Sara (Shiv) Denner
BS852 Course Project
Due 12/9/2024

Introduction

The framdat4 dataset includes 2000 observations from a subsample of data from the Framingham Heart Study, which is a longitudinal study. Subjects had data measured at exam 4, including age (AGE4), total cholesterol (CHOL4), whether the subject smokes (SMOKE), number of cigarettes smoked per day (CIGS4), systolic and diastolic blood pressure (SPF4 and DPF4), weight in pounds (WGT4), pulmonary function (FVC4), BMI (BMI4), hypertension (HTN4), and menopause (MENO4). The incidence of CHD, type 2 diabetes, and death were recorded following 22 years of follow-up. We are interested in exploring the association between systolic blood pressure (SPF4) and overall survival.

Methods

Since this is a longitudinal study. survival analysis was used in all following analyses to account for time to death and censoring. First, variable selection techniques were used to create a final Cox Proportional Hazards (PH) model with SPF4 as the main predictor and including other relevant covariates. MENO4 was dropped due to its relevance only to females, while T2D and T2D_SURV were removed due to irrelevance to the analysis. Univariate analyses with Spearman correlations and survival analyses identified collinear variables and variables significantly associated with mortality. Potential confounders to SPF4 were identified using bivariate survival analyses and the 10% rule. A full Cox PH model was made including all variables that were significantly associated with mortality and that were confounders to SPF4, and a Wald Test at a significance level of 0.15 determined which predictors were significant while adjusting for all other predictors. Non-significant variables were removed, and all remaining variables were tested for collinearity via a correlation test. The final Cox PH model included non-collinear variables that were significant in the Wald test. Results were validated with forward and stepwise automatic variable selection at a 0.15 significance level.

Once the final model was generated, a new dataset was created with just the final covariates, predictor of interest, and censoring and time to event variables. CHD and CHD_SURV were included but not used, as CHD development over 22 years is a different outcome from death, which is the focus here. Rows with missing values for any covariates, DTH, or SURV variables were removed, resulting in 1,964 non-missing observations. The Schoenfeld residuals method tested the proportional hazards assumption on the final model. Continuous covariates that violated the PH assumption were converted to categorical variables, and all following survival analyses were stratified accordingly. Summary statistics were generated using a proc means procedure.

A crude, stratified survival analysis was used to assess the association between SPF4 and mortality. Analysis of Schoenfeld residuals showed no violation of the PH assumption. The null hypothesis states that SPF4 is not associated with mortality. Confounding by each covariate in the final model was assessed using bivariate Cox PH analysis applying the 10% rule to crude and adjusted hazard ratios for SPF4. We tested the null hypothesis that there is no interaction by sex to explore whether the association between SPF4 and mortality differs between males and females. An interaction term for SPF4 and SEX was added to the final Cox PH model and was tested for significance at a .05 significance level. Hazard ratios for SPF4's effect on mortality, adjusted by all covariates, were calculated separately for males and females and compared.

Survival analysis was used to assess if CHD (coded as 1 if CHD occurred during the 22 years of follow up) is associated with total mortality, using exam 4 as the baseline. A Kaplan-Meier plot compared the survival functions of 2 groups classified by the dichotomous CHD variable, and the Log-Rank test formally tested the association between developing CHD and survival under the null hypothesis of no difference between survival curves. A limitation of this analysis is that CHD is a time-varying variable which may cause the proportional hazard assumption to fail. Therefore, results of the analysis testing the association between CHD and mortality may be biased.

Results

Correlation analysis showed that BMI and WGT4 (Spearman correlation coefficient r = 0.79), CIGS4 and SMOKE (r = 0.95), and SPF4 and DPF4 (r = 0.76) were collinear. Univariate survival analysis showed that all variables (SEX, AGE4, CHOL4, CIGS4, WGT4, FVC4, BMI4, HTN4) were significantly associated with mortality except SMOKE. DPF4 was dropped due to collinearity with predictor of interest SPF4, and SMOKE was dropped due to collinearity and insignificance. Bivariate survival analyses showed that no variables confounded the association between SPF4 and mortality. Remaining variables were included in a Cox PH model and adjusted significance was determined via a Wald test for specific variables (Table 5). Significant predictors (SEX, AGE4, CIGS4, SPF4, FVC4) were retained, while insignificant ones (CHOL4, HTN4, WGT4, BMI4) were dropped. Correlation analysis between each of the final covariates showed that none of the predictors were collinear (no coefficient > .6). The final model was validated using forward and stepwise selection at a 0.15 significance level.

Analysis of Schoenfeld residuals on the final model indicated that AGE4 violated the Proportional Hazards assumption, so it was converted into a categorical variable based on age groups (Table 6). All remaining analyses were stratified by the age group variable. Results of the crude survival analysis showed that SPF4 is significantly associated with mortality (Likelihood Ratio Chi-Square = 73.49 on 1DF and p<.0001, Wald Chi-Square = 83.26 on 1DF and p<.0001), and that with every unit increase in SPF4, the hazard of death increases by 1.014 units within the same age stratum (Table 2). Note that we cannot compare hazard of death based on SPF4 for subjects in different age strata. Age group stratified bivariate survival analysis and the 10% rule applied to crude and adjusted hazard ratios showed no covariate confounded the relationship between SPF4 and mortality (Table 2). Analysis of the interaction term included in the final Cox PH model showed that there is no interaction by sex on the association between SPF4 and mortality (Wald Chi-Square = 0.0435 and p = 0.8347), and therefore that the association between SPF4 and mortality is the same in males and females (Table 3).

Survival analysis was used to determine whether CHD is associated with mortality. The Kaplan-Meier plot illustrates the difference in survival between individuals who developed CHD over 22 years of follow up and those who did not, and that the two groups seem to satisfy the PH assumption (Figure 1). Those who did not develop CHD over follow up appear more likely to survive longer than those who did. Results of the Log-Rank test are consistent with the results of the Kaplan-Meier curves and suggest that there is a significant association between CHD and mortality (Log-Rank Chi-Square = 170.17 on 1df and p<.0001), therefore we reject our null hypothesis (Table 7). To include time-to-event variables for CHD alongside time-to-death variables in our survival analysis with SPF4 as the main predictor, we could use a competing risk analysis. The Fine and Gray Subdistribution Hazard Model would calculate the hazard of death in the presence of CHD as a competing event.

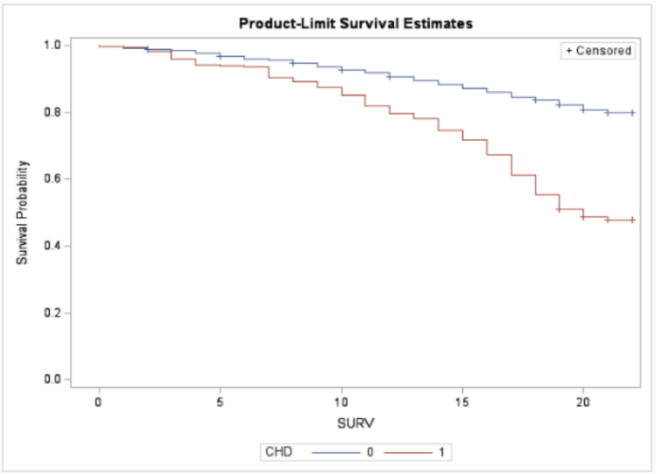

Figure 1: Kaplan-Meier curves illustrating the difference in survival for those who developed CHD over follow up and those who did not.

Discussion

We performed survival analysis to assess the association between systolic blood pressure (SPF4) at exam 4 and mortality over 22 years of follow up. Before analyses occurred, variable selection techniques were used to determine a final Cox PH model for the question of interest. No confounding was found between any of the chosen covariates and SPF4, and no interaction by sex on the association between SPF4 and mortality was found. Survival analysis was also used to show that CHD is significantly associated with mortality, though this analysis is likely biased due to the time-varying nature of the CHD variable. Finally, we found that competing risk analyses could account for the addition of more complex time-to-CHD variables as a competing risk in the survival analysis at hand.

Continued Results– Additional Page of Figures and Tables

SEX	*Number of Observations*	*Variable Name*	*Minimum*	*Maximum*	*Median*	*Mean*	*Std Deviation*
1 = Male	850	AGE4	34.0	69.0	49.0	49.7	8.6
		CIGS4	0.0	43.0	9.0	12.5	13.3
		SPF4	90.0	252.0	130.0	133.0	21.7
		FVC4	227.0	833.0	539.0	542.5	93.4
2 = Female	1114	AGE4	34.0	68.0	49.0	49.9	8.4
		CIGS4	0.0	43.0	0.0	5.2	8.7
		SPF4	88.0	290.0	130.0	134.6	25.9
		FVC4	82.0	636.0	419.0	413.2	85.6

Table 1: summary of patient characteristics for all variables included in the final model (generated with PROC MEANS)

Analysis Type	*Predictors Included*	*p-value of Association*	*Chi-Squared Value of Association*	*Effect (Beta) Coefficient of SPF4*	*HR for 1 unit increase of SPF4 within each age stratum*	*Confounding? (Y/N)*
Crude Cox PH	SPF4	<.0001	83.3	0.01424	1.014	N/A
Adjusted Cox PH	SPF4 SEX	<.0001	104.7	0.01600	1.016	N
Adjusted Cox PH	SPF4 AGE4	<.0001	68.8	0.01311	1.013	N
Adjusted Cox PH	SPF4 CIGS4	<.0001	91.1	0.01509	1.015	N
Adjusted Cox PH	SPF4 FVC4	<.0001	81.8	0.01460	1.015	N
Adjusted Cox PH	SPF4 SEX AGE4 CIGS4 FVC4	<.0001	72.3	0.01384	1.014	N

Table 2: Measures of crude and adjusted associations between SPF4 and mortality within each age stratum, generated with bivariate Cox PH analyses with SPF4 as predictor of interest.

Parameter	*Parameter Estimate*	*Standard Error*	*Chi-Square*	*p-value of Term*
SPF4	0.01418	0.00230	38.1026	<.0001
SEX	-0.73051	0.45743	2.5504	0.1103
SPF4*SEX	-0.00063	0.00302	0.0435	0.8347
AGE4	0.07397	0.01728	18.3349	<.0001
CIGS4	0.01722	0.00394	19.0606	<.0001
FVC4	-0.00233	0.00055	16.9370	<.0001

Hazard Ratios for SPF4			
Description	*Point Estimate*	*95% Wald Confidence Limits*	
SPF4 at SEX = Male (1)	1.014	1.010	1.019
SPF4 at SEX = Female (2)	1.014	1.009	1.018

Table 3: gender-specific associations between SPF4 and mortality, adjusting for all covariates. Results show that the interaction term is not significant, and that the hazard ratio for SPF4 is the same between males and females.

Appendix 1: Supplemental Materials

Spearman Correlation Coefficients		
	BMI4	***WGT4***
BMI4	1.0000	0.7905
WGT4	0.7905	1.0000
	CIGS4	***SMOKE***
CIGS4	1.0000	0.9449
SMOKE	0.9449	1.0000
	SPF4	***DPF4***
SPF4	1.0000	0.7586
DPF4	0.7586	1.0000

Table 4: Spearman Correlation Coefficients from correlated variables at exam 4 generated during univariate analysis portion of model selection.

Analysis of Maximum Likelihood Estimates								
Parameter	*DF*	*Parameter Estimate*	*Standard Error*	*Chi-Square*	*PR > ChiSq*	*Hazard Ratio*	*95% Hazard Ratio Confidence Limits*	
SEX	1	-0.793	0.128	38.293	<.0001	0.453	0.352	0.582
AGE4	1	0.075	0.007	125.462	<.0001	1.078	1.064	1.092
CHOL4	1	3.55E-5	0.001	0.001	0.975	1.000	0.998	1.002
CIGS4	1	0.018	0.004	18.199	<.0001	1.018	1.010	1.026
SPF4	1	0.013	0.002	32.022	<.0001	1.014	1.009	1.018
FVC4	1	-0.002	5.884E-5	13.339	0.0003	0.998	0.998	0.999
BMI4	1	-0.004	0.012	0.104	0.747	0.996	0.974	1.019
HTN4	1	0.039	0.129	0.089	0.765	1.039	0.806	1.340

Table 5: Results of Wald Test for specific variables from the Cox PH Model showing insignificance of BMI4, HTN4, CHOL4. Note WGT4 was not included since it is correlated with BMI4.

Zph Tests for Nonproportional Hazards						
Transform	*Predictor Variable*	*Correlation*	*Chi Square*	*Pr > ChiSquare*	*T Value*	*Pr > |t|*
RANK	SPF4	-0.0194	0.2071	0.6490	-0.44	0.6567
RANK	SEX	0.0593	1.9540	0.1622	1.36	0.1738
RANK	AGE4	0.0866	4.1643	0.0413	1.99	0.0467
RANK	CIGS4	0.0364	0.7382	0.3902	0.84	0.4036
RANK	FVC4	0.0095	0.0570	0.8113	0.22	0.8272

Table 6: Results from Schoenfeld Residuals Analysis of final Cox PH Model with SPF4, SEX, AGE4, CIGS4, FVC4 illustrating AGE4's departure from the proportional hazards assumption. These results showed that it was necessary to stratify the following survival analyses by age group.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	170.1727	1	<.0001

Table 7: Result of the Log-Rank test testing the null hypothesis that there is no association between development of CHD over 22 years follow-up and risk of mortality. Results should be interpreted in tandem with Figure 1.

Appendix 2: SAS Scripts

```
libname project 'C:\Users\sarad\OneDrive\Desktop\School\MSAB\BS852\Project';

proc import file = "C:\Users\sarad\OneDrive\Desktop\School\MSAB\BS852\Project\framdat4.csv"
out = project.fd
dbms = CSV REPLACE;
getnames = YES; datarow = 2;
run;

proc print data = project.fd;
run;

**Model Building - identify covariates and confounders**;

*Drop MENO4 because it is only relevant for females;
*Drop T2D and T2D_SURV because we do not need these variables for our analysis;
data project.fd1;
set project.fd (drop = MENO4 T2D T2D_SURV);
run;

proc print data = project.fd1;
run;

*BMI and WGT4 likely correlated - test with proc corr;
proc corr data = project.fd1 spearman;
var BMI4 WGT4;
run;
*correlation coeff is .79 --> they are collinear so choose 1 over other

*Then see which one creates the best model fit using AIC;

*CIGS4 and SMOKE;

proc corr data = project.fd1 spearman;
var CIGS4 SMOKE;
run;
*correlation coeff is .945 --> they are collinear so choose 1 over other

*SPF4 and DPF4;
proc corr data = project.fd1 spearman;
var SPF4 DPF4;
run;
*correlation coeff is .759 --> they are collinear. Since SPF4 is predictor of interest, drop DPF4;

data project.fd2;
set project.fd1 (drop = DPF4);
run;

proc print data = project.fd2 (obs = 25);
run;

*Univariate survival analysis to determine which variables are associated with mortality;

proc phreg data = project.fd2;
MODEL SURV*DTH(0) = SEX / RL TIES = EFRON;
RUN;
*sex is significant, p<.0001 so keep;

proc phreg data = project.fd2;
MODEL SURV*DTH(0) = AGE4 / RL TIES = EFRON;
RUN;
*AGE IS SIGNIFICANT, P<.0001 SO KEEP;
```

```
proc phreg data = project.fd2;
MODEL SURV*DTH(0) = CHOL4 / RL TIES = EFRON;
RUN;
*CHOL4 IS SIGNIFICANT, P=.0006 SO KEEP;

proc phreg data = project.fd2;
MODEL SURV*DTH(0) = CIGS4 / RL TIES = EFRON;
RUN;
*CIGS4 significant, p = 0.0335;

proc phreg data = project.fd2;
MODEL SURV*DTH(0) = WGT4 / RL TIES = EFRON;
RUN;
*WGT4 is significantly associated with death, p<.0001 - run 1 final model with BMI4 and 1 with WGT4 and compare AIC;

proc phreg data = project.fd2;
MODEL SURV*DTH(0) = FVC4 / RL TIES = EFRON;
RUN;
*FVC4 IS SIGNIFICANT, P<.0001, SO KEEP;

proc phreg data = project.fd2;
model SURV*DTH(0) = BMI4 / RL TIES = EFRON;
RUN;
*BMI4 is significantly associated with death, p<.0001;

proc phreg data = project.fd2;
MODEL SURV*DTH(0) = HTN4 / RL TIES = EFRON;
RUN;
*HTN4 IS SIGNIFICANT, P<.0001, SO KEEP;
```

```
proc phreg data = project.fd2;
MODEL SURV*DTH(0) = SMOKE / RL TIES = EFRON;
RUN;
*SMOKE is not significant, p = .7028 - since CIGS4 and SMOKE are collinear, drop SMOKE;

DATA PROJECT.FD3;
SET PROJECT.FD2 (DROP = SMOKE);
RUN;

PROC PRINT DATA = PROJECT.FD3 (OBS = 25);
RUN;


*CHECK IF ANY VARIABLES CONFOUND THE ASSOCIATION BETWEEN SPF4 AND DEATH;
PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SPF4 / RL TIES = EFRON;
RUN;
*CRUDE HR = 1.022;

PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SPF4 SEX / RL TIES = EFRON;
RUN;
*ADJ HR = 1.023 - NO CONFOUNDING;

PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SPF4 AGE4 / RL TIES = EFRON;
RUN;
*ADJ HR = 1.013 - NO CONFOUNDING;
```

```
PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SPF4 CHOL4 / RL TIES = EFRON;
RUN;
*ADJ HR = 1.022 - NO CONFOUNDING;

PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SPF4 CIGS4 / RL TIES = EFRON;
RUN;
*ADJ HR = 1.023 - NO CONFOUNDING;

PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SPF4 WGT4 / RL TIES = EFRON;
RUN;
*ADJ HR = 1.021 - NO CONFOUNDING;

PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SPF4 FVC4 / RL TIES = EFRON;
RUN;
*ADJ HR = 1.020 - NO CONFOUNDING;

PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SPF4 BMI4 / RL TIES = EFRON;
RUN;
*ADJ HR = 1.023 - NO CONFOUNDING;

PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SPF4 HTN4 / RL TIES = EFRON;
RUN;
*ADJ HR = 1.021 - NO CONFOUNDING;

*NO VARIABLES CONFOUND THE ASSOCIATION BETWEEN SPF4 AND DEATH.
*SMOKE WAS REMOVED BECAUSE IT IS CORRELATED WITH CIGS4, AND WHILE WE FOUND VIA UNVARIATE ANALYSIS THAT CIGS4
*IS SIGNIFICANTLY ASSOCIATED WITH DEATH, SMOKE WAS NOT. SO SMOKE WAS REMOVED.

*SINCE ALL OTHER VARIABLES ARE SIGNIFICANT VIA UNIVARIATE ANALYSIS, TEST FOR OVERALL SIGNIFICANCE
* WHEN ALL INCLUDED TOGETHER IN A COX PH MODEL AND SEE WHICH ARE SIGNIFICANT VIA WALD TEST

 *RUN FULL MODEL WITH ALL COVARIATES INCLUDED WITH BMI AND GET AIC;
PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SEX AGE4 CHOL4 CIGS4 SPF4 FVC4 BMI4 HTN4 / RL TIES = EFRON;
RUN;
*AIC = 7243.438;

*RUN FULL MODEL WITH ALL COVARIATES INCLUDED WITH WGT4 AND GET AIC;
PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SEX AGE4 CHOL4 CIGS4 SPF4 FVC4 WGT4 HTN4 / RL TIES = EFRON;
RUN;
*AIC = 7243.438 - THEY'RE THE SAME

*IN THE FULL MODEL, THE FIT OF THE MODEL IS THE SAME WHETHER BMI4 OR WGT4 WAS INCLUDED. SINCE THEY'RE COLLINEAR
*WITH ONE ANOTHER AND NOT SIGNIFICANT BY WALD TEST, DROP BOTH OF THEM.;

PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SEX AGE4 CHOL4 CIGS4 SPF4 FVC4 HTN4 / RL TIES = EFRON;
RUN;

*USING THE WALD TEST FOR SPECIFIC VARIABLES, WE FIND THAT CHOL4 AND HTN4 ARE NOT SIGNIFICANT WHILE
*ALL OF THE OTHER VARIABLES ARE. SO REMOVE CHOL4 AND HTN4;
```

```
PROC PHREG DATA = PROJECT.FD3;
MODEL SURV*DTH(0) = SEX AGE4 CIGS4 SPF4 FVC4 / RL TIES = EFRON;
RUN;

*IN THIS MODEL, ALL COVARIATES ARE SIGNIFICANT. ;
*CHECK THAT NONE OF THE FINAL COVARIATES ARE CORRELATED;

PROC CORR DATA = PROJECT.FD3 SPEARMAN;
VAR SEX AGE4 CIGS4 SPF4 FVC4;
RUN;

*WE FIND THAT NO CORRELATION COEFFICIENTS ARE GREATER THAN .6 THEREFORE THEY ARE NOT COLLINEAR
*WE FIND THAT SPF4 AND SEX HAVE A SPEARMAN CORRELATION COEFFICIENT OF .582 - SO THEY ARE HIGHLY ASSOCIATED, BUT
*NOT COLLINEAR.;

PROC PRINT DATA = PROJECT.FD1 (OBS = 10);
RUN;

*CHECK FINAL MODEL SELECTION WITH FORWARD AND STEPWISE ON DATA THAT HAS NOT HAD ANY VARIABLES REMOVED;
PROC PHREG DATA = PROJECT.FD1;
MODEL SURV*DTH(0) = SEX AGE4 CHOL4 CIGS4 SPF4 DPF4 WGT4 FVC4 BMI4 HTN4 SMOKE / RL TIES = EFRON
SELECTION = STEPWISE INCLUDE = 1 SLENTRY = 0.15 SLSTAY = 0.15;
RUN;
*KEEP SEX, AGE4, CIGS4, SPF4, FVC4;

PROC PHREG DATA = PROJECT.FD1;
MODEL SURV*DTH(0) = SEX AGE4 CHOL4 CIGS4 SPF4 DPF4 WGT4 FVC4 BMI4 HTN4 SMOKE / RL TIES = EFRON
SELECTION = FORWARD INCLUDE = 1 SLENTRY = 0.15;
RUN;
*KEEP SEX, AGE4, CIGS4, SPF4, FVC4;

*CONCLUSION - FINAL MODEL WILL INCLUDE SEX, AGE4, CIGS4, SPF4, FVC4;
PROC PRINT DATA = PROJECT.FD3 (OBS = 10);
RUN;



*MAKE A NEW DATASET WITH ONLY PREDICTORS FROM FINAL MODEL;
DATA PROJECT.FD4;
SET PROJECT.FD3 (DROP = CHOL4 WGT4 BMI4 HTN4);
RUN;

PROC PRINT DATA = PROJECT.FD4 (OBS = 10);
RUN;

proc means data=project.fd4 nmiss;
run;

*Check that I'm deleting values correctly between two methods;
data project.clean1;
set project.fd4 (drop = CHD CHD_SURV);
if cmiss(of _all_) then delete;
run;

proc sql;
select count(*) as row_count
from project.clean1;
quit;
```

```
**Delete missing data for all of the chosen covariates**;
data project.clean;
set project.fd4;
if not missing(SEX) and not missing(AGE4) and not missing(CIGS4) and not missing(SPF4) and not missing(FVC4)
and not missing(DTH) and not missing(SURV); /*keep rows where all of the covariates in the model are not missing*/
run;

proc print data = project.clean;
run;

proc sql;
select count(*) as row_count
from project.clean;
quit;
*looks like both ways gives 1964 rows;

*CHECK PH ASSUMPTION FOR FINAL MODEL;
PROC PHREG DATA = PROJECT.CLEAN ZPH;
MODEL SURV*DTH(0) = SPF4 SEX AGE4 CIGS4 FVC4 / RL TIES = EFRON;
RUN;
*AGE FAILS THE PH ASSUMPTION;
*SINCE AGE IS NOT THE PREDICTOR OF INTEREST, AND IT IS A CONTINUOUS VARIABLE,
*WE WILL CREATE A STRATA AND DO STRATIFIED ANALYSIS;

*visualize the age variable using a histogram;
proc sgplot data = project.clean;
histogram AGE4;
run;


PROC PHREG DATA = PROJECT.CLEAN ZPH;
MODEL SURV*DTH(0) = SPF4 / RL TIES = EFRON;
STRATA AGEGRP;
RUN;
*FAIL TO REJECT NULL, SO NO DEPARTURE FROM PH ASSUMPTION;
PROC PHREG DATA = PROJECT.CLEAN;
MODEL SURV*DTH(0) = SPF4 / RL TIES = EFRON;
STRATA agegrp;
RUN;


*WALD TEST OF SPECIFIC VARIABLE AND GLOBAL NULL HYPOTHESIS ARE CONSISTENT, SPF4 IS SIGNIFICANTLY ASSOCIATED
*WITH MORTALITY. WITH EVERY UNIT INCREASE IN SPF4, THE HAZARD OF DEATH INCREASES BY 1.014 UNITS within the same
* age stratum. Note that we cannot compare hazard of death based on SPF4 unit increases for subjects in
* different age strata;

/* QUESTION 2 - WHAT FACTORS CAN CONFOUND THE ASSOCIATION BETWEEN SPF4 AND MORTALITY */
*bivariate analysis for confounding with the predictors included in the final model*;

proc phreg data = project.clean;
model SURV*DTH(0) = SPF4 / rl ties = efron;
STRATA agegrp;
run;
*crude HR for death comparing 1 unit increase in SPF4 is 1.014 within the same age stratum.;

proc phreg data = project.clean;
model SURV*DTH(0) = SPF4 SEX / rl ties = efron;
STRATA agegrp;
run;
*adjusted HR for death comparing 1 unit increase in SPF4 is 1.016 within the same age stratum;
```

```
proc phreg data = project.clean;
model SURV*DTH(0) = SPF4 AGE4 / rl ties = efron;
STRATA agegrp;
run;
*adjusted HR for death comparing 1 unit increase in SPF4 is 1.013 within the same age stratum;

proc phreg data = project.clean;
model SURV*DTH(0) = SPF4 CIGS4 / rl ties = efron;
STRATA agegrp;
run;
*adjusted HR for death comparing 1 unit increase in SPF4 is 1.015 within the same age stratum;

proc phreg data = project.clean;
model SURV*DTH(0) = SPF4 FVC4 / rl ties = efron;
STRATA agegrp;
run;
*adjusted HR for death comparing 1 unit increase in SPF4 is 1.015 within the same age stratum;

proc phreg data = project.clean;
model SURV*DTH(0) = SPF4 SEX AGE4 CIGS4 FVC4 / rl ties = efron;
STRATA agegrp;
run;
*adjusted HR for death comparing 1 unit increase in SPF4 is 1.014 within the same age stratum;

*CONCLUSION - no confounding by any of the covariates in the final model;

/* Question 3 - Is the association between SPF4 and mortality the same in males and females? */

proc print data = project.clean (obs = 10);
run;

proc phreg data = project.clean;
class SEX (REF = "1");
model SURV*DTH(0) = SPF4|SEX AGE4 CIGS4 FVC4 / RL TIES = EFRON;
HAZARDRATIO SPF4 / AT (SEX = ALL);
STRATA AGEGRP;
RUN;

*CONCLUSION: WE FIND THAT THE INTERACTION TERM BETWEEN SPF4 AND SEX IS NOT SIGNIFICANT AT A SIGNIFICANCE
* LEVEL OF .05 (CHISQ = 0.0435 AND P = .8347 ON 1 DF), AFTER ADJUSTING FOR AGE4, CIGS4, AND FVC4. THEREFORE,
* WE CONCLUDE THAT THE HAZARD OF DEATH FOR 1 UNIT INCREASE IN SPF4 DOES NOT CHANGE BASED ON SEX. THE HAZARD
* OF DEATH FOR 1 UNIT INCREASE IN SPF4 INCREASES BY 1.014 IN BOTH MALES AND FEMALES (95% CI = 1.009,1.019)

/* Question 4 - is CHD associated with mortality? */

*Create new dataset excluding rows where there are missing values for CHD, SURV, and DTH;

proc print data = project.fdl (obs = 10);
run;

data project.chd;
set project.fdl;
IF NOT MISSING (CHD) AND NOT MISSING (DTH) AND NOT MISSING (SURV); /* KEEP ROWS WITHOUT MISSING VALUES FOR CHD, SURV,
DTH */
RUN;

proc print data = project.chd (obs = 100);
run;
```

```
proc lifetest method = KM plot = survival data = project.chd;
time SURV*DTH(0);
STRATA CHD;
RUN;

proc phreg data = project.chd zph;
model SURV*DTH(0) = CHD / rl ties = efron;
run;
```