FINAL EXAMINATION
Analysis of survival data (7.5hp) 2025-02-15

INFORMATION:

A. Allowed means of assistance:
Handouts from lectures L1-L7 (handwritten notes allowed), ethical guidelines by ISI, RSS, ASA, and Svenska statistikfrämjandet. Calculator, ruler, dictionary.

B. Writing time: 5 hours.

C. The examination consists of 5 tasks, for a total of 80 points.
Including any bonus points, at least 48 points are needed to pass (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, 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 Studium.

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!

(11) **Task 1**

A tech company is studying the tenure of employees, focusing on how long employees stay with the company before resigning or retiring. The study includes employees who were hired on or after January 1, 2015. Their tenure is measured in months from their hire date to either the date they left the company or the last recorded follow-up.

(2) **A** Describe the type(s) of truncation represented in this study. Motivate your answer. If you find that no truncation is present, provide an example of truncation that could have been present in a study like this one.

(6) **B** For the three following employees, identify the type(s) of censoring present. Motivate your answer.

i) An employee hired in January 2018 who is still employed as of February 2025.

ii) An employee hired in June 2016 who resigned in October 2020.

iii) An employee hired in September 2017, whose employment records were lost due to a database error in October 2021, but who was later discovered to have resigned before the error was fixed in December 2021.

(3) **C** Confining your attention to the three employees described above, write down the likelihood for this portion of the study.

(26) **Task 2**

Inspired by the study in Task 1, a larger study on employee retention in the tech industry was performed to examine the relationship between company size and the time until an employee voluntarily left their job. Data was collected for 450 software engineers who started their careers between 2010 and 2020. Employees still working at their respective companies at the study's conclusion in December 2024 were considered censored.

The following variables are included in the data:

- *time*: Time in years from the start of employment to resignation or the study's conclusion
- *censored*: Indicator for whether the employee was still employed at the end of the study (1 = yes, 0 = no)
- *company_size*: Size of the employer (Small, Medium, Large)

Source: Ficticious statistics

Use the SAS output on the following pages to answer the questions below. Note that survival estimates are presented only for selected time points (all necessary information for answering the questions is included).

NOTE: In each of the questions, interpret "survival" and "event" in the context of this study.

(2) **A** Are there any competing risks present in this study? If so, how does that affect the results of the analysis?

(8) **B** Present, interpret, and compare the following estimates for the two company sizes with the lowest survival probabilities (i.e., only two of the three company size groups):

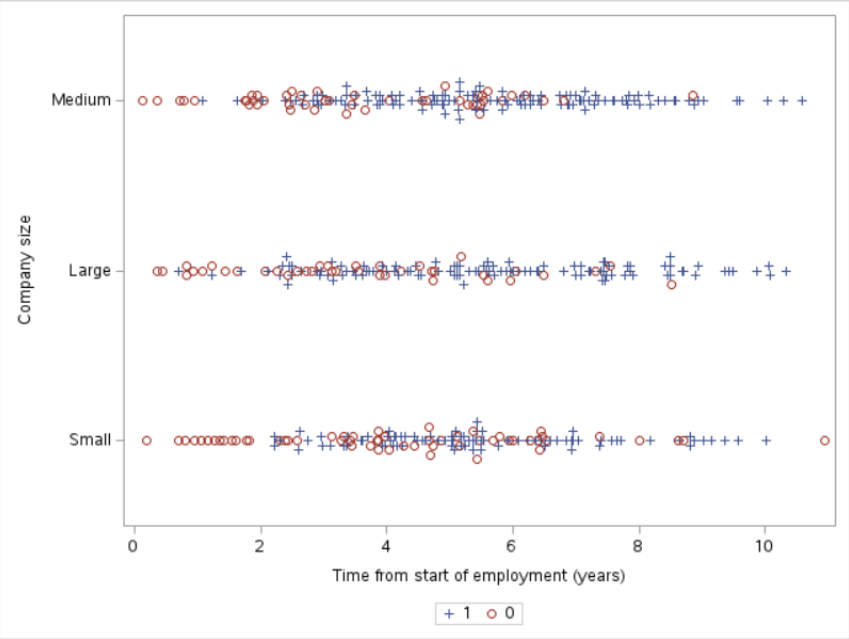
- The 25th percentile, 75th percentile and median (50th percentile) of "survival" times.

If any measures cannot be estimated, explain why and present the minimum time for those measures.

(3) **C** Present and compare the approximate probability of "surviving" at least 6 years for the two company size groups in part B above.

(13) **D** Is there a significant difference in the probability of leaving the company between the three company size groups? Perform an appropriate hypothesis test to determine this. Remember to follow the instructions on the front page.

SAS OUTPUT Task 2



Summary of the Number of Censored and Uncensored Values

Stratum	company_size	Total	Failed	Censored	Percent Censored
1	Large	143	38	105	73.43
2	Medium	166	47	119	71.69
3	Small	141	57	84	59.57
Total		450	142	308	68.44

Stratum 1: Company size = Large

Product-Limit Survival Estimates

time		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000		1.0000	0	0	0	143
0.3720		0.9930	0.00699	0.00697	1	142
0.4390		0.9860	0.0140	0.00982	2	141
0.6955	*	.	.	.	2	140
5.1000		.	.	.	29	68
5.1667	*	.	.	.	29	67
5.1850		0.7550	0.2450	0.0398	30	66
5.2120	*	.	.	.	30	65
5.4466	*	.	.	.	30	62
5.5311		0.7428	0.2572	0.0410	31	61
5.5314	*	.	.	.	31	60
8.5043	*	.	.	.	37	12
8.5125		0.5824	0.4176	0.0716	38	11
8.6740	*	.	.	.	38	10
10.0511	*	.	.	.	38	2
10.0751	*	.	.	.	38	1
10.3243	*	.	.	.	38	0

Note: The marked survival times are censored observations.

Stratum 2: Company size = Medium

Product-Limit Survival Estimates

time		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000		1.0000	0	0	0	166
0.1373		0.9940	0.00602	0.00601	1	165
0.3543		0.9880	0.0120	0.00847	2	164
5.3674	*	.	.	.	33	78
5.3759		0.7673	0.2327	0.0359	34	77
5.4092	*	.	.	.	34	76
5.4186	*	.	.	.	34	75
5.4215	*	.	.	.	34	74
5.4409		0.7569	0.2431	0.0369	35	73
5.4427		0.7465	0.2535	0.0378	36	72
5.4694	*	.	.	.	36	71
5.4737		0.7360	0.2640	0.0387	37	70
8.7827	*	.	.	.	46	9
8.8588		0.5582	0.4418	0.0781	47	8
8.8654	*	.	.	.	47	7
8.9365	*	.	.	.	47	6
9.0223	*	.	.	.	47	5
9.5446	*	.	.	.	47	4
9.6012	*	.	.	.	47	3
10.0430	*	.	.	.	47	2
10.2861	*	.	.	.	47	1
10.5906	*	.	.	.	47	0

Note: The marked survival times are censored observations.

Stratum 3: Company size = Small

Product-Limit Survival Estimates

time		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000		1.0000	0	0	0	141
0.1904		0.9929	0.00709	0.00707	1	140
0.7031		0.9858	0.0142	0.00996	2	139
4.2387	*	.	.	.	30	85
4.2613		0.7662	0.2338	0.0370	31	84
4.2805	*	.	.	.	31	83
4.4467	*	.	.	.	31	82
4.4480		0.7568	0.2432	0.0377	32	81
4.4733	*	.	.	.	32	80
4.5555	*	.	.	.	32	79
4.6046	*	.	.	.	32	78
4.6663		0.7471	0.2529	0.0385	33	77
4.6813		0.7374	0.2626	0.0392	34	76
6.4353	*	.	.	.	49	32
6.4444		0.5232	0.4768	0.0543	50	31
6.4699		0.5064	0.4936	0.0551	51	30
6.4989	*	.	.	.	51	29
6.5222		0.4889	0.5111	0.0559	52	28
6.6022	*	.	.	.	52	27
8.6317	*	.	.	.	55	10
8.7110		0.3548	0.6452	0.0711	56	9
8.8020	*	.	.	.	56	8
10.0147	*	.	.	.	56	1
10.9445		0	1.0000	.	57	0

Note: The marked survival times are censored observations.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	5.9823	2	0.0502
Wilcoxon	2.6432	2	0.2667
-2Log(LR)	5.8094	2	0.0548

END OF TASK 2

(30) **Task 3**

In the study on employee retention described in Task 2, the researchers also investigated the relationship between a number of possible covariates and employment time.

The following variables are included in the data:

- *time*: Time in years from the start of employment to resignation or the study's conclusion
- *censored*: Indicator for whether the employee was still employed at the end of the study (1 = yes, 0 = no)
- *company_size*: Size of the employer (Small, Medium, Large)
- *education_level*: Employee's highest degree (Bachelor's, Master's, PhD)
- *remote_work*: Indicator for whether the employee primarily worked remotely (1 = yes, 0 = no)
- *starting_salary*: Initial salary in thousands of USD
- *salary_group*: Initial salary divided into groups (up to 80, 80-100, over 100)

Use the output below, generated from SAS PROC PHREG, and answer the questions below.

NOTE: Remember to interpret "survival" and "event" in the context of this study.

(3) **A** Do any of the covariates above need to be handled in a special way (e.g. by recoding or transforming them), or can they be used as they are in the regression analysis? Motivate your answer.
(You do not need to take the PH assumption into account here.)

(15) **B** Is the presented Cox regression model appropriate? Motivate your answer carefully. If you find that the model is not appropriate, suggest a model that you would consider evaluating.

(8) **C** Interpret the relationships between the covariates in the presented model and time to resignation.

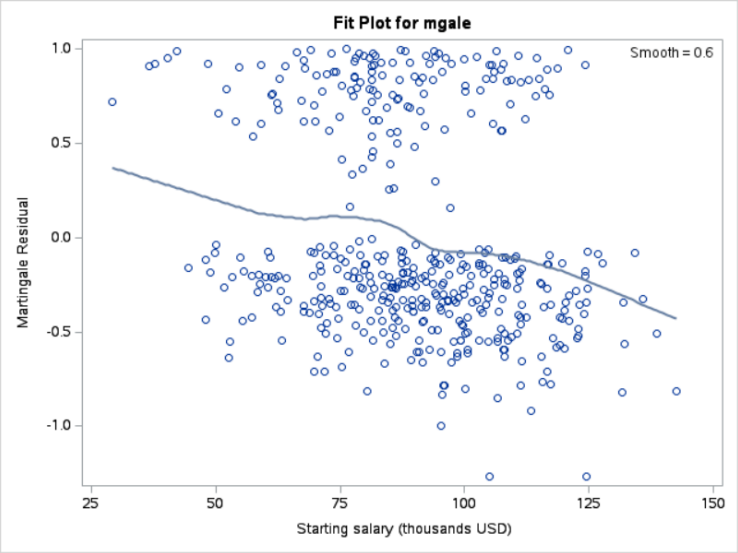
If you find that the model is stratified, remember to also interpret the stratifying variable(s).

If you find that the model includes a time-dependent coefficient, calculate the hazard ratio at 2, 5, and 8 years, respectively for that coefficient (confidence intervals are not required).

(2) **D** Calculate and interpret the relative risk of an "event" for employees at large companies compared to employees at medium sized companies.

(2) **E** Calculate and interpret the generalized R^2 for the presented model.

SAS OUTPUT Task 3

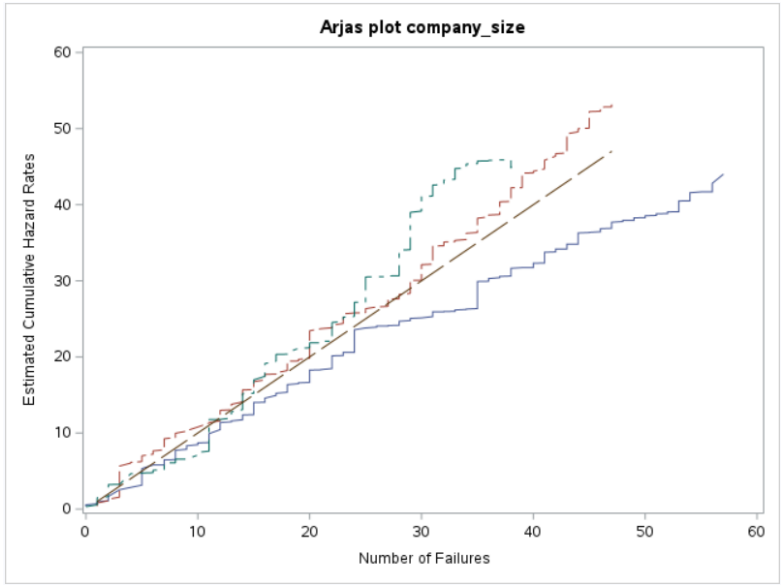


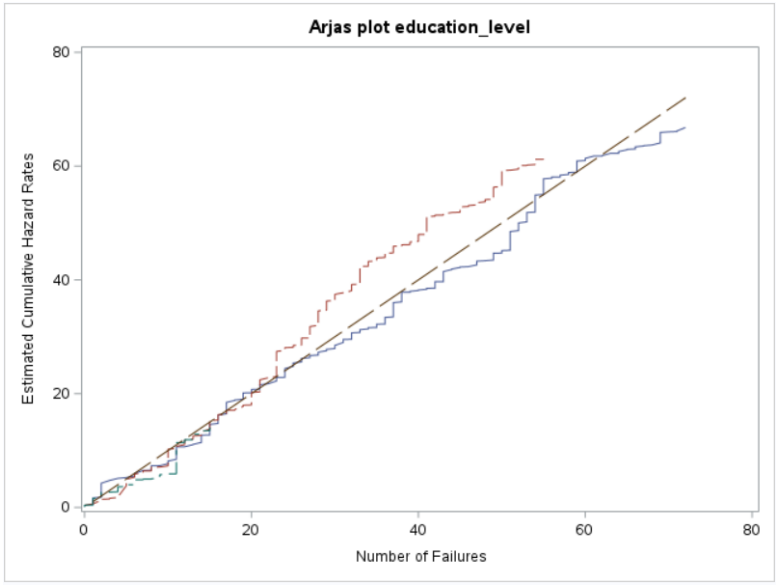
Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
lnt_remote_work (yes)	1	0.18543	0.22115	0.7030	0.4018	1.204
lnt_starting_salary	1	-0.00741	0.00576	51.6568	<.0001	0.993

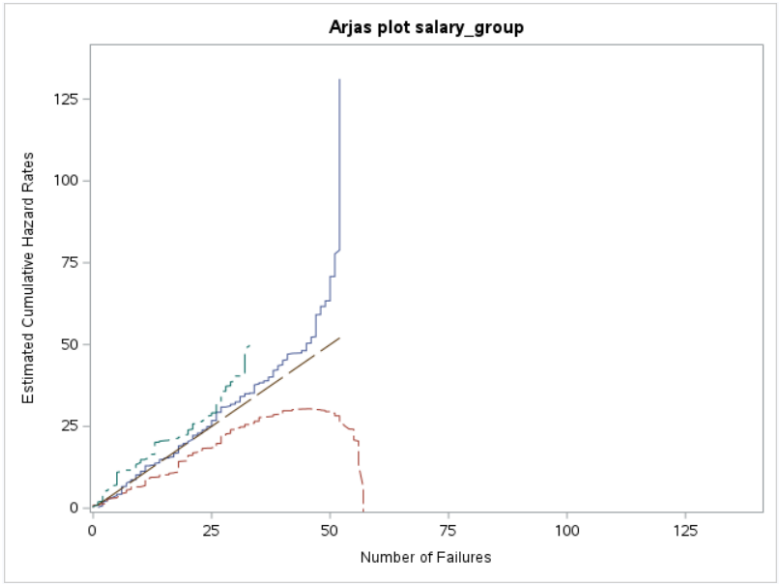
NOTE: "lnt_x" denotes a time-dependent covariate, calculated as x*log(time)

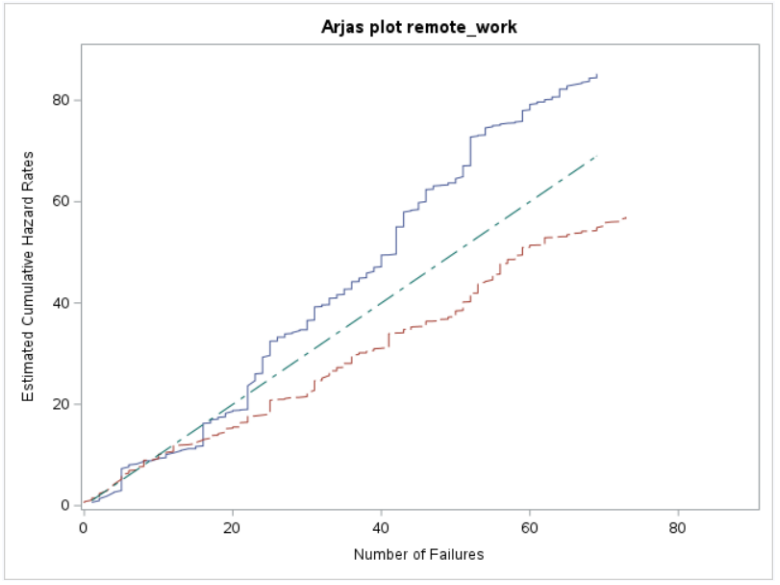
Linear Hypotheses Testing Results			
Test	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
lnt_company_medium = lnt_company_large = 0	2	1.6703	0.4338
lnt_education_Master = lnt_education_PhD = 0	2	49.0725	<.0001
lnt_salary_80_to_100 = lnt_salary_over_100 = 0	2	0.8315	0.6599

NOTE: Dummy variables have been created above for two of the three categories of the variables *company_size*, *education_level*, and *salary_group*.


Arjas plot company_size
Estimated Cumulative Hazard Rates
Number of Failures
0
10
20
30
40
50
60


Arjas plot education_level
Estimated Cumulative Hazard Rates
Number of Failures
0
20
40
60
80


Arjas plot salary_group
Estimated Cumulative Hazard Rates
Number of Failures
0
25
50
75
100
125


Arjas plot remote_work
Estimated Cumulative Hazard Rates
Number of Failures
0
20
40
60
80

Estimated model: *company_size, remote_work, starting_salary, lnt_starting_salary*, stratified by *education_level*

Summary of the Number of Event and Censored Values					
Stratum	education_level	Total	Event	Censored	Percent Censored
1	Bachelor's	214	72	142	66.36
2	Master's	192	55	137	71.35
3	PhD	44	15	29	65.91
Total		450	142	308	68.44

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1297.853	1260.920
AIC	1297.853	1270.920
SBC	1297.853	1285.699

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
company_size	2	9.5525	0.0084
remote_work	1	8.1764	0.0042
starting_salary	1	2.8368	0.0921
lnt_starting_salary	1	1.8311	0.1760

education_level=Bachelor's

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	681.428	653.427
AIC	681.428	663.427
SBC	681.428	674.810

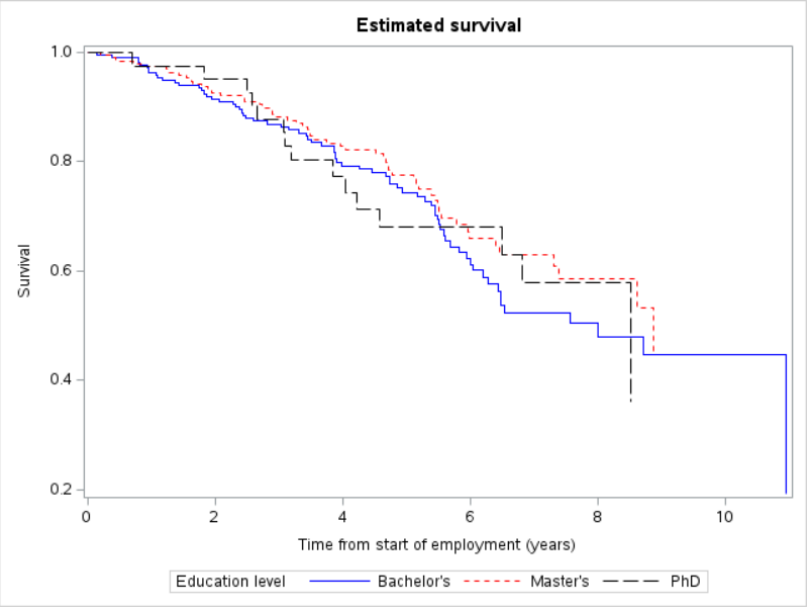
education_level=Master's

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	517.210	503.093
AIC	517.210	513.093
SBC	517.210	523.130

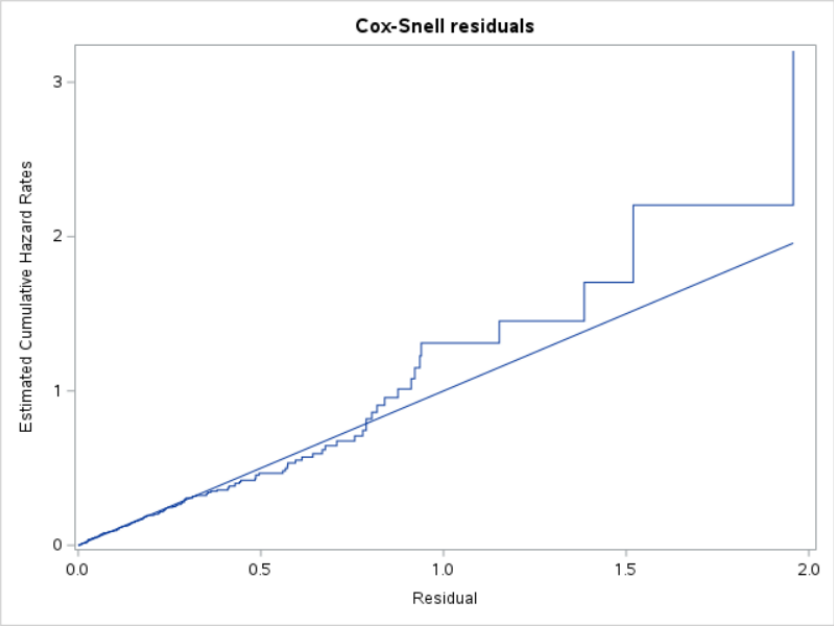
education_level=PhD

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	99.215	90.015
AIC	99.215	100.015
SBC	99.215	103.556

Analysis of Maximum Likelihood Estimates									
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Profile Likelihood Confidence Limits	
company_size	Large	1	-0.60683	0.21374	8.0606	0.0045	0.545	0.356	0.826
company_size	Medium	1	-0.46541	0.19918	5.4598	0.0195	0.628	0.423	0.927
remote_work		1	0.49246	0.17222	8.1764	0.0042	1.636	1.168	2.297
starting_salary		1	-0.01289	0.00765	2.8368	0.0921	0.987	0.973	1.002
lnt_starting_salary		1	-0.00790	0.00584	1.8311	0.1760	0.992	0.981	1.003



The estimated survival above shows the predicted survival for reference value(s) of any categorcal variable(s), and mean value(s) of any continuous variable(s). Any time-dependent variable(s) are excluded.



END OF TASK 3

(9) **Task 4**

In a study examining career longevity in professional basketball, researchers investigated the relationship between team performance and the time until retirement. Data was collected for 2500 randomly selected players who began their professional careers between 2005 and 2015. Players who were still active at the study's conclusion in December 2020 were considered censored.

Variable specification:

- *time:* Time in years from the start of the professional career to retirement or the study's conclusion
- *censored*: Indicator for whether the player was still active at the end of the study (1 = yes, 0 = no)
- *team_performance*: Team's average performance level (High, Medium, Low)
- *nationality:* Player's nationality (1 = domestic, 2 = international)
- *age*: Age at debut (years)

Source: Ficticious statistics

An accelerated failure time model is used to estimate the relationship between the variables above, where the Weibull distribution is used.

Use the SAS output on the following pages to answer the questions below.

(2) **A** Which model is being estimated here? Specify the model by using the variable names above and appropriate symbols representing regression coefficients.

(2) **B** What does it mean that "the Weibull distribution is used"?

(5) **C** Interpret the relationships between the covariates in the presented model and time to retirement.

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	3.2223	0.0858	3.0541	3.3905	1409.97	<.0001
performance_medium	1	0.0944	0.0387	0.0185	0.1703	5.94	0.0148
performance_high	1	0.3096	0.0436	0.2241	0.3951	50.37	<.0001
nationality_internat	1	0.0024	0.0326	-0.0614	0.0662	0.01	0.9417
age	1	-0.0360	0.0029	-0.0417	-0.0304	157.68	<.0001
Scale	1	0.6134	0.0129	0.5886	0.6392		
Weibull Shape	1	1.6304	0.0343	1.5644	1.6991		

Transformed Parameter Estimate

	Parameter Estimate	Standard Error	Exp Estimate
Intercept	0.005	0.000	1.005
performance_medium	-0.154	0.000	0.857
performance_high	-0.505	0.000	0.604
nationality_internat	-0.004	0.000	0.996
age	0.059	0.000	1.061
Scale	1.630	0.000	5.106

END OF TASK 4

(4) **Task 5**

You are working at a telecommunications company analyzing customer retention using survival analysis. The company has recently introduced a new premium plan, and you are asked to model time until churn (when the customer leaves) for different customer segments.

Upon inspecting the data, you realize that customers who upgraded to the new premium plan are censored in the dataset, meaning their churn status is treated as unknown instead of following up what happens in the new segment. This could introduce bias into your estimates, making the new plan appear more effective than it actually is.

When you raise this issue with your manager, you are told, *"The leadership team is keen on showcasing the success of the premium plan. Just proceed with the analysis as planned—no need to overcomplicate things."*

What do you do?
What would be the ethically correct thing to do in this situation?
Refer to four relevant sections of one or more codes of ethics documents.
Specify the document(s) and the exact section(s) you are referencing. Remember to use quotation marks if you are directly quoting from the text.

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
11	17.27501	19.67515	24.72502	26.75686	31.26351
12	18.54934	21.02606	26.21696	28.29966	32.90923
13	19.81193	22.36203	27.68818	29.81932	34.52737
14	21.06414	23.68478	29.14116	31.31943	36.12387
15	22.30712	24.99580	30.57795	32.80149	37.69777
16	23.54182	26.29622	31.99986	34.26705	39.25178
17	24.76903	27.58710	33.40872	35.71838	40.79111
18	25.98942	28.86932	34.80524	37.15639	42.31195
19	27.20356	30.14351	36.19077	38.58212	43.81936
20	28.41197	31.41042	37.56627	39.99686	45.31422
21	29.61509	32.67056	38.93223	41.40094	46.79627
22	30.81329	33.92446	40.28945	42.79566	48.26762
23	32.00689	35.17246	41.63833	44.18139	49.72764
24	33.19624	36.41503	42.97978	45.55836	51.17897
25	34.38158	37.65249	44.31401	46.92797	52.61874
26	35.56316	38.88513	45.64164	48.28978	54.05114
27	36.74123	40.11327	46.96284	49.64504	55.47508
28	37.91591	41.33715	48.27817	50.99356	56.89176
29	39.08748	42.55695	49.58783	52.33550	58.30064
30	40.25602	43.77295	50.89218	53.67187	59.70221
31	41.42175	44.98534	52.19135	55.00248	61.09799
32	42.58473	46.19424	53.48566	56.32799	62.48728
33	43.74518	47.39990	54.77545	57.64831	63.86936
34	44.90316	48.60236	56.06085	58.96371	65.24710
35	46.05877	49.80183	57.34199	60.27459	66.61917
36	47.21217	50.99848	58.61915	61.58107	67.98495
37	48.36339	52.19229	59.89256	62.88317	69.34759
38	49.51258	53.38351	61.16202	64.18123	70.70393
39	50.65978	54.57224	62.42809	65.47532	72.05504