchers should incorporate other explanatory variables if a more robust and stronger model is needed.

Improvements that could be done:

- Integration of new models. In the meantime, the ensemble methods with time-varying covariates [3] (package `LTRCforests` in `R`) have already been fit in the code, but as time is limited, I decided to not go further.
- The optimzation approach on random forest was done naively using grid search by setting range of parameters by hand. Other approaches should be considered as the calibrated is only slightly better here.

A Appendix

A.1 Supplementary tables

Variable	Category	Count	Proportion (%)
`death`	censoring	279	59.74
	death	188	40.26
`sex`	female	45	9.63
	male	422	90.37
`treatment`	ddC	237	50.75
	ddI	230	49.25
`prev_infection`	AIDS	307	65.74
	noAIDS	160	34.26
`azt`	failure	175	37.47
	intolerance	292	62.53

TABLE 3: Count and proportion of each level occurrence for categorical variables

Total subjects	467 (Train: 375, Test: 92)	
	Death	**Censored**
Initial data	188	279
Proportion	40.26%	59.74%
Train data	150	225
Proportion	40%	60%
Test data	38	54
Proportion of death	41.30%	58.70%

TABLE 4: Death and censored data counts and proportions

Predictor	cd4, prev	cd4, prev, sex	cd4, prev, treat	cd4, prev, sex, treat	cd4, prev, sex, treat, azt	pspline(cd4), prev, sex, treat
AIC	1641.798	1641.477	1640.829	1640.234	1641.565	1643.561
BIC	1647.910	1650.646	1649.998	1652.459	1656.847	1664.866

TABLE 5: AIC/BIC criterion of cox linear and non-linear (last column) model

A.2 Supplementary figures

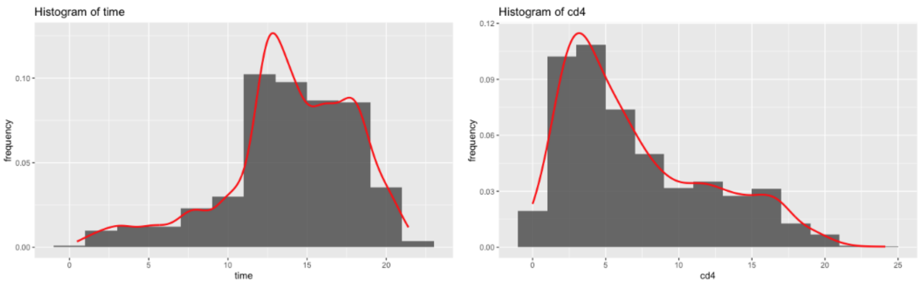

FIGURE 7: Distribution of `time` (left) and `cd4` (right)

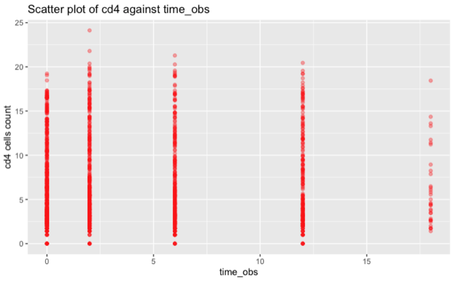

FIGURE 8: Scatter plot of `cd4` by different time observations (`time_obs`)

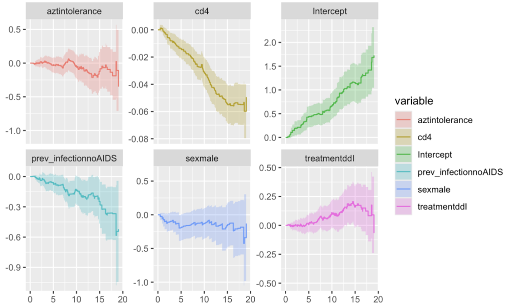

FIGURE 9: Estimated cumulative regression functions based on Aalen's Additive Regression

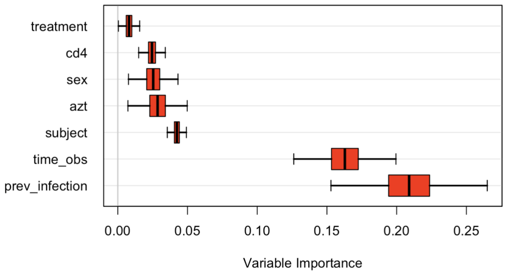

FIGURE 10: Variable importance from random forest

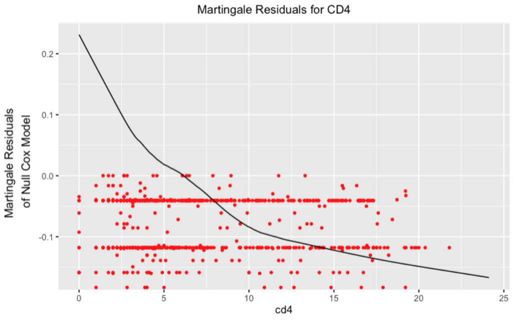

FIGURE 11: Plot of martingale residuals against `cd4`

A.3 Code

This project was done in `R` with the help of different sources: [4], [5], [6], [7], slides provided by Dr. PARK Juhyun. Link to source code: https://github.com/richardcheam

References

[1] The PySurvival Team. (n.d.). Brier score - PySurvival. Retrieved from https://square.github.io/pysurvival/metrics/brier_score.html

[2] The PySurvival Team. (n.d.). C-index. Retrieved from https://square.github.io/pysurvival/metrics/c_index.html

[3] Yao, W., Frydman, H., Larocque, D., Simonoff, J. S. (2022). Ensemble methods for survival function estimation with time-varying covariates. Statistical methods in medical research, 31(11), 2217–2236. https://doi.org/10.1177/09622802221111549

[4] Views, R. (2017, September 25). Survival analysis with R. Retrieved from https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r

[5] Random survival forests. (2023, May 29). Retrieved from https://www.randomforestsrc.org/articles/survival.html

[6] Başar, E. (2017). Aalen's additive, Cox proportional hazards and the Cox-Aalen Model: Application to kidney transplant data. Sains Malaysiana, 46(3), 469-476. https://www.ukm.my/jsm/pdf_files/SM-PDF-46-3-2017/15%20Aditif%20Aalen.pdf

[7] Dietrich, S., Floegel, A., Troll, M., Kühn, T., Rathmann, W., Peters, A., Sookthai, D., von Bergen, M., Kaaks, R., Adamski, J., Prehn, C., Boeing, H., Schulze, M. B., Illig, T., Pischon, T., Knüppel, S., Wang-Sattler, R., & Drogan, D. (2016). Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. International journal of epidemiology, 45(5), 1406–1420. https://doi.org/10.1093/ije/dyw145