Homework 4

Elana Nelson

April 22, 2019 at 11:59pm

We consider in this homework the relationship of cardiometabolic risk factors with the occurrence of colorectal cancer in the Physicians' Health Study. The data contain information on 13 variables in a sample of 16,018 participants in the Physicians' Health Study. These participants were randomized in 1982-1983 and followed until they died, dropped out, developed colorectal cancer, or until 12/31/1995.

Variables are:

Name	Description
age	Age in years at time of Randomization
asa	0 - placebo, 1 - aspirin
bmi	Body Mass Index (kg/m^2)
hypert	1 - Hypertensive at baseline, 0 - Not
alcohol	0 - less than monthly, 1 - monthly to less than daily, 2 - daily consumption
dm	0 = No diabetes Mellitus, 1 - diabetes Mellitus
sbp	Systolic BP (mmHg)
exer	0 - No regular, 1 - Sweat at least once per week
csmoke	0 - Not currently, 1 - $<$ 1 pack per day, 2 - $\geq$ 1 pack per day
psmoke	0 - never smoked, 1 - former $<$ 1 pack per day, 2 - former $\geq$ 1 pack per day
pkyrs	Total lifetime packs of cigarettes smoked
crc	0 - No colorectal Cancer, 1 - Colorectal cancer
cayrs	Years to colorectal cancer, or death, or end of follow-up.

1. Fit a Poisson regression model that predicts the incidence of CRC with the following independent variables: continuous age, continuous SBP, pack-years of cigarettes consumed, continuous BMI and an indicator of diabetes mellitus. Provide interpretation of the relationship of SBP with rates of CRC from this model, based on comparison of the adjusted relative rate in two men differing by 10 mmHg, with a relevant 95% confidence interval.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
age	1.068814	0.0068344	9.737337	0.0000000	1.0545735	1.083218
sbp	1.015298	0.0051549	2.945183	0.0032276	1.0050080	1.025517
pkyrs	1.007921	0.0025494	3.094875	0.0019690	1.0027334	1.012816
bmi	1.032369	0.0197568	1.612432	0.1068679	0.9919775	1.071772
dm	1.425564	0.2876446	1.232658	0.2177033	0.7711639	2.403583

For two men of the same age, pack years, bmi, and diabetes mellitus status, a 10 mmHg increase in sbp yiels a 10% increase in rate of CRC (1% increase for 1 mmHg increase). The confidence interval for this estimate is also relatively small suggesting a good estimate.

2. Obtain and plot Kaplan-Meier estimates of the probability of developing CRC among those with and without elevated SBP ($\geq$ 140 mmHg). Give estimated probabilities of developing CRC at 1, 5, and 10 years in each SBP group.

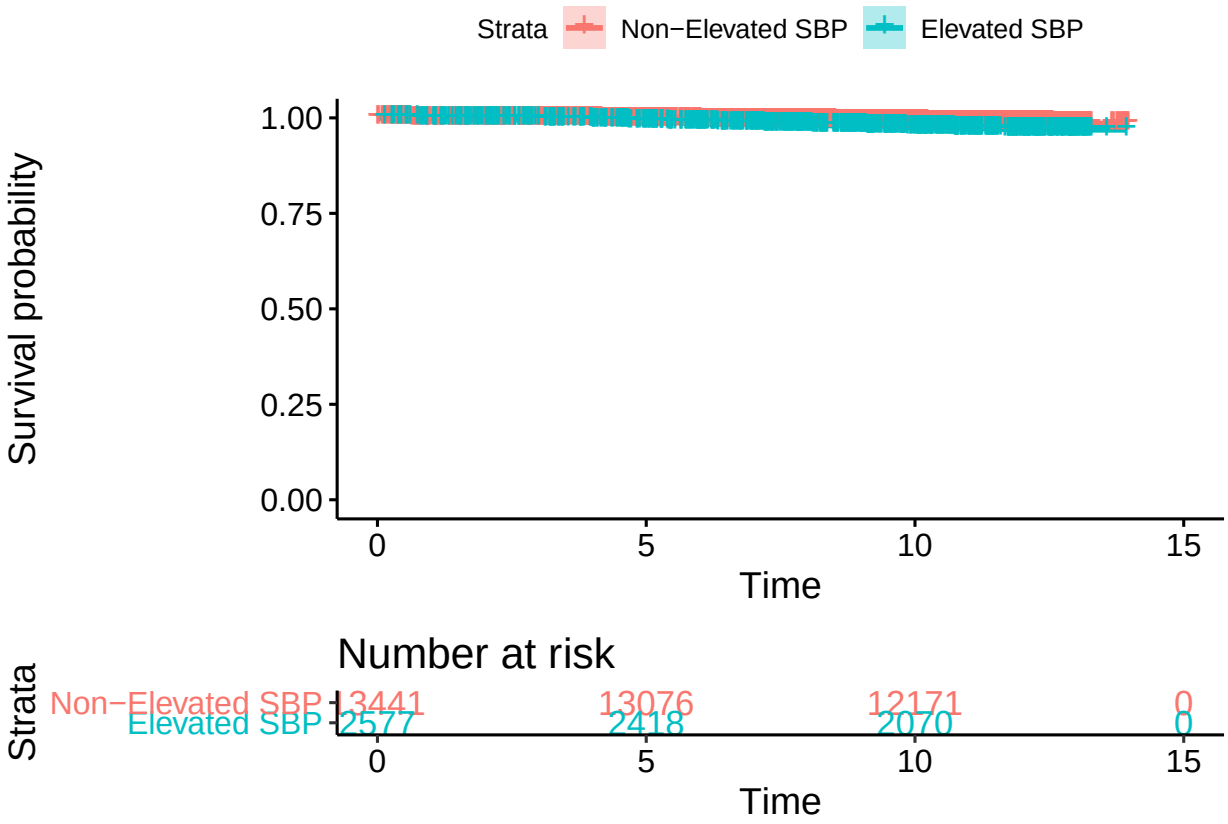


Table 3: Estimated probabilities of developing colorectal cancer at different times

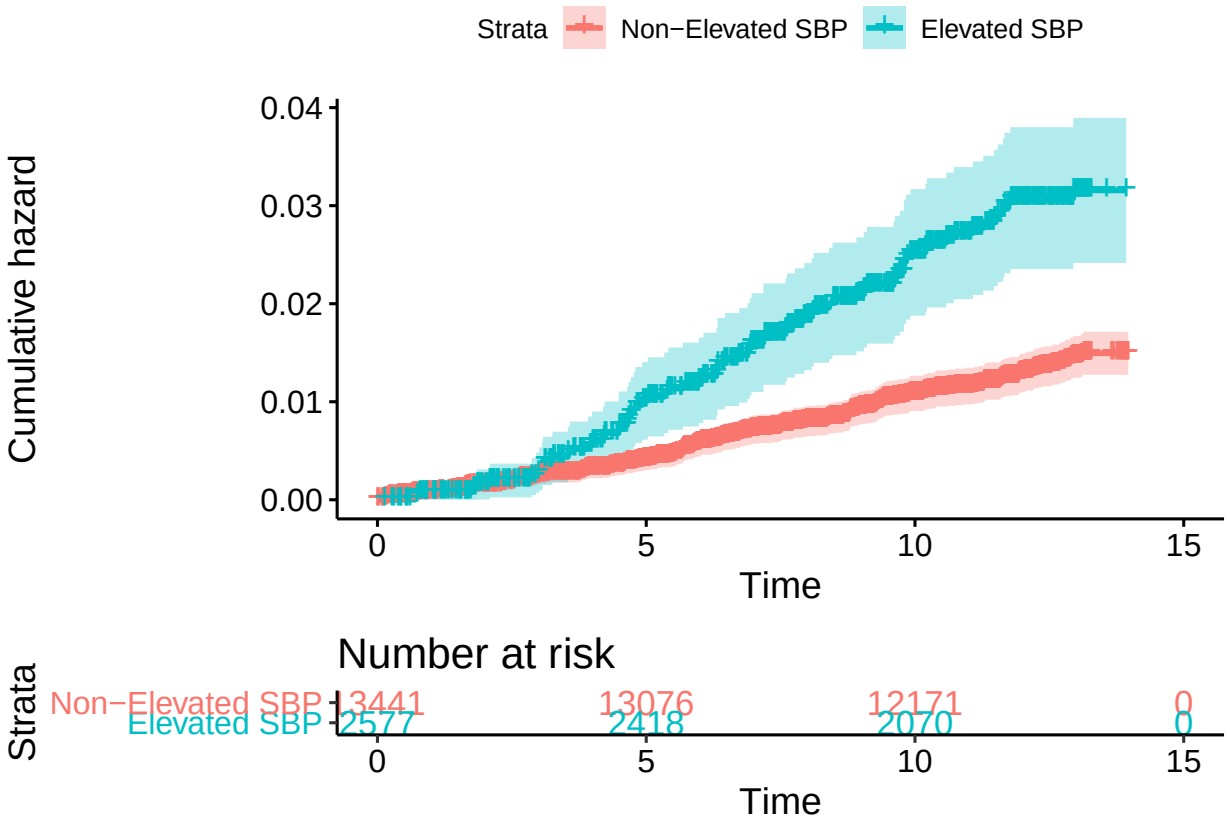
	Non-Elevated SBP Survival	Elevated SBP Survival
1 year	0.9992550	0.9992210
5 years	0.9959392	0.9899638
10 years	0.9892036	0.9750711

As we can see from the plot of the Kaplan-Meier estimates over time, not many people are developing CRC. As a result, the plot is very concentrated near a probability of survival = 1.

Furthermore, the direct output of survival by sbp group at different time steps also shows that in both groups, not only arre the survival rates similar but also very close to 1.

3. In question 2, you compared the probability of developing colorectal cancer at 3 different follow-up times (1, 5, and 10 years) between men with and without elevated SBP ($\geq$ 140 mmHg). Continue with this same dichotomous exposure variable, and perform the log-rank test to compare the hazard of colorectal cancer between the two groups of men. Make sure you state the null and alternative hypotheses.

The null hypothesis of the log-rank test is that there is no true difference between the two groups, while the alternative hypothesis supports a significant difference between the two groups.



Call: survdiff(formula = Surv(cayrs, crc) ~ sbp.high, data = phscrc)

n=16018, 16 observations deleted due to missingness.

```
                N Observed Expected (O-E)^2/E (O-E)^2/V
```

sbp.high=0 13441 183 215.6 4.93 32.7 sbp.high=1 2577 71 38.4 27.72 32.7

Chisq= 32.7 on 1 degrees of freedom, p= 1e-08

Our plot of the cumulative hazard indicates that the groups could be significantly different, as their cofidence intervals are not overlapping too much. Furthermore, they are not completely straight lines, indicating that Kaplan-Meier is a sound method to use. We can further test the difference between the two groups with a log rank test, shown above. By the log-rank test, we have a significant difference in the rate of developing crc by those men with or without elevated SBP.

4. Older age is a powerful risk factor for colorectal cancer that also influences SBP, and is therefore a probable confounder. Form four age groups (40-49, 50-59, 60-59, and 70-84 years) and perform a stratified log-rank test of whether elevated SBP ($\geq$ 140 mmHg) is related to the hazard of colorectal cancer, adjusted for age.

Call: survdiff(formula = Surv(cayrs, crc) ~ sbp.high + strata(age.cat), data = phscrc)

n=16010, 24 observations deleted due to missingness.

```
                N Observed Expected (O-E)^2/E (O-E)^2/V
```

sbp.high=0 13439 183 198.3 1.18 5.76 sbp.high=1 2571 71 55.7 4.21 5.76

Chisq= 5.8 on 1 degrees of freedom, p= 0.02

When adjusting for age, the log-rank test shows that the two groups are still significantly different (p-value = 0.02). Thus, there is a significant difference in the hazard of developing colorectal cancer between those

people with elevated or non-elevated SBP, adjusted for age.

5. Evaluate the shape of the relationship of SBP with rates of CRC, *adjusting for age* in a cox proportional hazards model. You can use either continuous age or the four categories described above. Create 4 models:

- **Model 1**: Use continuous SBP (We will refer to this as the simple linear model)
- **Model 2**: Use clinical cutpoints of SBP (<120, 120-129, 130-139, $\geq$ 140 mmHg)
- **Model 3**: Quartiles of the distribution
- **Model 4**: Linear plus quadratic term of SBP.

Does a simple linear model (on the log scale) adequately describe the relationship controlling for age? You can use Likelihood Ratio tests to compare Models 2-4 vs Model 1.

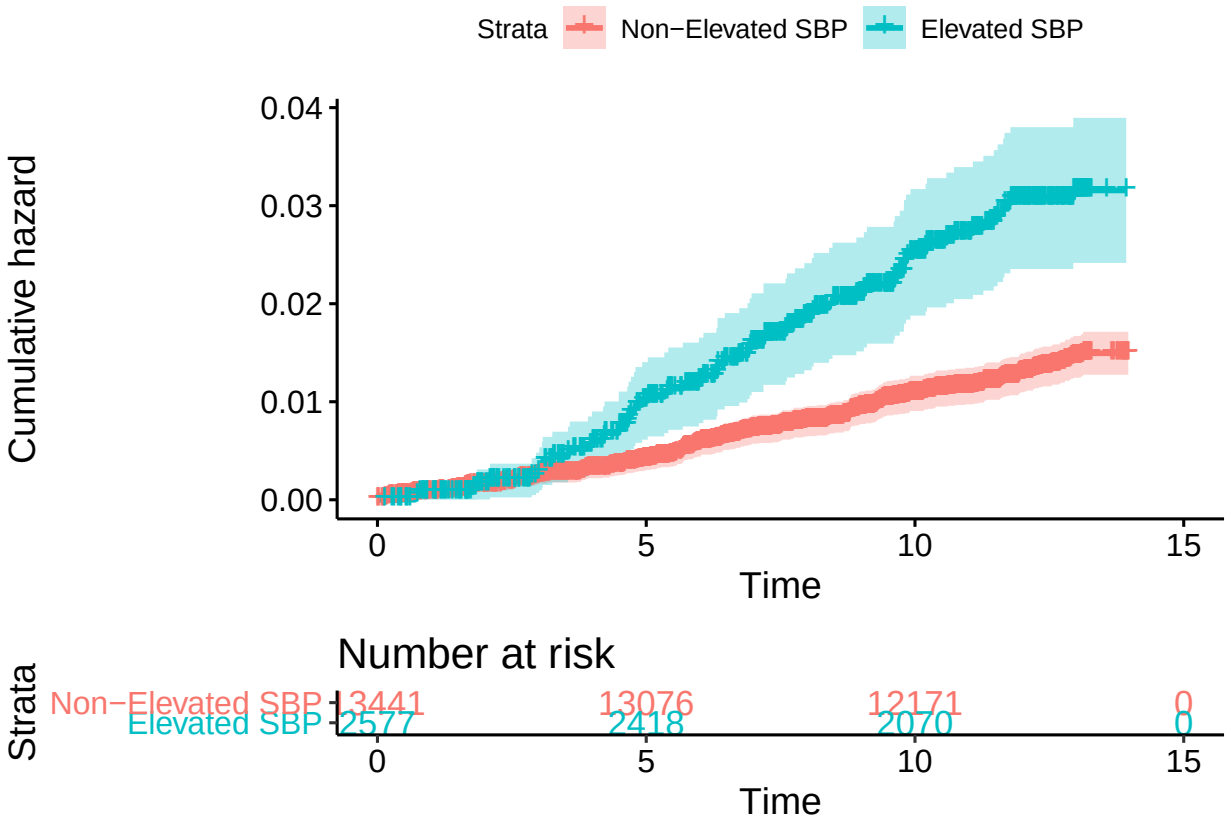
term	estimate	p.value	conf.low	conf.high
sbp	1.018	0.000	1.008	1.028
age	1.072	0.000	1.058	1.086
sbp.cat[120,130)	1.370	0.156	0.887	2.114
sbp.cat[130,140)	1.785	0.008	1.167	2.731
sbp.cat[140,200]	1.956	0.003	1.252	3.056
age	1.074	0.000	1.060	1.088
sbp.quart[121,126)	1.707	0.028	1.059	2.753
sbp.quart[126,133)	1.595	0.008	1.132	2.247
sbp.quart[133,200]	1.770	0.001	1.271	2.465
age	1.074	0.000	1.060	1.088
sbp	1.018	0.000	1.008	1.028
age	1.072	0.000	1.058	1.086

loglik	Chisq	Df	P(>|Chi|)
-2354.630	NA	NA	NA
-2354.919	0.5778919	2	0.7490527
-2353.934	1.9710161	0	0.0000000
-2354.630	1.3931242	2	0.4982954

Considering the table of 4 models (note that the cutoff for each model is at each successive age predictor), we see that age is significant across each model, meaning it is good that we adjusted for it. Each model (2-4) is nested within model 1, but just larger. Thus, we can test for a significant difference between these models with a likelihood ratio test, shown above. By the likelihood ratio test, we see that model 3 (sbp quartiles) is the only model that is significantly different from the reference model (simple linear model). In particular, model 3 is significantly *better* than model 1.

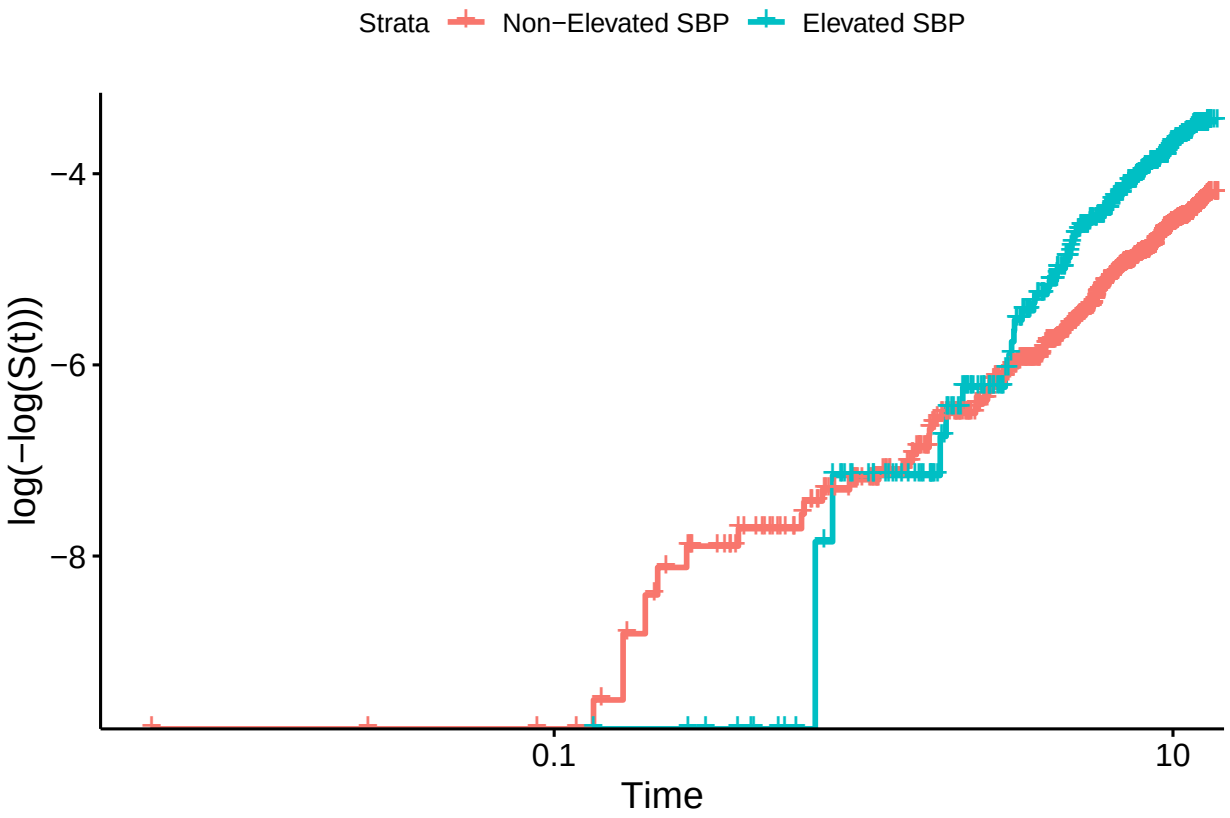
6. Which graphical approaches can you use to evaluate whether the assumptions of the proportional hazards appear to hold? Obtain these graphs and briefly interpret them.

Again, we refer to our plot of cumulative hazards for each group. If a straight line could be fit to this plot, then we could settle with a different model (perhaps poisson). If the groups are clearly not straight, then Kaplan-Meier survival model is a good way to go.



The two groups in this cumulative hazards plot are relatively straight, indicating that the cumulative hazard could be constantly increasing (derivative is constant), suggesting that an alternative model could be used (such as poisson). However, they are not completely straight, indicating that a survival model could also be used because the hazard is increasing over time. This moreso applies to the Elevated SBP group, which seems to have a cumulative hazard that is increasing more and more over time. The non-elevated SBP group seems pretty consistent. We can then check our proportional hazards assumption for the created survival model.

Under the proportional hazards assumption, we should see that log(-log()) of our Kaplan-Meier estimates from our original survival model are parallel over time.



We see that our two estimates eventually become relatively parallel, however there is a lot of overlapping in our groups, which is definitely not supportive of a proportional hazard model. We should use a hypothesis test next to check for proprtional hazards.

7. Perform a hypothesis test to evaluate the proportional hazards assumption.

The test will be performed on model 3, with SBP quartile groups, as it was previously shown to be a better fit of the data.

	rho	chisq	p
sbp.quart[121,126)	-0.0462761	0.5417025	0.4617280
sbp.quart[126,133)	-0.0002777	0.0000194	0.9964826
sbp.quart[133,200]	-0.0022004	0.0012207	0.9721286
age	-0.0586832	0.7391059	0.3899475
GLOBAL	NA	1.3714844	0.8491357

The table above shows only insignificant p values for all predictors in the model, meaning that we can reject the alternative hypothesis in favor of the null hypothesis that the hazard rate over time for the individuals are relatively constant. Our proportional hazards assumption holds, indicating that the use of a Cox Proportional Hazards model is sound.