Statistical Methods for Observational Studies Survival Analysis Project
Nicholas Cvercko

APPENDIX WITH SAS CODE AND OUTPUTS SUBMITTED AS SEPARATE ATTACHMENT

The data was collected from a study of longevity conducted on subjects who were enrolled at an age greater than 25 years. Subjects' ages were last censored in 2021 resulting in a follow up of around 8 years in the study. The goal of the collected data and subsequent analyses was to test the association between inflammation and death, the two log-transformed concentration variables of IL6 and hsCRP concentrations are indicators of inflammation, and the aim of the current statistical analysis is to determine if one possesses a statistically significant association with death, and to determine if other covariates from the phenotypic data listed above play a role in the mentioned association.

METHODS:

COMPLETE CASE ANALYSIS OF MISSING VARIABLES

The frequency of missing variables was observed for each variable after data collection was completed. It was necessary to eliminate missing data by either imputation or deletion in a fashion ensuring the contribution of all variables to the analysis would not be hindered by missing information.

PRELIMINARY INSPECTION WITH KAPLAN MEIER CURVES AND TESTS

The two inflammation protein concentrations were evaluated to determine if they visually presented evidence of association with death. Furthermore, the Kaplan Meier test statistics were produced to evaluate whether there was a difference in the resulting curves representing the two variables. The results of the Kaplan Meier curves and test helped to determine which variable to use as the predictor variable of interest.

PROPORTIONAL HAZARDS ANALYSIS

The PHREG procedure in SAS using time-varying effects was applied in order to evaluate if the assumption of proportional hazards held true for all variables used as predictors. Furthermore it was used to generate a Cox Proportional Hazards model in order to determine the Hazard Ratio, as well as test for a statistically significant association.

ANALYSIS FOR CONFOUNDING

The Hazard Ratios collected were used to evaluate whether or not there was evidence of confounding between the group of variables used for analysis in the full model, and the crude model which only used the predictor variable of interest.

TESTS FOR INTERACTION

Tests for interaction were used in order to further evaluate the role the covariates played in the association of the predictor variable of interest. The Cox Models were run with interaction variables included in order to determine if they held statistically significant associations with survival.

RESULTS:

The method of complete case analysis was used as each observation containing missing information for a variable used in the analysis was removed from the data. The first employment of complete case analysis was conducted initially to remove missingness of the Alive variable as well as the variable for age at last contact which was used to calculate time to censorship. Next, evaluating the frequency of missing data, it was determined that the majority (over 68%) of missingness was accounted for by DSST and Z.fev1.7 which is the measure of lung function. These variables were therefore completely excluded from future analysis as their inclusion would involve a large number of subjects being removed from the data. Moving forward every variable was used in the complete adjusted analysis and therefore every instance of missing data required the removal of the subject.

Preliminary analysis: the two variables of Il6 concentration and hsCRP were both temporarily converted into categorical variables of low and high levels of concentration separated by the median. Therefore it would be easy to visually assess the Kaplan Meier Curves.. The resulting curves for hsCRP were very similar to each other, and it was difficult to visually evaluate the proportional hazards assumption. The resulting curves for IL6 were also very similar, yet it was more obvious that the increase in IL6 would lead to increased probability of survival as age increased. The proportional hazards assumption did not appear to be violated visually, however the Log-Rank test was conducted for the Kaplan Meier with the following hypotheses: H0: the survival curves are the same, and HA: the survival curves are not the same. IL6 produced a Log-Rank Chi-Square test statistic of 11.7434 with 1 df and a p-value of 0.0006, this being less than the threshold of significance indicated sufficient evidence to reject the null

hypothesis and conclude the two curves are statistically significantly not the same. hsCRP returned a Log-Rank Chi-Square test statistic of 2.5717 with 1 df and a p-value of 0.1086, indicating that there was not sufficient evidence to reject the null hypothesis and therefore there is no statistically significant difference in the curves. For these reasons IL6 was chosen as the predictor variable of interest to evaluate for association with death. Only one inflammation protein concentration is included in analysis, to only have the main predictor protein evaluated for association and create a model with an inflammatory protein analysis and hazard independent of the other protein.

CRUDE ANALYSIS OF IL6 ASSOCIATION:

A model was created to test the assumption of proportional hazards using IL6 and a time-varying effect variable of IL6. The global Likelihood Ratio hypothesis test was conducted on the model with the following hypotheses: H0: neither IL6 nor the time-varying effect of IL6 have an effect on the hazard. HA: at least one of IL6 or the time-varying effect of IL6 has an effect on the hazard. The Likelihood Ratio results in a Chi-Square test statistic of 6.18 with 2 df and a p-value of 0.0455 showing there is sufficient evidence to reject the null hypothesis and conclude that at least one has a statistically significant impact on the hazard. Hypotheses tested for assumption of proportional hazards: H0: proportional hazards assumption is not violated, HA: the model fails the proportional hazards assumption. The likelihood ratio test returns a chi-square test static of 0.0142 with 1df and a p-value of 0.9052 showing insufficient evidence to reject the null hypothesis, therefore concluding that the proportional hazards assumption is not violated. The Cox proportional hazards model was then created to determine the hazard ratio and association. This model also produced a statistically significant global p-value of 0.0150 with a chi-square test statistic of 5.9115 at 1 df. The following hypotheses were tested for the parameter of log_IL6: H0: There is no association between IL6 and the risk of death occurring, the hazard ratio is equal to 1. HA: There is an association between IL6 and the risk of death occurring, the hazard ratio is not equal to 1. The Likelihood Ratio Test returns a chi-square test statistic of 5.8251 with 1 df and a p-value of 0.0158, indicating there is sufficient evidence to reject the null hypothesis and conclude that there is a statistically significant association between IL6 and the risk of death occurring, the ratio is not 1. The resulting Hazard ratio is 0.950 indicating the risk of death was 5% lower in the group exposed to IL6 compared to those not.

ADJUSTED ANALYSIS OF IL6 ASSOCIATION AND CONCLUSION:

The global models for all tests were evaluated and the result was a p-value below the threshold of significance and subsequent rejection of the null hypothesis of no impact on the hazard. Hypothesis tests from crude analysis were repeated for adjusted models. Every variable in the adjusted model and their time-varying effects failed to reject the null hypothesis of statistically significant compliance with the proportional hazards assumption. The resulting chi-square statistic for log_il6 from the Cox Proportional Hazard Model was 2.8748 with 1df and a p-value of 0.09 which showed insufficient evidence to reject the null hypothesis of no association. Z_grip_strength and educ both indicated statistically significant associations with the risk of death occurring with p-values of <0.0001 and 0.0003. To evaluate the drastic change in p-value for log_il6, interaction was evaluated by entering interaction terms for log_il6 with grip strength and educ and they were evaluated with the same hypothesis testing. A statistically significant association between the interaction term for log_il6 and educ was observed, but not grip strength. This led to an adjusted model of all covariates plus an interaction term for log_il6 and educ with a chi-square statistic of 5.5061 at 1df with a p-value of 0.019. The log_il6 chi-square statistic in the final model was 7.314 with 1df and a p-value of 0.0068, concluding statistically significant association between log_il6 and the hazard of the event of death. The hazard ratio for log_il6 from the final model with interaction was 0.960 indicating a 4% less likely hazard for 1 unit change in log_il6 with the event of death .

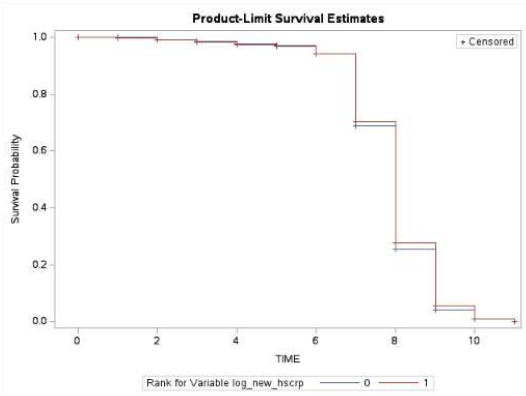
To evaluate confounding the 10% rule was evaluated with crude ratio of 0.95 and adjusted of 0.963, $|0.95-0.960|/0.95 = 0.0105$ which is < 10% indicating no need to adjust for confounding between the group of covariates of sex, education, grip strength, bmi, gait speed, or systolic bp.

The model which included interaction between log_il6 and educ indicated a statistically significant association between log_il6 and the hazard of death. The full model without interaction returned a hazard ratio that suggested a decreased hazard of death with an increase in inflammation factor IL6 with less than 10% confounding caused by the group of covariates. IL6 is not the only protein responsible for inflation, more analyses could be conducted to evaluate if other factors or proportions are associated with hazard.

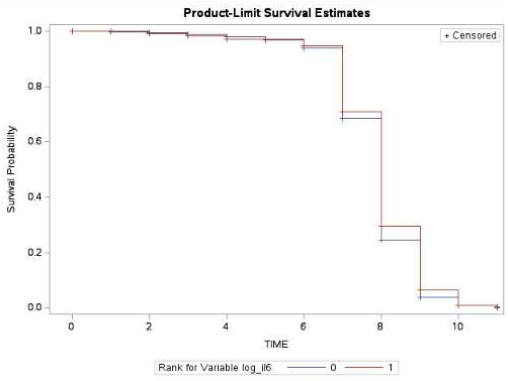
Statistical Methods for Observational Studies Survival Analysis Project
Nicholas Cvercko

KAPLAN MEIER CURVES AND TESTS

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	2.5717	1	0.1086
Wilcoxon	1.8467	1	0.1742
-2Log(LR)	10.3922	1	0.0013



Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	11.7434	1	0.0006
Wilcoxon	8.3596	1	0.0038
-2Log(LR)	64.0635	1	<.0001



CRUDE VS ADJUSTED ANALYSIS

	log_il6 Chi-Sqr	df	log_il6 p-value	Hazard Ratio
Crude	5.8251	1	0.0158	0.950
Adjusted	2.8748	1	0.09	0.963
Interaction	7.3140	0.0068		
*educ (interaction term)	5.5081	0.0190		

SUMMARY AND FREQUENCY OF DATA AFTER REMOVAL OF MISSING DATA

The FREQ Procedure

Alive	Frequency
Not Missing	4309

educ	Frequency
Not Missing	4309

DSST	Frequency
Missing	183
Not Missing	4126

sex	Frequency
Not Missing	4309

log_il6	Frequency
Not Missing	4309

Age_last_contact	Frequency
Not Missing	4309

Age_enrollment	Frequency
Not Missing	4309

log_new_hscrp	Frequency
Not Missing	4309

Z_grip_strength	Frequency
Not Missing	4309

Z_bmi	Frequency
Not Missing	4309

Z_gait_speed	Frequency
Not Missing	4309

Z_fev1_7	Frequency
Missing	415
Not Missing	3894

Z_sysbp	Frequency
Not Missing	4309

TIME	Frequency
Not Missing	4309

EVENT	Frequency
Not Missing	4309

Variable	N	Mean	Minimum	Maximum
sex	4309	1.5388721	1.0000000	2.0000000
log_il6	4309	0.0911336	-2.5257289	4.8828019
educ	4309	11.6539800	0	17.0000000
Z_grip_strength	4309	0.0187399	-3.6646254	5.8807748
Z_bmi	4309	0.0118473	-3.0300328	3.9505092
Z_gait_speed	4309	0.0257414	-4.4444483	4.5799524
Z_sysbp	4309	0.0229385	-2.9511742	5.3621917

Z_fev1_7	Frequency
Missing	415
Not Missing	3894

HR FOR FINAL MODEL

HR: 1 unit difference: Hazard Ratios for log_il6			
Description	Point Estimate	95% Wald Confidence Limits	
log_il6 Unit=1 At educ=11.65398	0.960	0.919	1.003