FINAL EXAMINATION
Analysis of survival data (7.5hp) 2018-01-10

INFORMATION:

A. Allowed means of assistance:
Book, handouts from lectures (where handwritten notes are allowed), Declaration on professional ethics, calculator, ruler, dictionary.

B. Writing time: 9.00-14.00 (5 hours).

C. The examination consists of 2 tasks, for a total of 80 points. Including any attendance and home assignment bonus points, at least 48 points are needed to pass the course (G), and 72 points to pass with distinction (VG).

D. For every task the maximum score is shown (for every part of the task). Sometimes the parts cannot be judged independent of each other, which means that points might not be able to be set for a later part if the previous part has not been solved in a correct way (in principle). Negative points will never be set.

E. You can write your answers in English or Swedish.

F. If you desire clarification regarding the test, especially the wording of a problem, then please contact an examination proctor. The examination proctors can contact the responsible teacher.

G. After turning in your test, you will keep the test pages with the question statements (not to be handed in!). Preliminary solutions will be posted at Studentportalen.

INSTRUCTIONS:

A. Follow the instructions on the front page to be stapled to your solutions. E.g., the solutions for each task should be started on a new sheet.

B. Present all your solutions in a way that makes it easy to follow your way of thinking! What is unclearly presented is assumed to be unclearly thought. Motivate all important steps of your solution, including any assumptions that need to be fulfilled (and check if they are).

C. When constructing confidence intervals you must (besides what is presented in B above) state what the interval is intended to cover, and present the formula for the interval before you present the calculation (if needed), and interpret the calculated interval.

D. When performing hypothesis testing you must (besides what is presented in B above) present null and alternative hypotheses, choice of significance level, choice of test, *P*-value, result, and conclusion.

Good luck!

(27) **Task 1**

Hulegårdh et al (2014) reports data on patients with acute lymphoblastic leukemia (ALL) during the period 2000-2009 in Sweden.

Reference: E. Hulegardh et al (2014). Outcome after HSCT in Philadelphia chromosome positive acute lymphoblastic leukemia in Sweden: a population-based study. Medical Oncology, Vol. 31:66, No. 8, 2014.

This task is based on a fictitious follow-up of the study above, including a total of 408 patients.

All patients underwent stem cell transplantation and the interest of the study is to investigate the relationship between chronic graft versus host disease (explained below) and survival time.

The variables represented in the dataset are as follows:

surv_time: Time from stem cell transplant to death or on-study time (months)

status: Event indicator (0 = alive, 1 = death in relapse, 2 = death without relapse)

cGVHD: Chronic graft versus host disease (1 = yes, 0 = no), a disease that patients can get after a stem cell transplantation where immune cells (white blood cells) in the transplanted tissue (the graft) recognize the recipient (the host) as "foreign". The transplanted immune cells then attack the host's body cells.

The normal progress of ALL after transplantation is to get a relapse after some time, followed by death (eventually). The theoretical reasoning around cGVHD is that patients have a higher risk of dying without first getting a relapse, and the event of interest is therefore death without relapse.

Use the SAS output on the following pages, and answer the questions below.
NOTE that output from two different methods is presented.

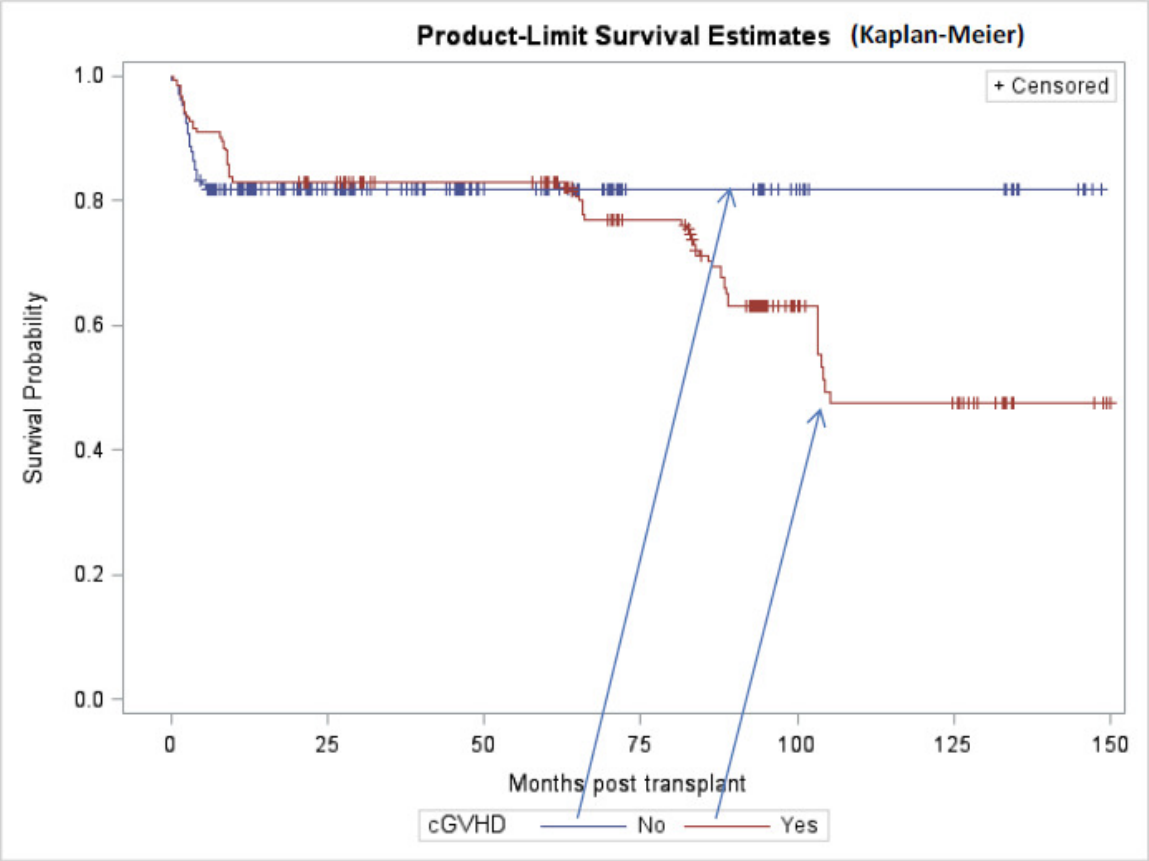
(2) **A** Which of the two presented methods is appropriate in this case? Motivate your choice.

(8) **B** Present, interpret, and compare for the two cGHVD groups:
The 25 percentile, median (50 percentile), and 75 percentile non-relapse survival times (approximate values). If the measures cannot be estimated, explain why and present the minimum value of the measures.

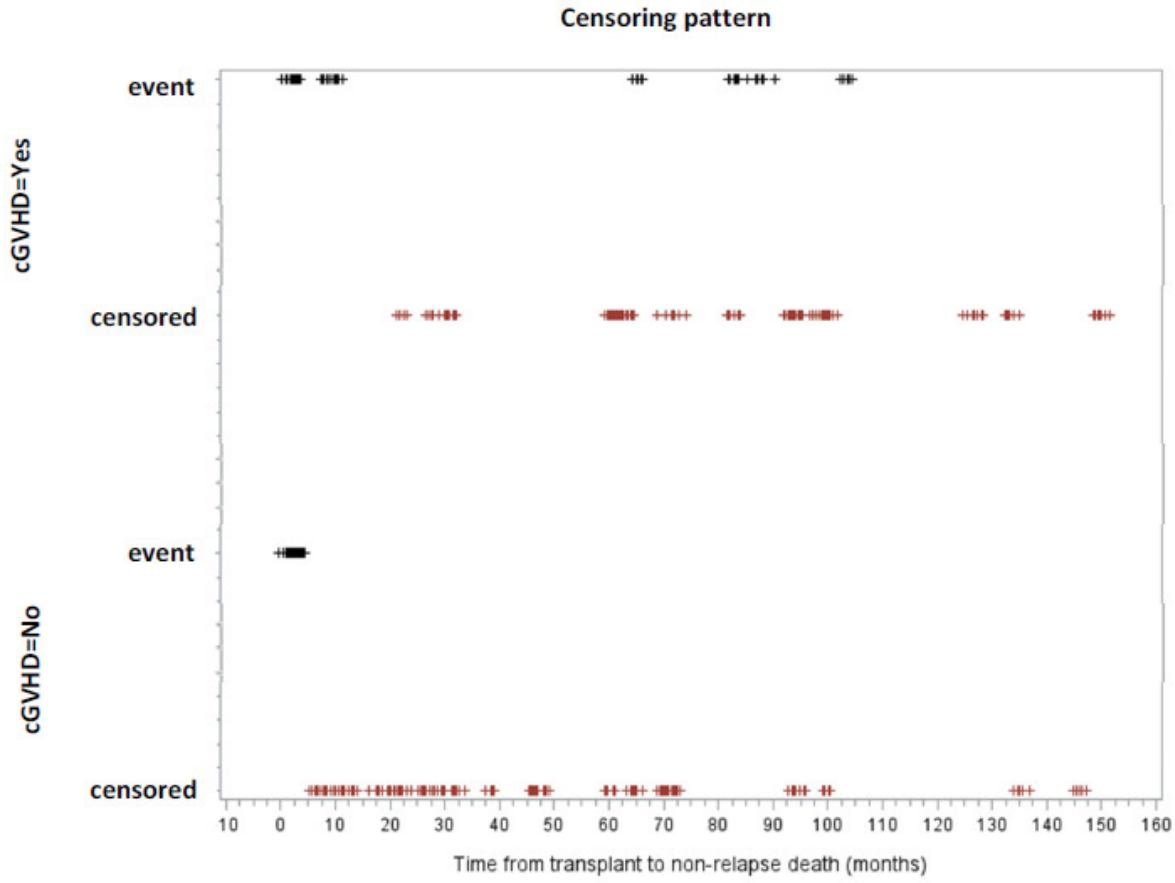
(4) **C** Estimate and compare the probability of non-relapse death at 50 months and 100 months after transplant for the two cGVHD groups.

(13) **D** Is there a significant difference in the risk of non-relapse death between the two groups? Perform a hypothesis test to support your answer. Remember to follow the instructions on
the front page.

Method 1



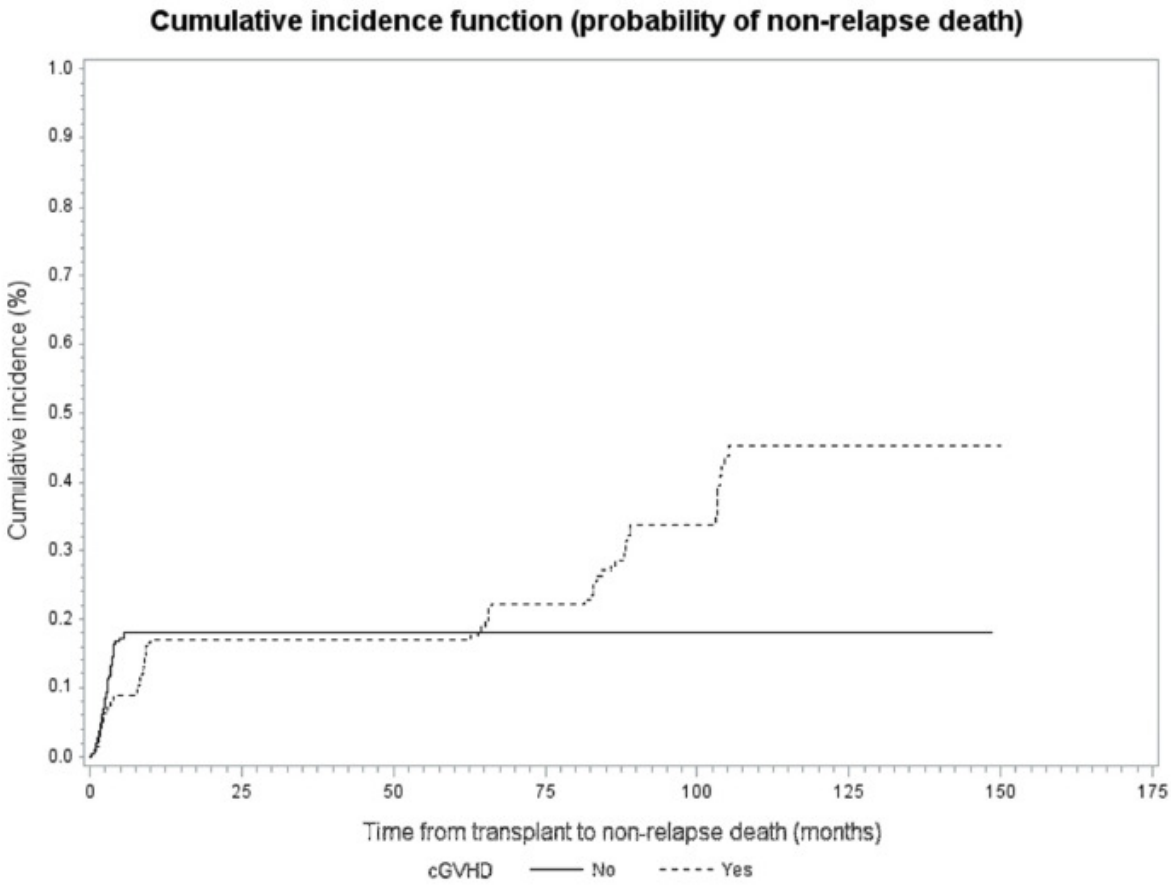
Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	2.3707	1	0.1236
Wilcoxon	0.0062	1	0.9374
-2Log(LR)	0.1305	1	0.7179



Summary of the Number of Censored and Uncensored Values					
Stratum	cGVHD	Total	Failed	Censored	Percent Censored
1	No	215	39	176	81.86
2	Yes	193	65	128	66.32
Total		408	104	304	74.51

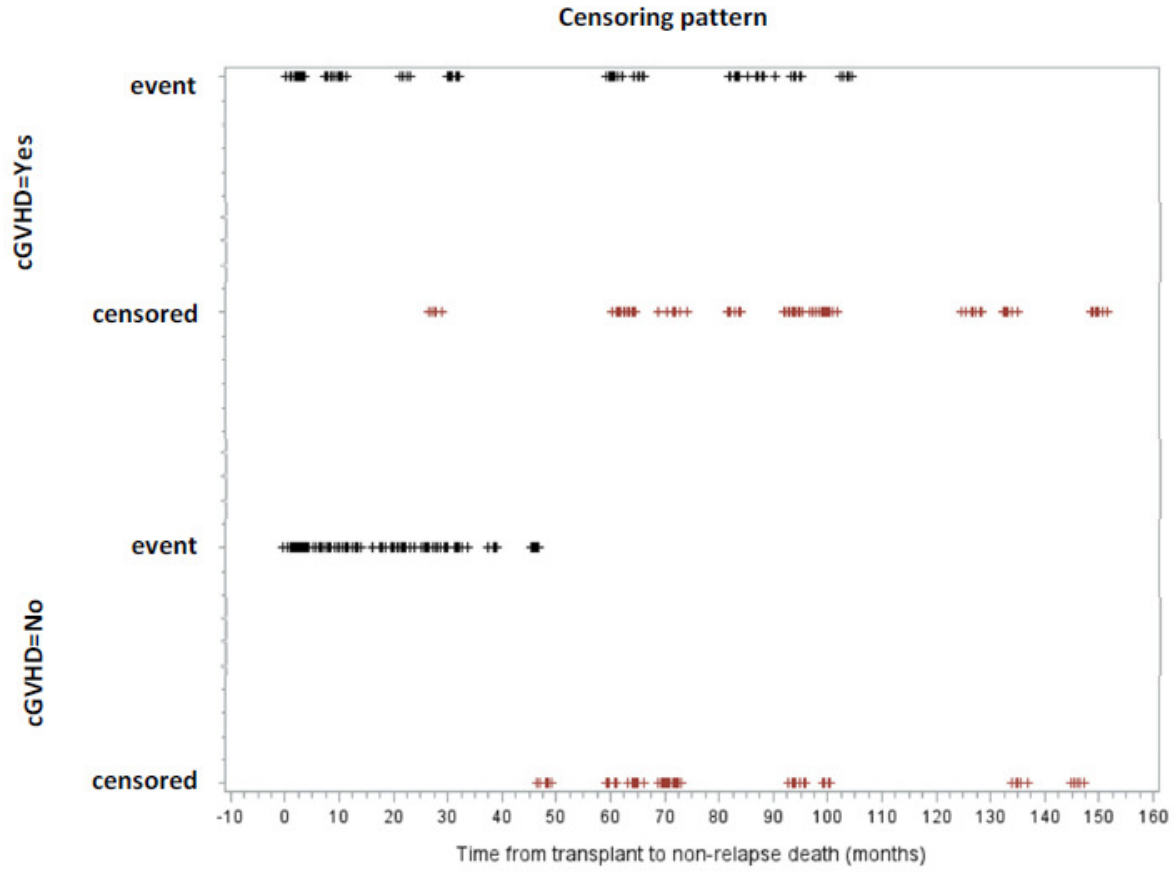
"Alive" or "Death in relapse" is being censored

Method 2



Gray's Test for Equality of Cumulative Incidence Functions

Chi-Square	DF	Pr > Chi-Square
6.69053	1	0.0097



Frequency Row Pct	Table of cGVHD by status			
	status			
cGVHD	0	1	2	Total
No	72	104	39	215
	33.49	48.37	18.14	
Yes	96	32	65	193
	49.74	16.58	33.68	
Total	168	136	104	408

"Alive" is being censored

END OF TASK 1

(43) **Task 2**

Rossi, Berk, and Lenihan (1980) present data from an experimental study of recidivism of 432 male prisoners, who were observed for a year after being released from prison.

Reference: Rossi, P. H., R. A. Berk & K. J. Lenihan. 1980. Money, Work and Crime: Some Experimental Results. New York: Academic Press

The researchers investigated the relationship between a number of covariates and time to rearrest. The following variables are included in the data:

time: time of first arrest after release, or censoring time (weeks).

arrest: event indicator (1 = arrested during the period of the study, 0 = not arrested).

fin_aid: financial aid indicator (1 = the individual received financial aid after release from prison, 0 = no financial aid); financial aid was a randomly assigned factor manipulated by the researchers.

age: age at the time of release (years.

married: married at the time of release (1 = married, 0 = not married).

prior_conv: number of prior convictions.

time_job: time to first job or time on study (weeks).

job: indicator of whether an individual manages to get a job or not (1 = job, 0 = no job). *This defenition was missing in the original exam.*

(2) **A** Fitting a Cox regression model to the data above, which method of handling ties would you choose and why?

(2) **B** Which method of constructing confidence intervals for the hazard ratios would you choose and why?

(2) **C** The effect of whether the individual gets a job after release depends on the time it takes before he/she gets a job (if ever). Which of the following options to handle the job covariate is correct? Motivate your choice.

i)
```
proc phreg data=jail;
model time*arrest(0)=fin_aid age married prior_conv time_job;
run;
```

ii)
```
proc phreg data=jail;
model time*arrest(0)=fin_aid age married prior_conv time_job job;
run;
```

iii)
```
proc phreg data=jail;
model time*arrest(0)=fin_aid age married prior_conv Zjob;
if time_job<=time and job=1 then Zjob=1; else Zjob=0;
run;
```

iv)
```
proc phreg data=jail;
model time*arrest(0)=fin_aid age married prior_conv
time_job job Zjob;
if time_job<=time and job=1 then Zjob=1; else Zjob=0;
run;
```

D-F Based on the output from SAS proc PHREG on the next pages, answer the questions below.

NOTE that five different models are presented at the end of the output, and that the models include all job components as listed in option *iv)* above. If you selected any other option above, ignore the parts of the output representing the components not included in your chosen option.

(22) **D** Which of the five models below would you choose to analyze the data? Motivate your choice carefully.

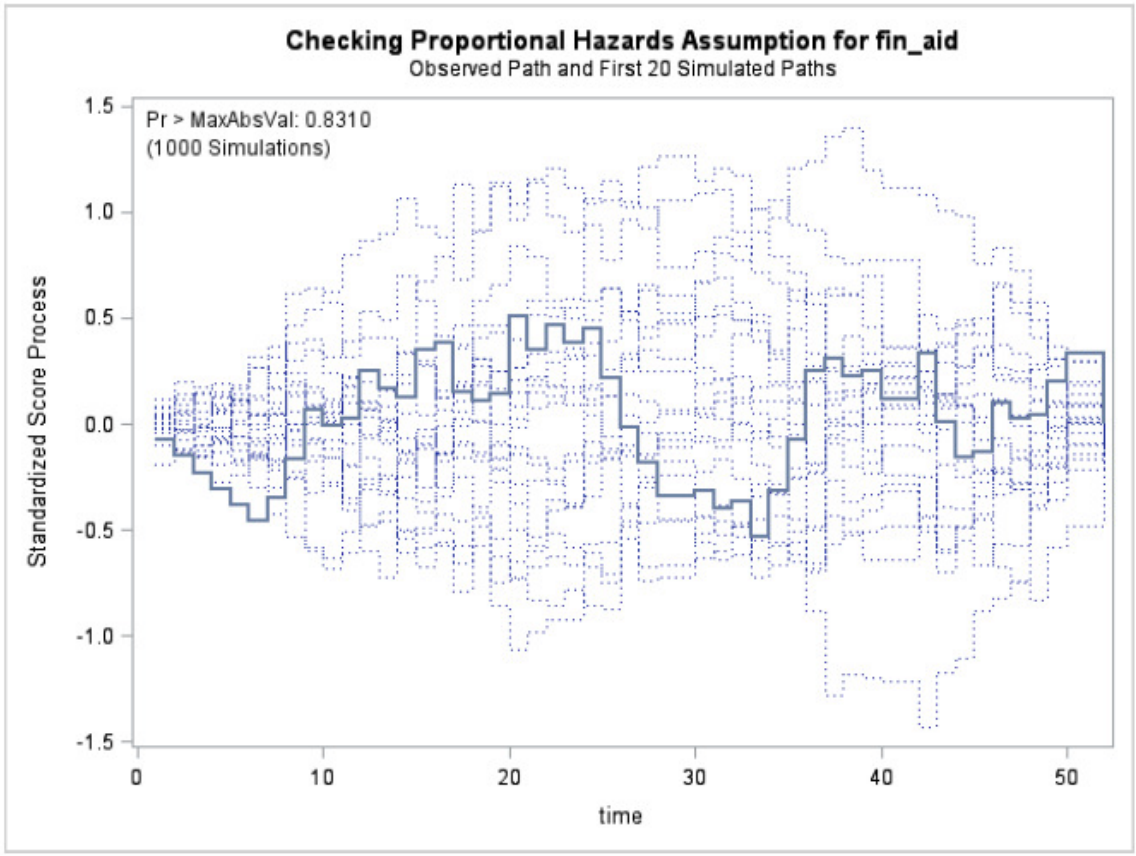
(12) **E** Explain/interpret the relationships between the covariates in the model you chose in D above and time to rearrest.

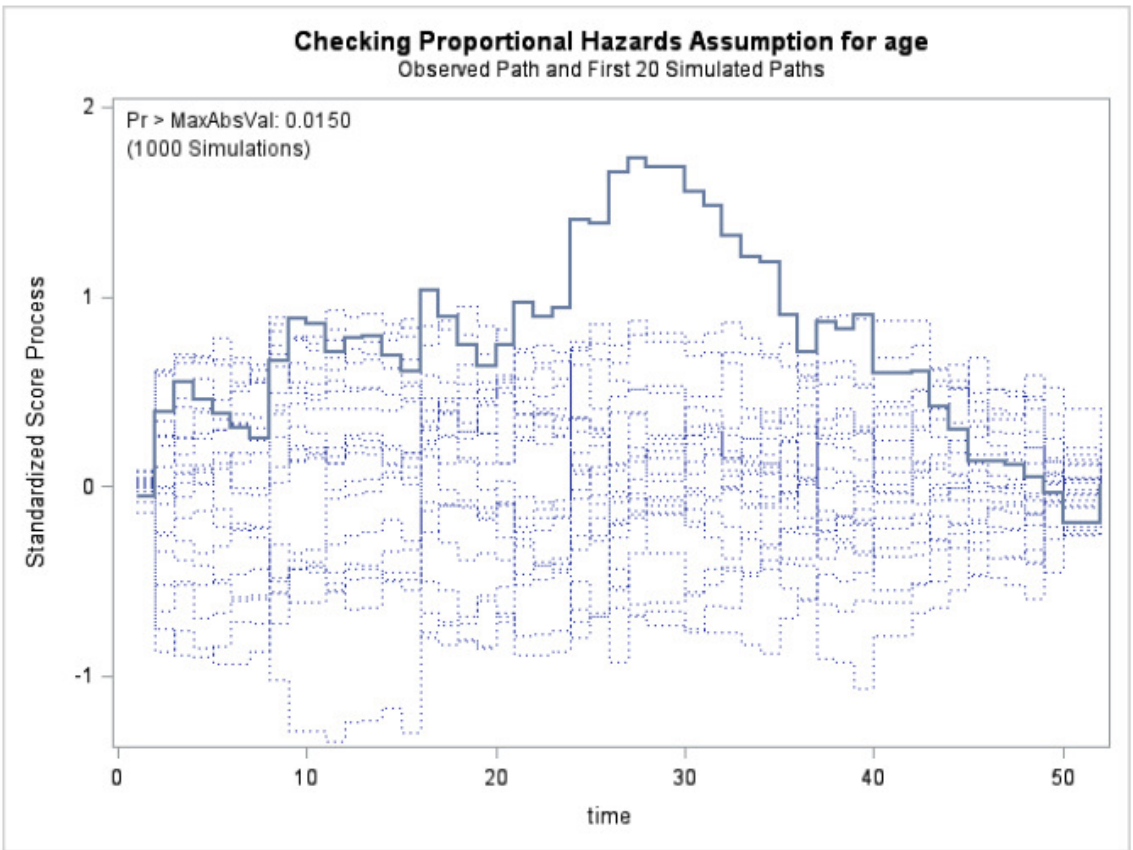
(3) **F** Calculate and interpret the relative risk of arrest for individuals with 10 more prior convictions than others.

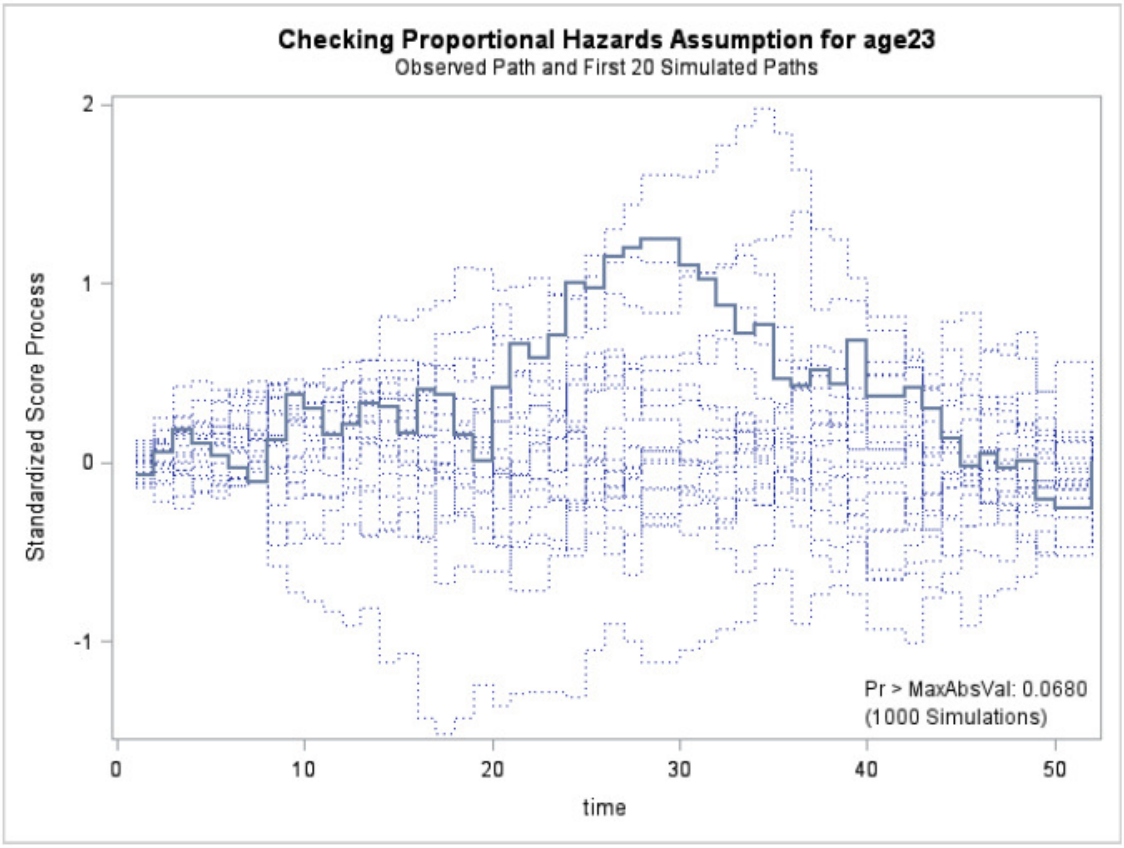
Analysis of Maximum Likelihood Estimates

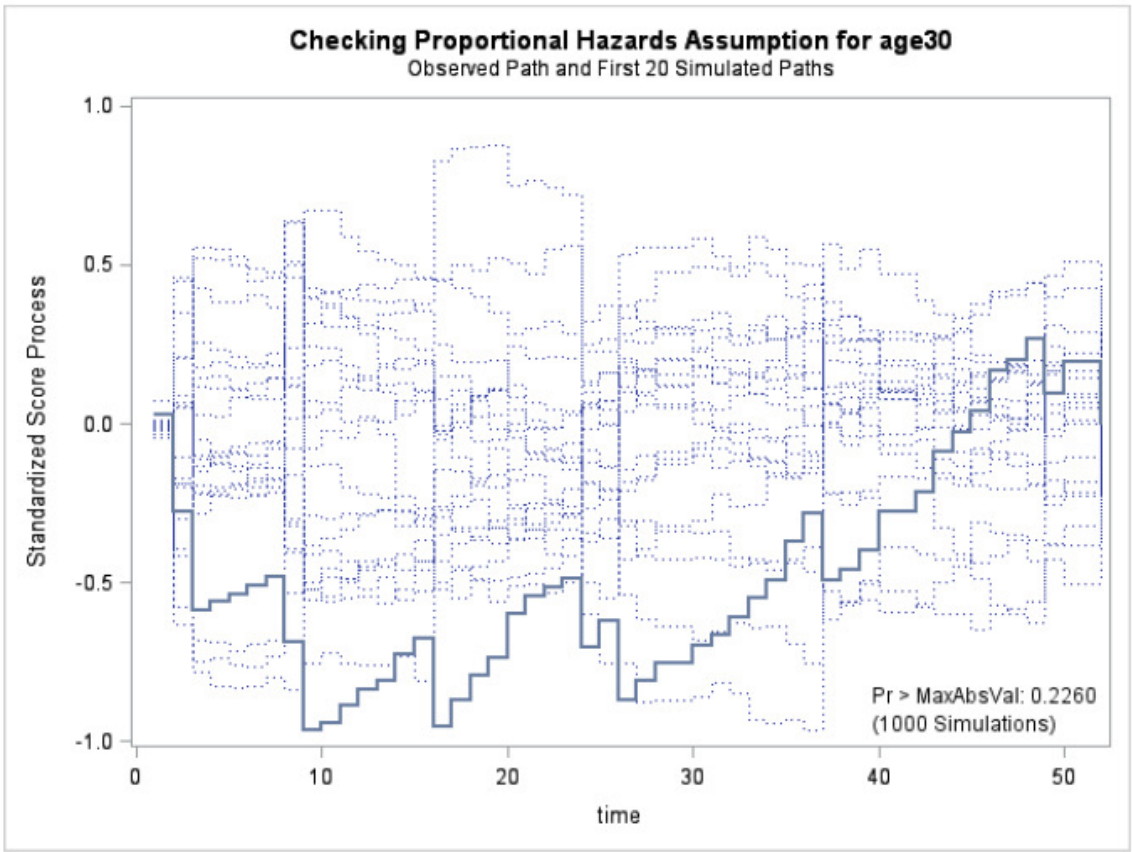
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
fin_aid	1	-0.74825	0.86067	0.7558	0.3846	0.473
lnt_fin_aid	1	0.11955	0.26413	0.2049	0.6508	1.127
age	1	0.10983	0.06610	2.7604	0.0966	1.116
lnt_age	1	-0.06053	0.02204	7.5418	0.0060	0.941
prior_conv	1	0.10130	0.10859	0.8702	0.3509	1.107
lnt_prior_conv	1	-0.0000985	0.03432	0.0000	0.9977	1.000
married	1	-3.32710	2.56290	1.6853	0.1942	0.036
lnt_married	1	0.78548	0.73707	1.1357	0.2866	2.193
age23	1	-0.26090	0.84136	0.0962	0.7565	0.770
lnt_age23	1	0.28247	0.26034	1.1772	0.2779	1.326
age30	1	-1.55607	1.06994	2.1151	0.1459	0.211
lnt_age30	1	0.80980	0.36206	5.0026	0.0253	2.247

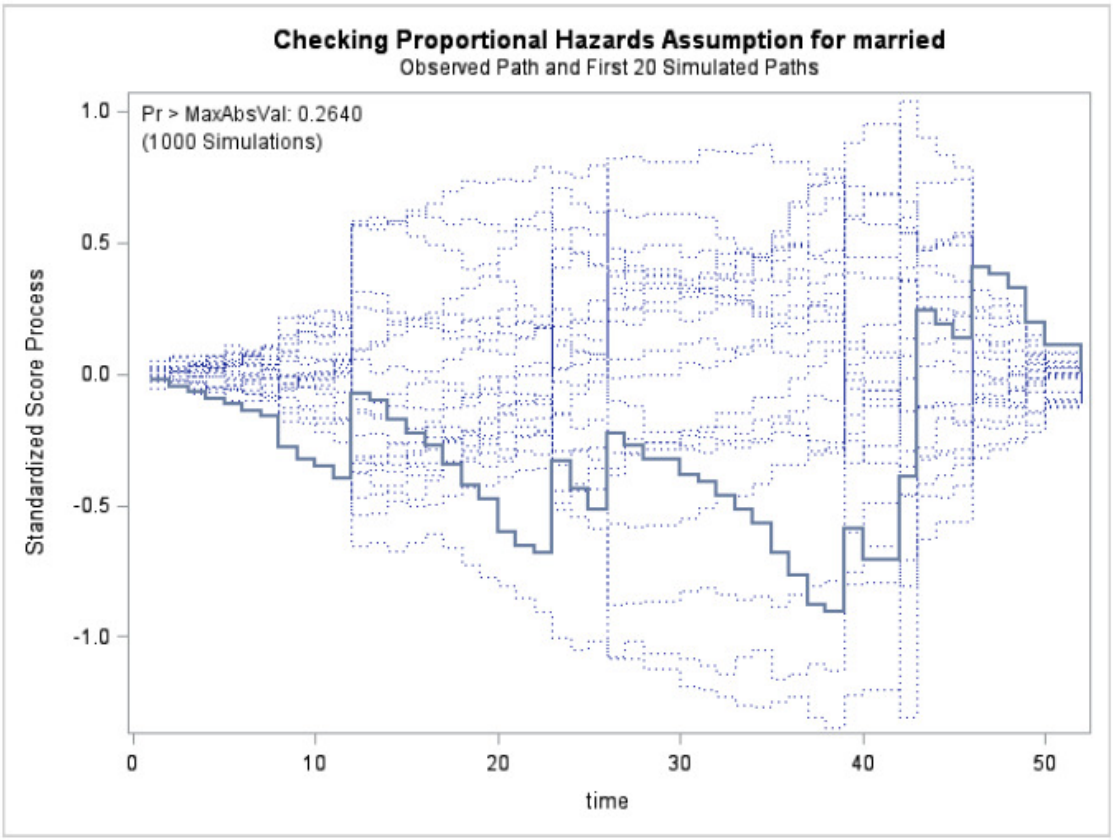
"lnt_x" denotes a time-dependent covariate, calculated as x*log(time)
age23=1 if age<23, 0 otherwise
age30=1 if age<30, 0 otherwise

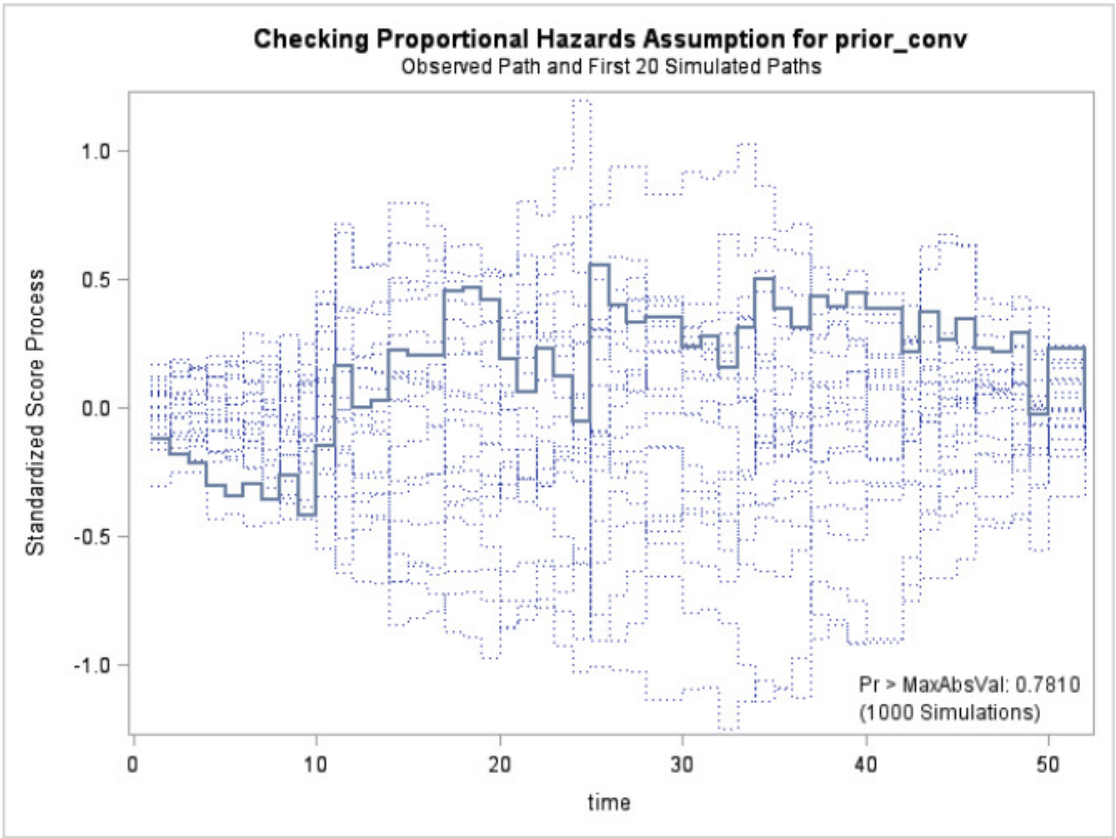

Checking Proportional Hazards Assumption for fin_aid
Observed Path and First 20 Simulated Paths
Pr > MaxAbsVal: 0.8310
(1000 Simulations)
Standardized Score Process
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
10
20
30
40
50
time

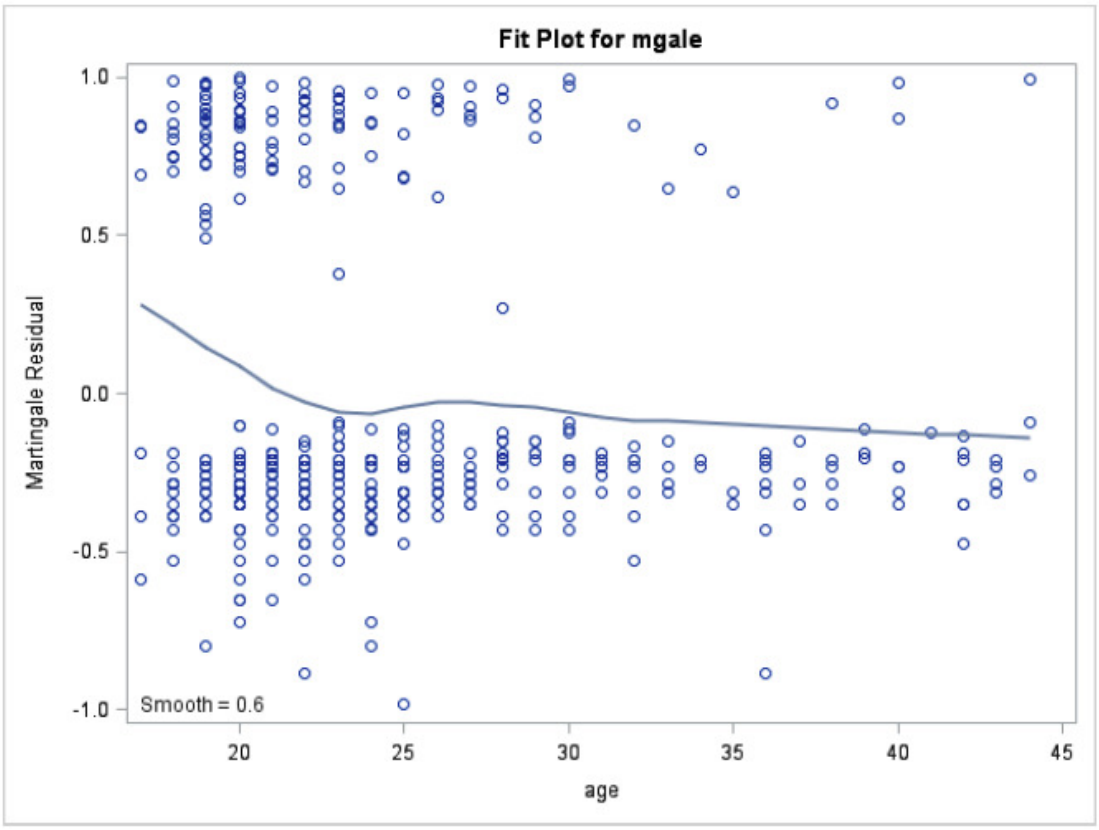

Checking Proportional Hazards Assumption for age
Observed Path and First 20 Simulated Paths
Pr > MaxAbsVal: 0.0150
(1000 Simulations)
Standardized Score Process
2
1
0
-1
0
10
20
30
40
50
time


Checking Proportional Hazards Assumption for age23
Observed Path and First 20 Simulated Paths
Standardized Score Process
2
1
0
-1
0
10
20
30
40
50
time
Pr > MaxAbsVal: 0.0680
(1000 Simulations)


Checking Proportional Hazards Assumption for age30
Observed Path and First 20 Simulated Paths
Standardized Score Process
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
time
Pr > MaxAbsVal: 0.2260
(1000 Simulations)


Checking Proportional Hazards Assumption for married
Observed Path and First 20 Simulated Paths
Pr > MaxAbsVal: 0.2640
(1000 Simulations)
Standardized Score Process
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
time

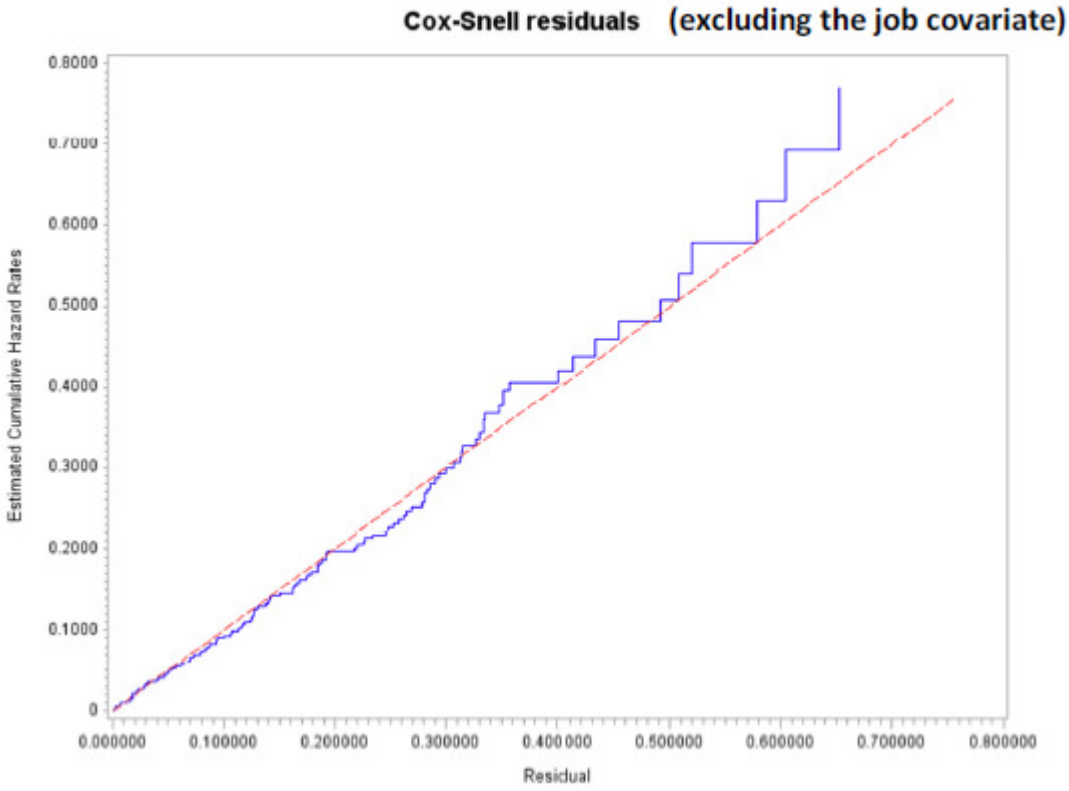

Checking Proportional Hazards Assumption for prior_conv
Observed Path and First 20 Simulated Paths
Standardized Score Process
1.0
0.5
0.0
-0.5
-1.0
Pr > MaxAbsVal: 0.7810
(1000 Simulations)
0
10
20
30
40
50
time


Fit Plot for mgale
Martingale Residual
1.0
0.5
0.0
-0.5
-1.0
Smooth = 0.6
20
25
30
35
40
45
age

Model 1: age as a continuous covariate

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1227.506	1191.894
AIC	1227.506	1201.894
SBC	1227.506	1215.575

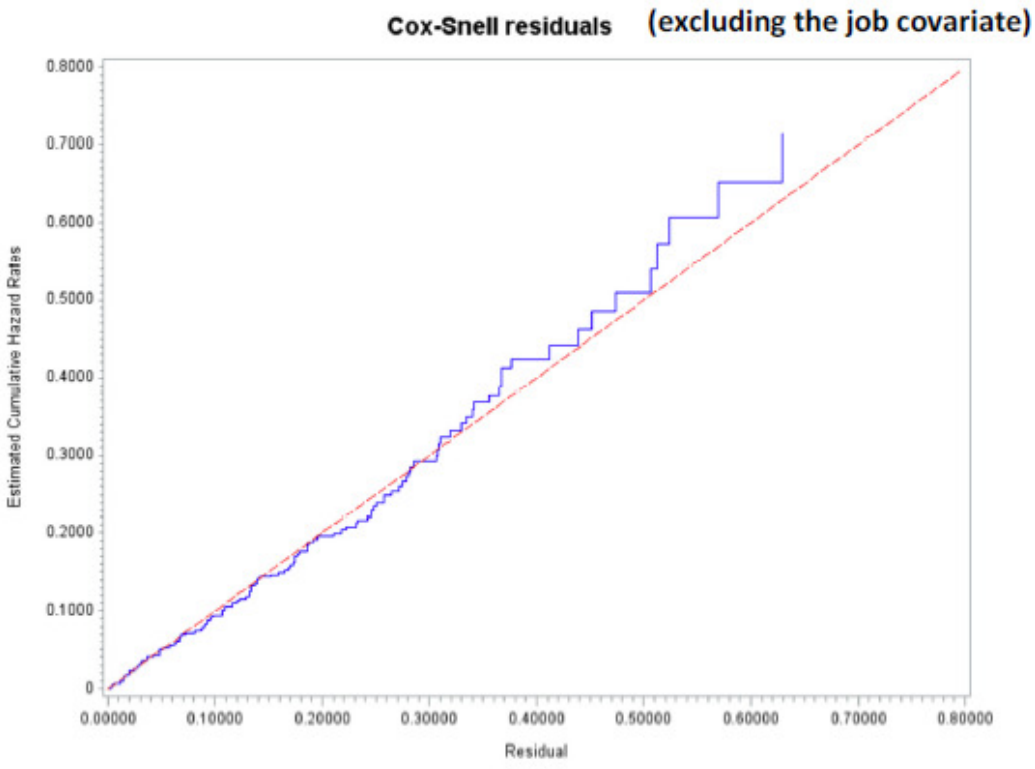
Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% HR Confidence Limits	
fin_aid	1	-0.36115	0.19046	3.5956	0.0579	0.697	0.478	1.010
Age	1	-0.05958	0.02075	8.2434	0.0041	0.942	0.902	0.979
Married	1	-0.45818	0.37436	1.4979	0.2210	0.632	0.280	1.238
prior_conv	1	0.09061	0.02763	10.7506	0.0010	1.095	1.034	1.153
Zjob	1	-0.45079	0.21475	4.4063	0.0358	0.637	0.421	0.980
time_job	1	-0.10450	0.01135	84.7906	<.0001	0.901	0.880	0.920
job	1	-4.74686	0.42501	124.7406	<.0001	0.009	0.004	0.020



Model 2: age <23 vs ≥ 23 years

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1227.506	1194.499
AIC	1227.506	1204.499
SBC	1227.506	1218.180

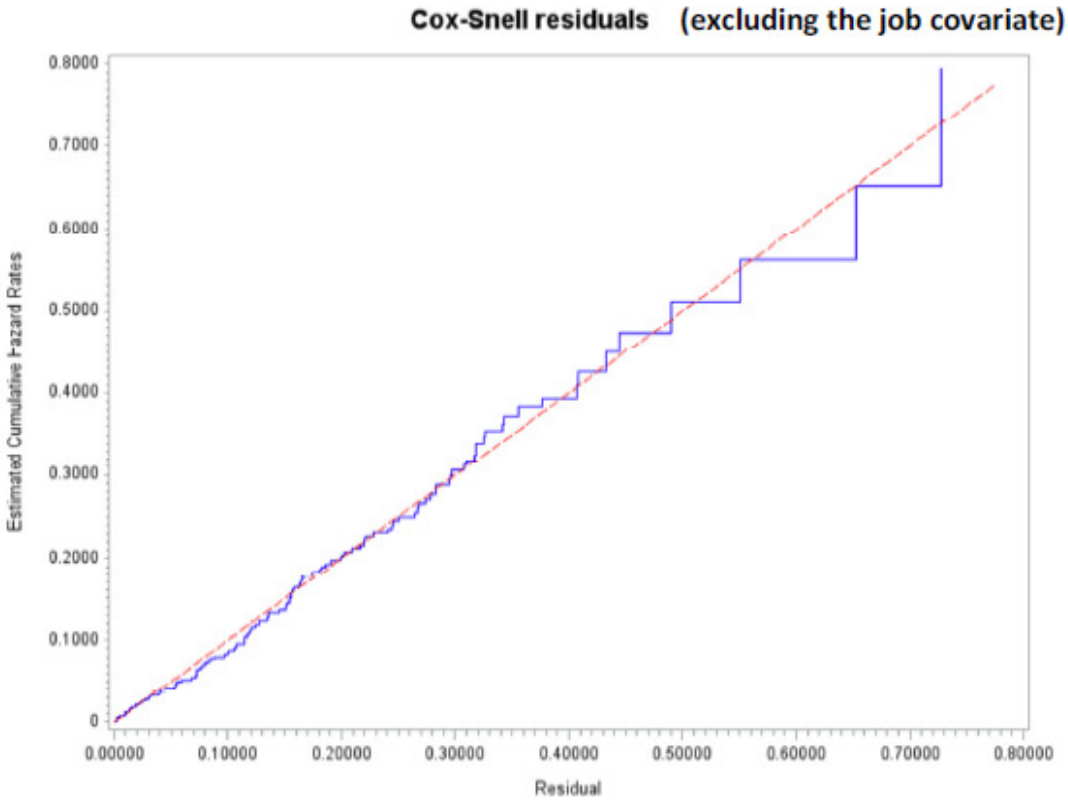
Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% HR Confidence Limits	
fin_aid	1	-0.39315	0.19011	4.2767	0.0386	0.675	0.463	0.977
age23	1	0.52137	0.19665	7.0289	0.0080	1.684	1.151	2.494
married	1	-0.45155	0.37649	1.4385	0.2304	0.637	0.281	1.252
prior_conv	1	0.09347	0.02751	11.5470	0.0007	1.098	1.038	1.156
Zjob	1	-0.45870	0.21466	4.5659	0.0326	0.632	0.418	0.972
time_job	1	-0.10347	0.01127	84.2514	<.0001	0.902	0.881	0.921
job	1	-4.72453	0.42276	124.8887	<.0001	0.009	0.004	0.020



Model 3: age <30 vs ≥ 30 years

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1227.506	1195.886
AIC	1227.506	1205.886
SBC	1227.506	1219.567

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% HR Confidence Limits	
fin_aid	1	-0.36595	0.19077	3.6797	0.0551	0.694	0.475	1.005
age30	1	0.73171	0.33327	4.8204	0.0281	2.079	1.138	4.262
married	1	-0.57983	0.37025	2.4526	0.1173	0.560	0.249	1.086
prior_conv	1	0.09049	0.02727	11.0129	0.0009	1.095	1.035	1.152
Zjob	1	-0.46748	0.21587	4.6895	0.0303	0.627	0.413	0.966
time_job	1	-0.10255	0.01138	81.2352	<.0001	0.903	0.882	0.922
job	1	-4.67220	0.42363	121.6355	<.0001	0.009	0.004	0.021



Model 4: Stratified by age <23 vs ≥ 23 years

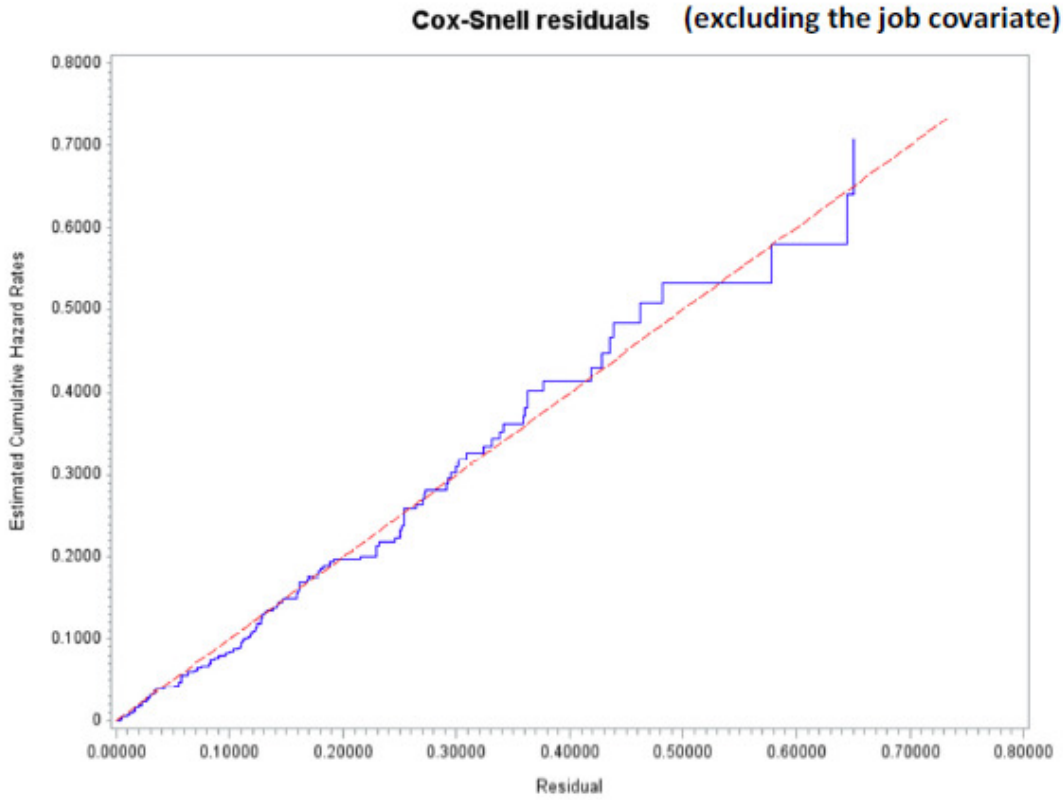
Model Fit Statistics				
Criterion	Without Covariates	With Covariates Full model	With Covariates age<23	With Covariates age≥23
-2 LOG L	1112.071	1090.100	659.964	424.324
AIC	1112.071	1098.100	667.964	432.324
SBC	1112.071	1109.045	676.958	439.461

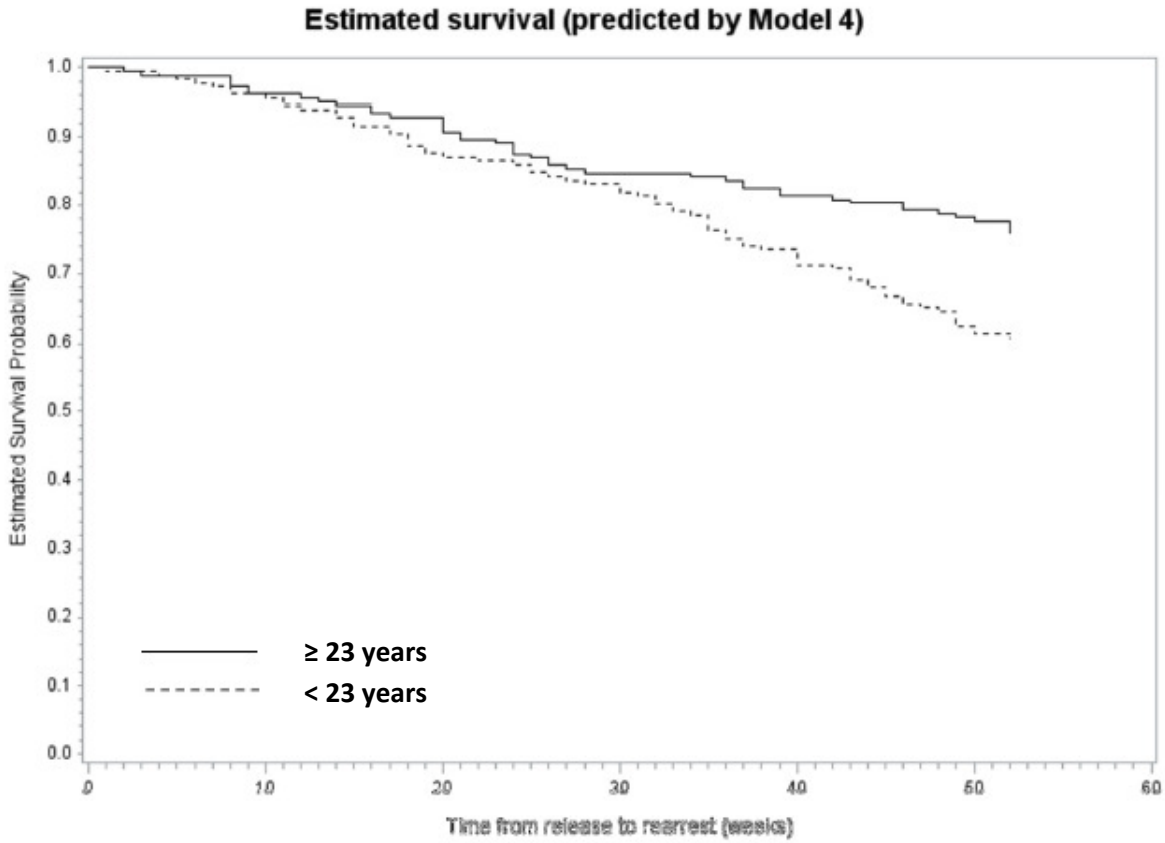
Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% HR Confidence Limits	
fin_aid	1	-0.39380	0.19006	4.2928	0.0383	0.674	0.463	0.976
married	1	-0.43284	0.37706	1.3178	0.2510	0.649	0.286	1.278
prior_conv	1	0.09243	0.02739	11.3840	0.0007	1.097	1.037	1.155
Zjob	1	-0.46240	0.21479	4.6345	0.0313	0.630	0.416	0.969
time_job	1	-0.10225	0.01124	82.7308	<.0001	0.903	0.882	0.922
job	1	-4.65077	0.41907	123.1642	<.0001	0.010	0.004	0.022

TABLE C.2

Upper Percentiles of a Chi-Square Distribution

Degrees of Freedom	*Upper Percentile*				
	0.1	*0.05*	*0.01*	*0.005*	*0.001*
1	2.70554	3.84146	6.63489	7.87940	10.82736
2	4.60518	5.99148	9.21035	10.59653	13.81500
3	6.25139	7.81472	11.34488	12.83807	16.26596
4	7.77943	9.48773	13.27670	14.86017	18.46623
5	9.23635	11.07048	15.08632	16.74965	20.51465
6	10.64464	12.59158	16.81187	18.54751	22.45748
7	12.01703	14.06713	18.47532	20.27774	24.32130
8	13.36156	15.50731	20.09016	21.95486	26.12393
9	14.68366	16.91896	21.66605	23.58927	27.87673
10	15.98717	18.30703	23.20929	25.18805	29.58789



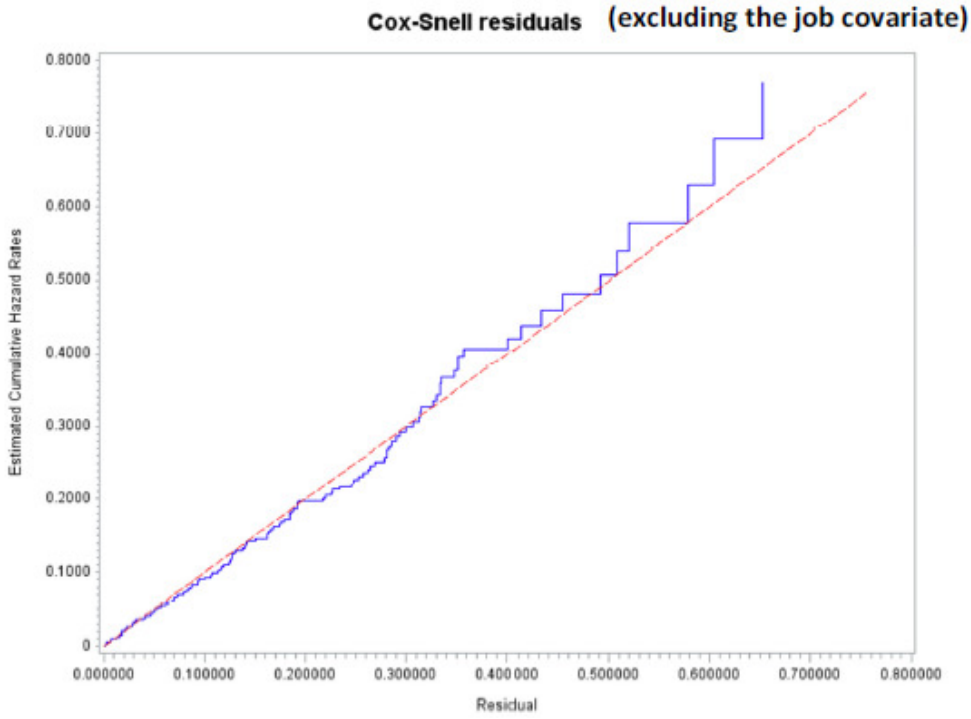


Predicted survival for reference values of categorical variable(s), and mean values of continuous variable(s). Time-dependent variable(s) excluded.

Model 5: time dependent covariate for age included

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1227.506	1185.060
AIC	1227.506	1197.060
SBC	1227.506	1213.477

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% HR Confidence Limits	
fin_aid	1	-0.36286	0.19043	3.6308	0.0567	0.696	0.477	1.008
age	1	0.12423	0.06578	3.5666	0.0590	1.132	0.986	1.285
lnt_age	1	-0.06076	0.02188	7.7131	0.0055	0.941	0.901	0.984
married	1	-0.42654	0.37479	1.2952	0.2551	0.653	0.289	1.279
prior_conv	1	0.09150	0.02770	10.9126	0.0010	1.096	1.035	1.154
Zjob	1	-0.45554	0.21385	4.5377	0.0332	0.634	0.420	0.973
time_job	1	-0.10535	0.01138	85.7374	<.0001	0.900	0.880	0.920
job	1	-4.76343	0.42531	125.4400	<.0001	0.009	0.004	0.019



END OF TASK 2

(4) **Task 3**

You are at your first job as a statistician, analyzing the data described in Task 2 using Cox regression. The dependent variable is time to rearrest, and the explanatory variable of main interest is whether individuals have received financial aid or not. When including this covariate as a single explanatory variable, the hazard ratio gets a value below 1 (indicating a lower risk of rearrest for individuals who had financial aid). In the different models presented in Task 2, with added covariates, you see the same result.

However, when adding other possible explanatory variables to the model (not shown in Task 2) you find one model where the hazard ratio for financial aid gets a value *above* 1.

You show your results to your employer who wants to use this model, the one that "shows" that financial aid actually leads to a *higher* risk of rearrest. This would mean that there is no use in providing financial aid.

What do you do (what would be the ethically correct thing to do)?

(6) **Task 4**

(3) **A** Why is the Cox regression method said to be a "semiparametric" method?

(3) **B** When would it be appropriate to use a parametric model to estimate time-to-event data?