KU LEUVEN

Statistical Analysis of Reliability and Survival Data: Rotterdam Dataset

Authors
Aksoy, Barış r0869901
Heller, Jack r0862809
Zdravković, Aleksandra r0869484

Leuven. May, 2022

1 Exploratory Data Analysis

These data sets are used in the paper by Royston and Altman that is referenced below. The Rotterdam data is used to create a fitted model, and the GBSG data for validation of the model. The paper gives references for the data source.

There are 43 subjects who have died without recurrence, but whose death time is greater than the censoring time for recurrence. A common way that this happens is that a death date is updated in the health record sometime after the research study ended, and said value is then picked up when a study data set is created. But it raises serious questions about censoring. For instance subject 40 is censored for recurrence at 4.2 years and died at 6.6 years; when creating the endpoint of recurrence free survival (earlier of recurrence or death), treating them as a death at 6.6 years implicitly assumes that they were recurrence free just before death. For this to be true we would have to assume that if they had progressed in the 2.4 year interval before death (while off study), that this information would also have been noted in their general medical record, and would also be captured in the study data set. However, that may be unlikely. Death information is often in a centralized location in electronic health records, easily accessed by a programmer and merged with the study data, while recurrence may require manual review. How best to address this is an open issue.

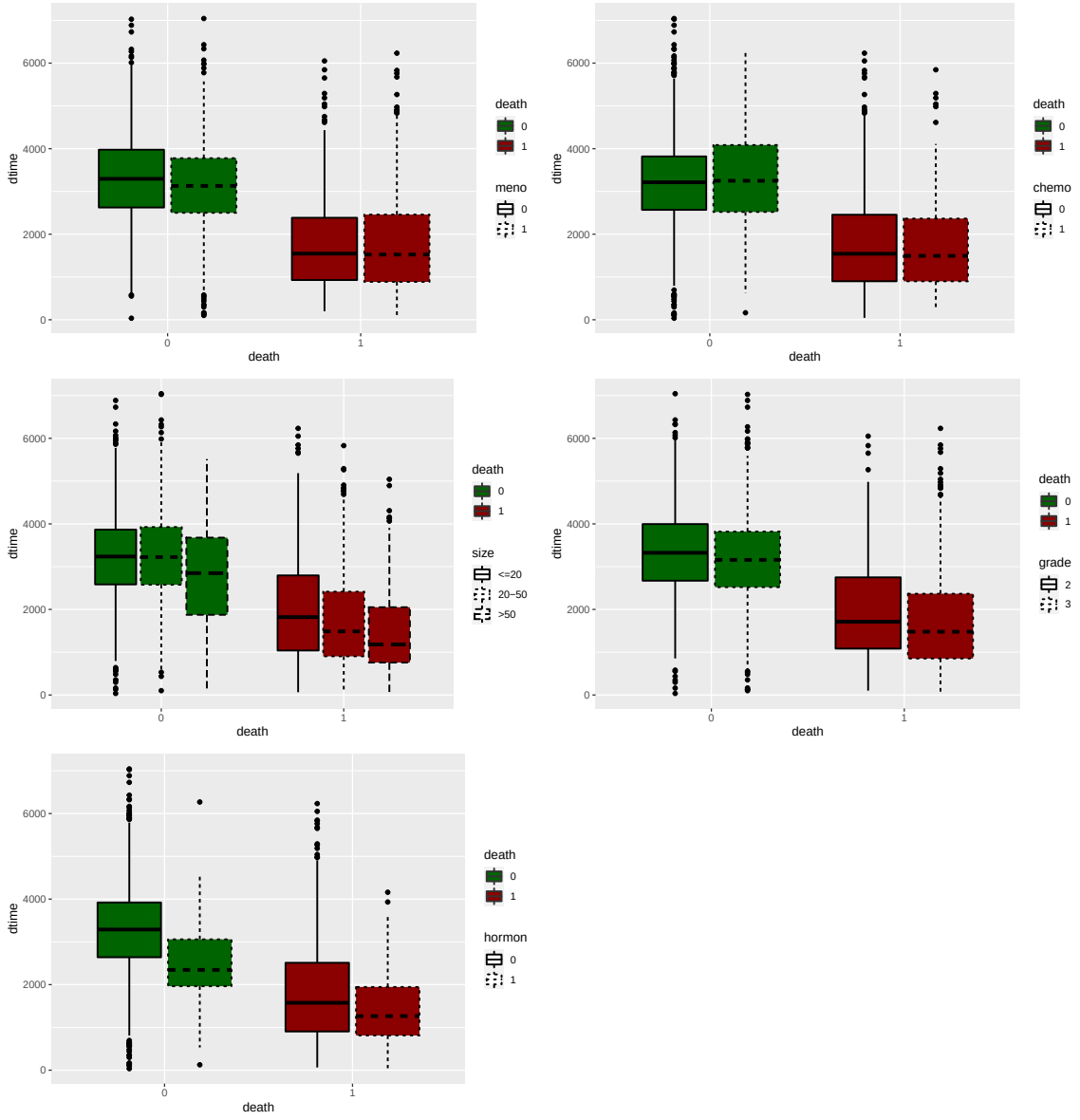
Table 1: Data description

pid	Patient identifier
year	Year of surgery
age	Age at surgery
meno	Menopausal status (0 = premenopausal, 1 = postmenopausal)
size	Tumor size, a factor with levels <= 20, 20-25, > 50
grade	Differentiation grade
nodes	Number of positive lymph nodes
pgr	Progesterone receptors (fmol/l)
er	Estrogen receptors (fmol/l)
hormon	Hormonal treatment (0=no, 1=yes)
chemo	Chemotherapy
rtime	Days to relapse or last follow-up
recur	0 = no relapse, 1 = relapse
dtime	Days to death or last follow-up
death	0 = alive, 1 = dead

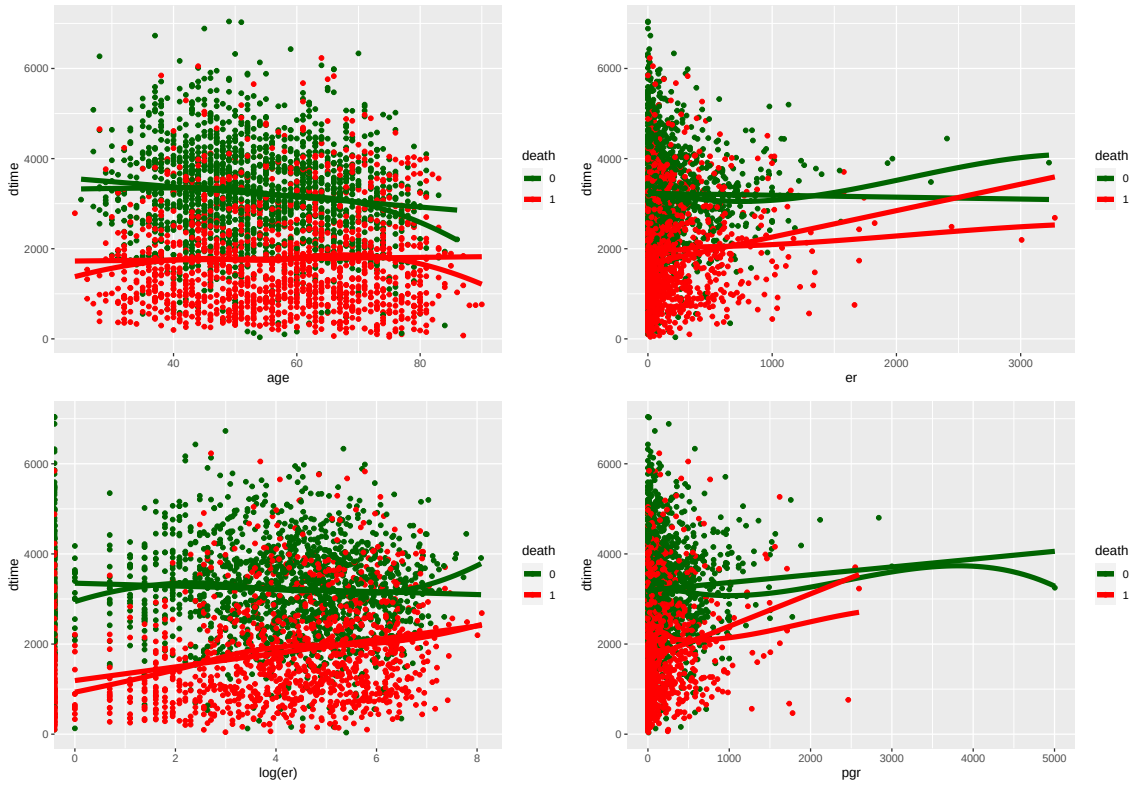
Table 1 explains the covariates in the Rotterdam dataset.

When we investigate the categorical variables, we see that some have a larger difference between time until death than others. When we first investigate menopause, we see that menopausal patients are very similar in their death time. A very similar pattern is observed for chemotherapy but the patients that did not die had a slightly higher time before leaving the study. While size, having three levels, needs to be interpreted slightly differently. With patients that died, the larger cancer cells certainly caused it to happen sooner. This makes intuitive sense that more severe cancer would have more long term health risks. Among patients that do not die, only the extremely large cancer patients had a much lower death time. While small and medium patients had a very similar response. Grade had a very similar response among all patients with higher grade patients living slightly shorter lives. Patients given hormonal therapy seem to have a smaller tail among patients

that died; meaning those who died are more likely to do it soon after treatment. Surviving patients that had hormonal therapy also appear to be censored more quickly than those who do not receive the treatment.



Among continuous variables, we first investigated age. We plotted a both the lowess curve and the linear model to see if they match. If so, they most likely have data with a limited skew and do not require transformation for modeling. We first evaluated age and there was not a difference between the linear and lowess curves. Next, we investigated the relationship between er and death time. Here we see a large deviation which was corrected for again with a log transform. Finally, we evaluated pgr and it had a good fit overall.



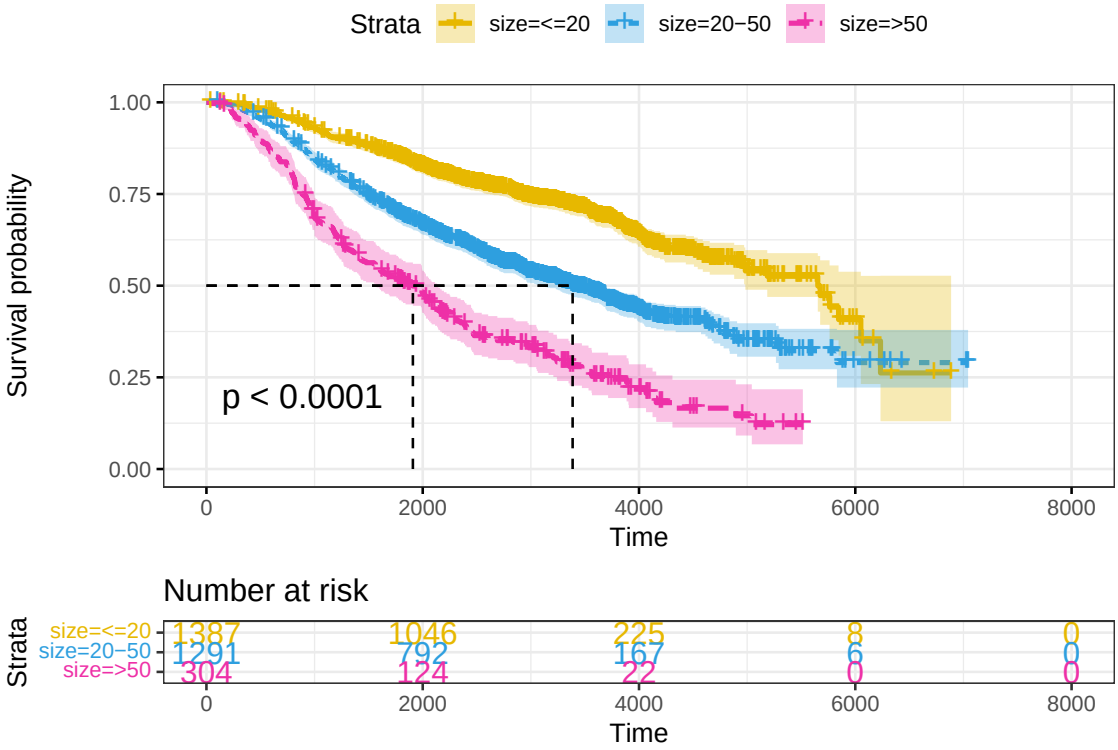
2 Further Analysis

Now the focus will be on the response variable, the censoring indicator, and the categorical variable.

2.1 Survival Distrubution by Levels of ...

(a) For each of the levels of the categorical variable, compute the survival distribution. Plot them on the same graph. What do the graphs suggest?

2.1.1 Size



The horizontal axis represents time in days, and the vertical axis shows the probability of surviving, or the proportion of people surviving. The lines represent survival curves of the three groups. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

Table 2: Summary of the model.

	records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
Size (≤20)	1387	1387	1387	414	4721.199	119.40158	5653	4983	
Size (20−50)	1291	1291	1291	646	3807.025	95.52799	3386	3084	3690
Size (>50)	304	304	304	212	2537.178	148.10071	1909	1566	2141

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2250, the probability of survival is approximately 0.625 for size>=50, and 0.85 for size<50.

From Table 2 can be seen that the median survival for Size 50 is 5653.0002289, for Size 20-50 is 3386.0000324, and for Size > 50 is 1908.9999793. This suggests worse survival for patients with tumor of larger size. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

2.1.2 Menopause

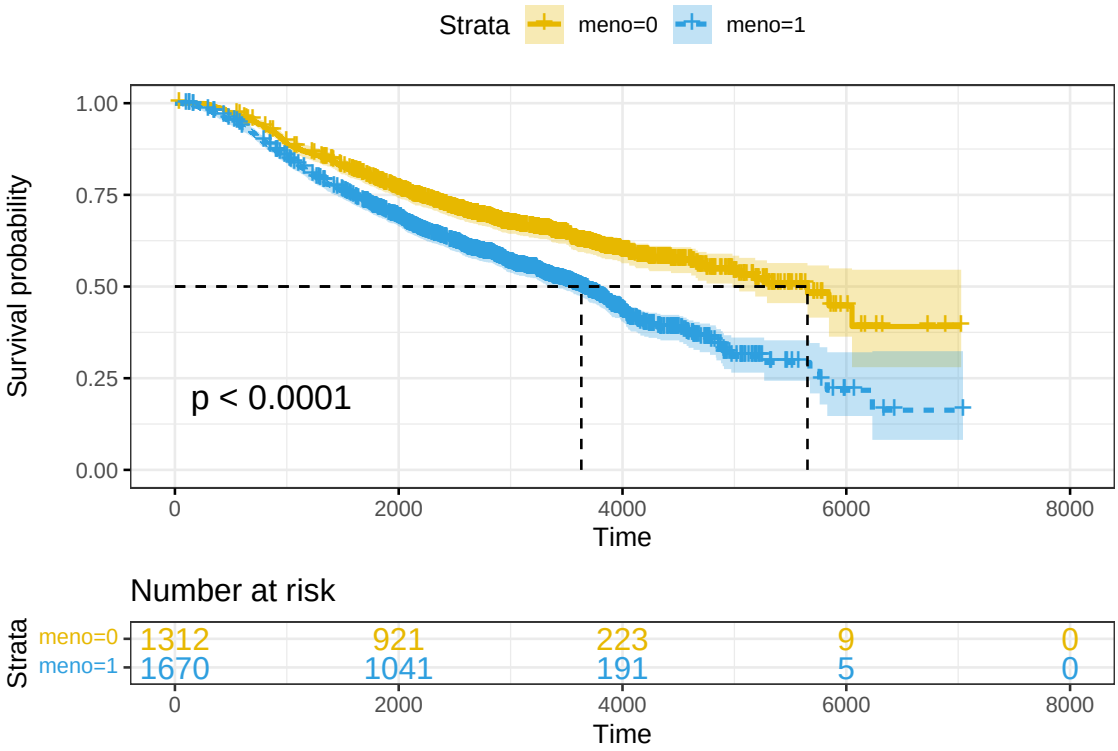


Table 3: Median survival times for each group.

	records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
Meno = 0	1312	1312	1312	468	4622.427	108.80722	5653	4983	
Meno = 1	1670	1670	1670	804	3672.887	92.46545	3632	3368	3813

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2200, the probability of survival is approximately 0.75 for non-menopausal patients, and 0.825 for menopausal patients.

From Table 3 can be seen that the median survival is 5653.0002289 for non-menopausal patients, and 3632.0001097 for menopausal, suggesting worse survival for patients that have gone through menopause.

2.1.3 Hormonal Treatment

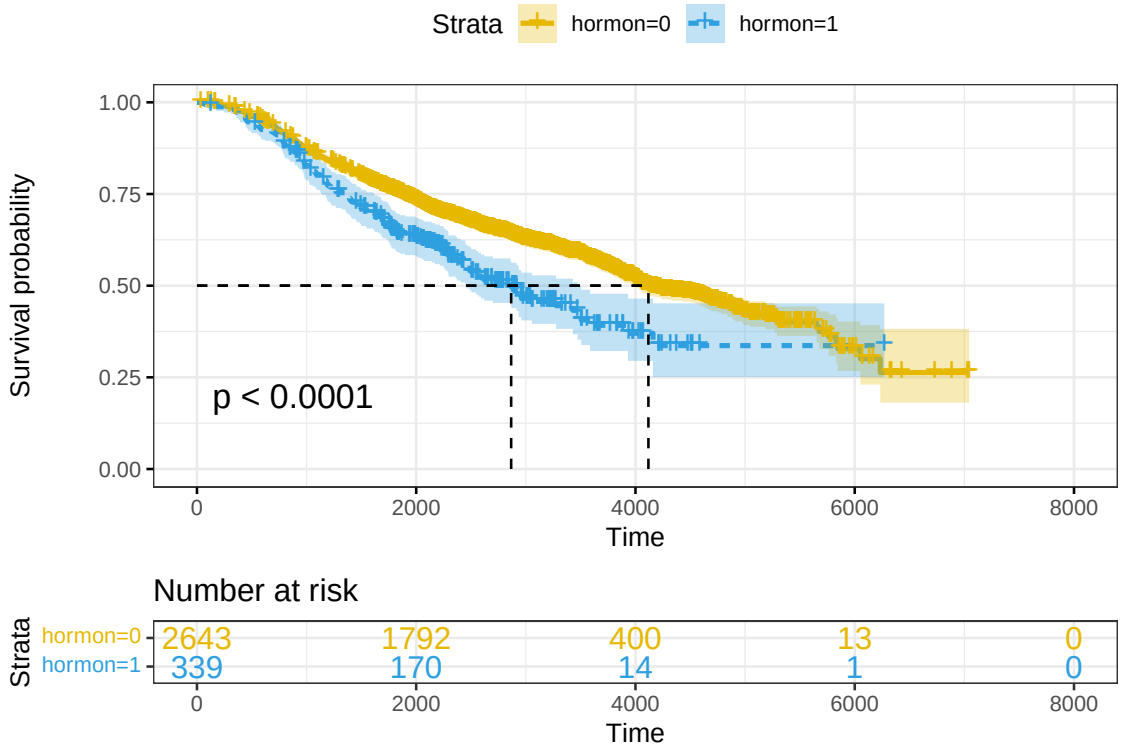


Table 4: Median survival times for each group.

	records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
hormon=0	2643	2643	2643	1113	4159.588	76.57099	4118	3988	4614
hormon=1	339	339	339	159	3659.665	203.38456	2866	2450	3472

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2200, the probability of survival is approximately 0.75 for premenopausal patients, and 0.825 for postmenopausal patients.

From Table 4 can be seen that the median survival is 4118.0001717 for patients that went through hormonal therapy, and 2865.9999881 for those who did not, suggesting slightly worse survival for patients that have gone through the hormonal therapy. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

2.1.4 Chemotherapy

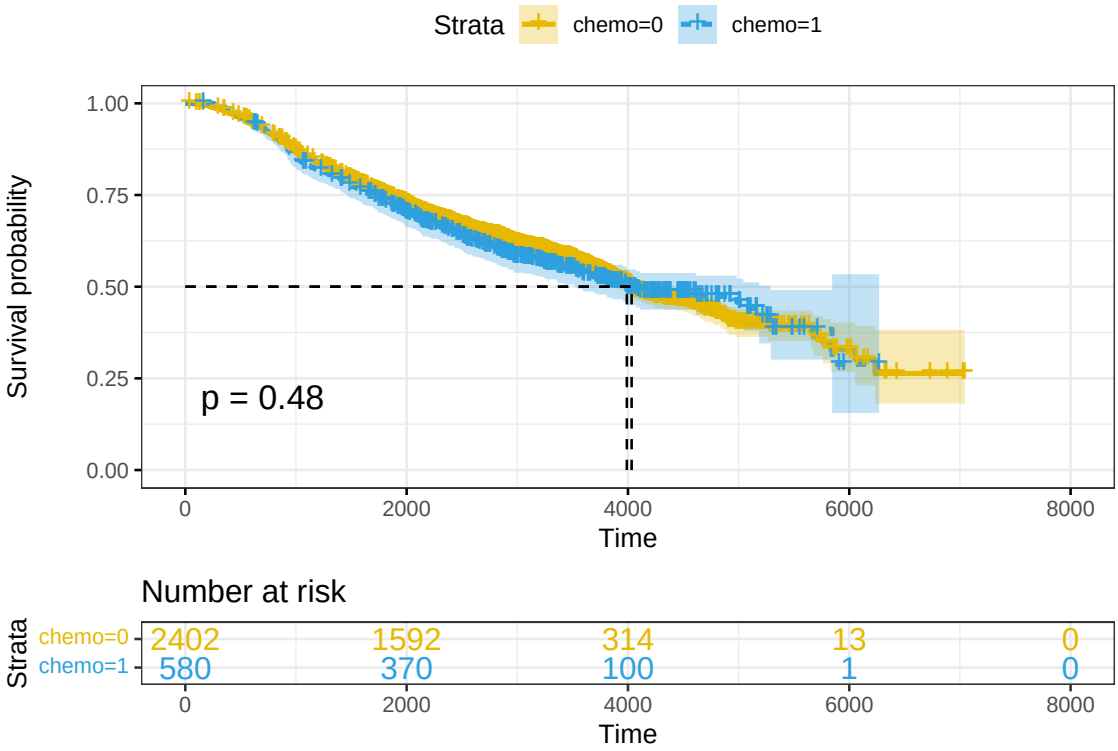


Table 5: Median survival times for each group.

	records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
chemo=0	2402	2402	2402	1014	4103.663	80.06765	4033	3885	4239
chemo=1	580	580	580	258	4080.311	162.59049	3990	3522	5291

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2200, the probability of survival is approximately 0.75 for premenopausal patients, and 0.825 for postmenopausal patients.

From Table 5 can be seen that the median survival is 4118.0001717 for patients that went through hormonal therapy, and 2865.9999881 for those who did not, suggesting slightly better survival for patients that have gone through chemotherapy. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

2.2 Confidence Intervals and Estimators by Levels of ...

(b) For each level obtain an appropriate estimator and confidence interval for the 3 quartiles of the survival curves. Interpret the results.

We first evaluate size and see similar results to what was observed in the EDA barcharts. We see that at the 25th percentile, size >50 has a very significant difference from the other two groups. With this being significantly smaller, their dtime is significantly smaller. This intuitively makes sense that a person who has larger cancer would have a worse life expectancy. However, this phenomenon has faded by the 50th percentile and persists through the 75th. This means that if a

Table 6: Estimates and confidence intervals for 3 quartiles for each level of covariates *size, chemo, meno, hormon.*

	$\widehat{q_1}$	5%	95%	$\widehat{q_2}$	5%	95%	$\widehat{q_3}$	5%	95%
Size ($\leq$50)	2880	2590	3315	5653	4983	*NA*	*NA*	6051	*NA*
Size (20−50)	1476	1361	1623	3386	3084	3690	*NA*	5830	*NA*
Size (>50)	890	809	999	1909	1566	2141	3714	3240	4309
Chemo = 0	1812	1677	1957	4033	3885	4239	*NA*	6051	*NA*
Chemo = 1	1699	1455	1954	3990	3522	5291	*NA*	5845	*NA*
Meno = 0	2115	1944	2371	5653	4983	*NA*	*NA*	*NA*	*NA*
Meno = 1	1571	1424	1723	3632	3368	3813	5762	5266	*NA*
Hormon = 0	1882	1742	1994	4118	3988	4614	*NA*	6051	*NA*
Hormon = 1	1361	1140	1618	2866	2450	3472	*NA*	*NA*	*NA*

person survives the initial hurdle after surgery, they are not expected to be significantly different from the rest of the population.

When we investigate the effect of menopause on survival, we see that generally people that are menopausal have lower survival times. This is supported by that at the 25th, 50th, and 75th percentiles, there is no overlap in the quantiles' confidence intervals. This means that at all stages, a menopausal person would have a shorter life. However, this could be due to other covariates such as the fact that all menopausal people are women or that they tend to be older.

We next evaluate the effect of hormonal treatment. We see here that those who received hormonal treatment generally have a lower survival time. We also see that there is no overlap between the confidence intervals indicating that not only did they have a large difference, but it sustained itself throughout the entire distribution of patients.

Finally, we evaluate the effect of chemotherapy. Here we see that generally chemotherapy patients live longer. However, this only becomes significant at the 75th percentile. This means that generally patients that receive chemotherapy are not significantly affected until they live for a long time.

2.3 Test of Differences Between the Survival Curves

(c) Conduct a single test of differences between the survival curves. Justify your choice of test.

Table 7: Stratified log-rank test for differences in menopause.

	χ^2	df	p-value
Stratified by Age	12.05	1	0.0005177
Stratified by Size	42.53	1	0.0000000
Stratified by Hormon	44.06	1	0.0000000

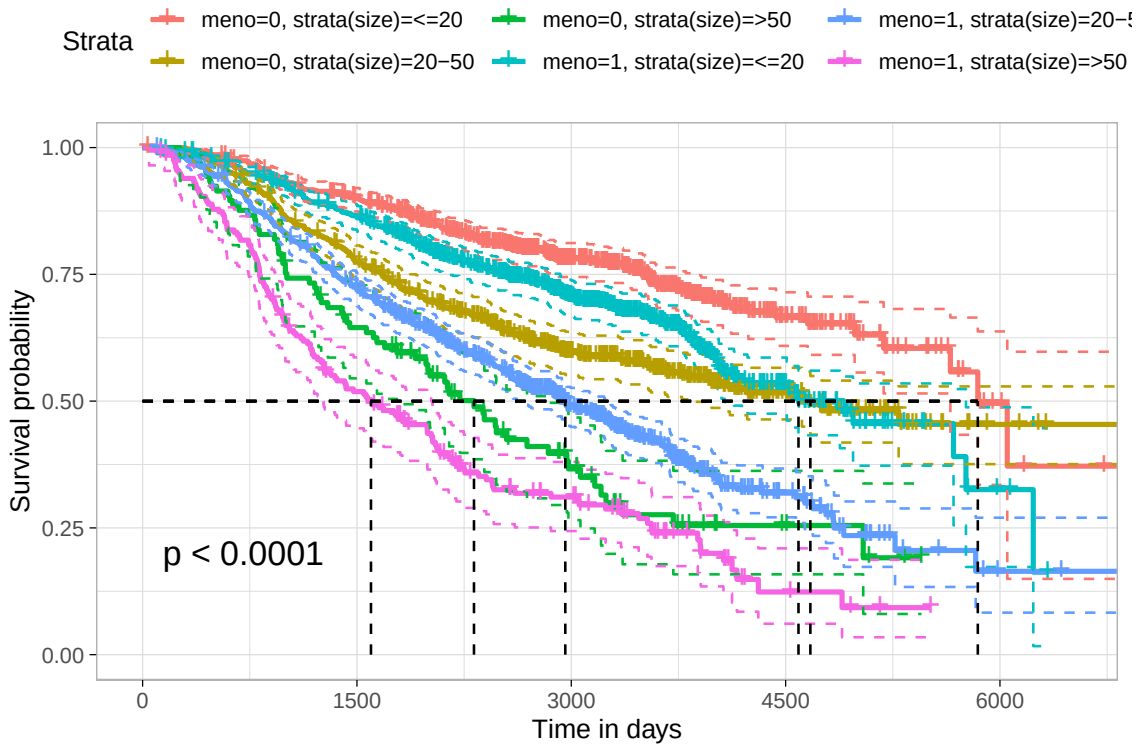


Figure 1: Independent variable meno stratified by size of the tumor.

In this setting we chose to use the log-rank test. This is because we do not want to weight any observations more heavily than others. Thus, we W(Y(j)) = 1. The log-rank test is the most powerful so it is the best choice. However, this test is only the most powerful when the hazard rates are proportional to each other within populations. For this reason, we stratify the model to identify its impacts and make sure no assumptions are violated.

What we see when we perform this stratification is a very large difference between someone without menopause and small cancer cells and someone who is menopausal and has very large cancer cells. We also see that stratifying on age has the best fit.

3 Modeling

3.1 Semi Parametric PH models

For Semi Parametric Proportional Hazards Model, We proceed as follows:

The variable size is a categorical variable with 3 levels so we recode the levels 20-50 and >50 as binary variables and take the level <=20 as reference level. The same holds for the variable grade which has 2 levels. So the level grade=3 is coded as a new binary variable and grade=2 as reference level.

In the dataset, there are 609 ties between times until death consisting of 1376 subjects in total. Considering the existence of tied observations, Efron approximation for the likelihood function will be implemented while estimating the PH model.

Lastly, rtime and recur variables are excluded from the model since they are not the chosen event of interest. We then estimate the full semi parametric proportional hazard model.

For model selection with all covariates, we first apply backward selection procedure to the fitted full model. By excluding the covariate with the largest p-value based on Chi-Square statistics in each iteration, we ended up with the model "dtime,death~meno+hormon+size+ age+grade+nodes+pgr" where all covariates are statistically significant. As an alternative we also run stepwise AIC selection in both direction which selecting the model with the lowest AIC among all possible combinations of models with one more or one less covariate in each iteration. It resulted the model with size, age, grade, nodes, pgr as variables. In order to decide which criteria should be used to choose the model, we approach from 2 different points. From a logical point of view, the model including hormonal treatment seems reasonable since it has an influence on both death time and censoring time of subjects as we explored in the exploratory data analysis part. If we compare AIC of both models, we see that there is no huge difference (18543 and 18546, respectively). So we decide to go with the model selected based on p-value since it includes hormon in exchange of acceptable increase in AIC.

Table 8: Relative Risks and CI for Every Pairs of Levels of Categorical Variables

	vs Level	Relative Risks ($exp(\hat{\beta})$)	95% CI Lower Bound	95% CI Upper Bound
meno=1	0	1.05556486724745	0.870128167355421	1.28052076782386
hormon=1	0	0.934820454679815	0.786310013406322	1.1113800760365
size20-50	size>=20	1.55734520613062	1.3701840278361	1.77007178728268
size>50	size>=20	2.28620352586905	1.91207238107565	2.73354011774165
age i	age i-1	1.0135491231301	1.0061191593931	1.02103395547844
grade=3	grade=2	1.37050188937117	1.19294731006471	1.5744831418146
nodes i	nodes i-1	1.07626939333738	1.0662294206514	1.08640390576272
pgr i	pgr i-1	0.999618567355978	0.999387452733015	0.999849735425649
size>50	size20-50	1.4680133324771	1.25212262928797	1.72112786233726

Estimation and CI of relative risks for every pair of levels of the categorical variables are calculated and shown in the table above. With the aim of checking constant proportional hazard assumption, we run a statistical test based on Schoenfeld residuals for each covariate included in the fitted model with the null hypothesis of residuals being independent from time. From the output below, one can see that test is statistically highly significant for the covariates pgr and age with nodes and menopause being a borderline case. It shows that relative risks for categorical variables are constant over time while null hypothesis of constant hazard ratio is rejected for continuous variables.

##The global test is also highly significant and the null hypothesis can be rejected. ##So we can conclude that proportional hazard assumption of constant hazard ratio over time is violated since there is significant dependency between Schoenfeld residuals and time. **??? Do you agree we should interpret the test outputs excluding cont variables as above since its stated in the project pdf that we are interested in relative risks for categorical variables or should we just look at global test and say that assumption is violated?**

3.2 Parametric Regression Models

We first evaluate parametric models based on variables already selected in the previous section. Here we are comparing log (normal), weibull, exponential, and log (logistic) distributions. This is done by using AIC and identifying if the additional parameters are worth it. What we see is that the log(normal) distribution has the best AIC at 2.4031876×10^4. It should be noted that these AIC values can not be directly compared to the semi-parametric values. We are able to estimate the linear coefficients to be the negative of the log(normal) coefficients due to the normality of this distribution. These can be see in Table 12.

Table 9: Proportional Hazard Test

	chisq	df	p
meno	4.3296389	1	0.0374542
hormon	0.7301267	1	0.3928421
as.factor(size)	5.2190506	2	0.0735695
age	13.9604002	1	0.0001867
as.factor(grade)	3.1603999	1	0.0754447
nodes	3.9615599	1	0.0465505
pgr	41.7855085	1	0.0000000
GLOBAL	60.9142139	8	0.0000000

Table 10: Parametric Model Evaluation

	AIC
log(normal)	24031.88
weibull	24120.92
exponential	24259.09
log(logistic)	24045.35

```
## Warning: package 'broom' was built under R version 4.0.5
```

Table 11: Parametric Model Coefficient Comparison

	estimate	std.error	statistic	p.value	2.5 %	97.5 %	Linear Coefficient
(Intercept)	9.36	0.14	65.91	0.0000000	9.08	9.64	-9.36
meno1	-0.07	0.08	-0.87	0.3856042	-0.22	0.09	0.07
hormon1	0.16	0.07	2.08	0.0379496	0.01	0.30	-0.16
as.factor(size)20-50	-0.34	0.05	-6.76	0.0000000	-0.44	-0.24	0.34
as.factor(size)>50	-0.61	0.08	-7.60	0.0000000	-0.77	-0.45	0.61
age	-0.01	0.00	-2.98	0.0029221	-0.01	0.00	0.01
as.factor(grade)3	-0.27	0.05	-4.97	0.0000007	-0.38	-0.17	0.27
nodes	-0.08	0.01	-14.38	0.0000000	-0.09	-0.07	0.08
pgr	0.00	0.00	4.81	0.0000015	0.00	0.00	0.00

When we look at our categorical variable, hormon, it is significant and the confidence interval confirms this as it does not contain 0. However, it should be noted that this is a borderline case.