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 Introduction

The Rotterdam dataset consists of observations taken from 2982 patients diagnosed with breast cancer. Table 1 explains the covariates in the dataset.

In many medical studies, time to death is the event of interest. However, in cancer, another important measure is the time between response to treatment and recurrence or relapse-free survival time. It is important to state what the event is and when the period of observation starts and finishes.

In this report we will analyze *overall survival* - the time from date of curative surgery to the time of death. In our dataset censoring variable is variable *death*, and the time variable is *dtime*. Hence, recurrence variables *recur*, and *rtime* will be excluded in the further analysis.

In the Rotterdam dataset we encounter right censored data only. Right censoring arises largely from the fact that only some individuals have experienced the event and, subsequently, survival times will be unknown for a subset of the study group. This phenomenon may arise in the following ways:

- A patient has not (yet) experienced the relevant outcome, such as relapse or death, by the time of the close of the study.
- A patient is lost to follow-up during the study period.
- A patient experiences a different event that makes further follow-up impossible.

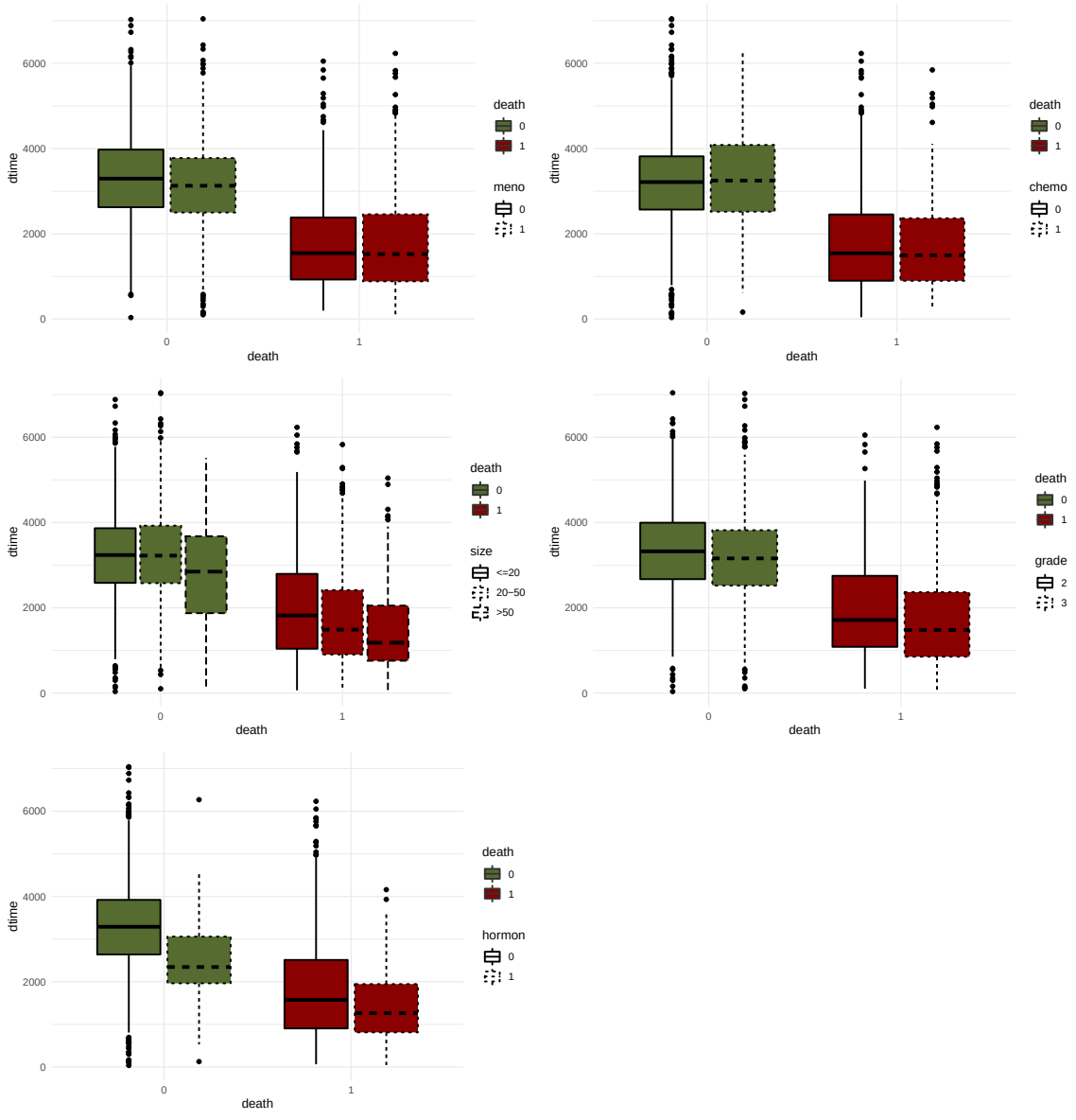
Table 1: Data description

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

2 Exploratory Data Analysis

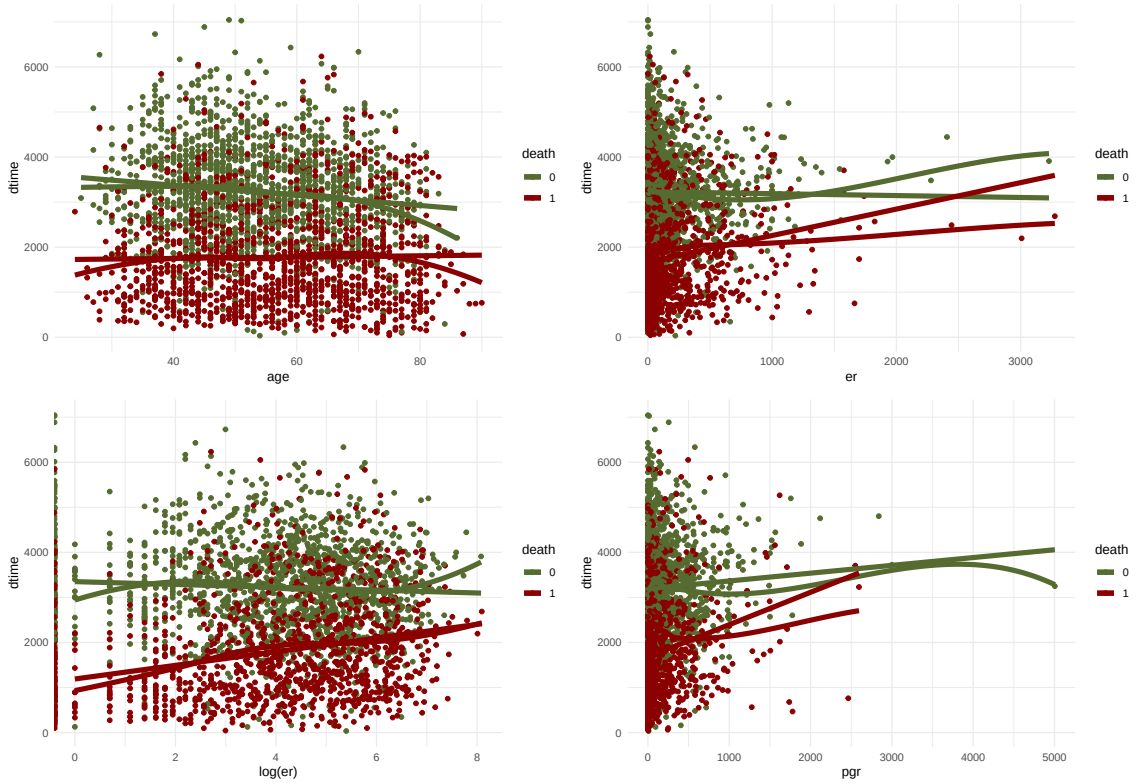
2.1 Categorical Variables

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 recurrence 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 relapse 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.



2.2 Continuous Variables

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.



3 Further Analysis

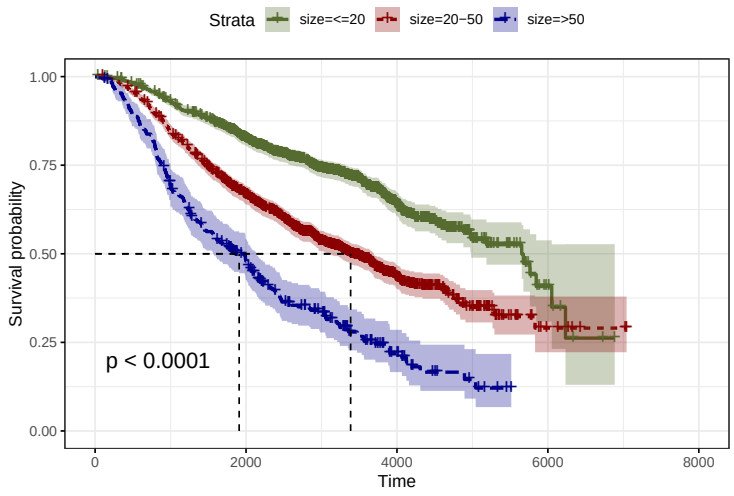
Now the focus will be on the response variable, the censoring indicator, and the categorical variable.

3.1 Survival Distrubution by Levels of *Size*, *Meno*, *Hormon*, and *Chemo*

For each of the levels of the categorical variables, we will compute the survival distribution and interpret the results.

The Kaplan-Meier curve illustrates the survival function. It's a step function illustrating the cumulative survival probability over time. The curve is horizontal over periods where no event occurs, then drops vertically corresponding to a change in the survival function at each time an event occurs.

3.1.1 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	119.40159	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). E.g. on the day 4000, the probability of survival for patients whose cancer size exceeds 50 is approximately 25%, while the same probability for the patients with cancer size less than 20 stands around 62%.

Table 2 shows that median survival times for patients with cancer size smaller than 20, between 20 and 50, and greater that 50 is respectively: 5653, 3386, and 1909. 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.

The same conclusion can be made upon visual inspection of the Kaplan-Meier curve. The blue curve displaying survival probability of patients with cancer size greater than 50 is always below the curves representing other patient groups, once again indicating worse survival time for patients with larger tumors.

3.1.2 Menopause

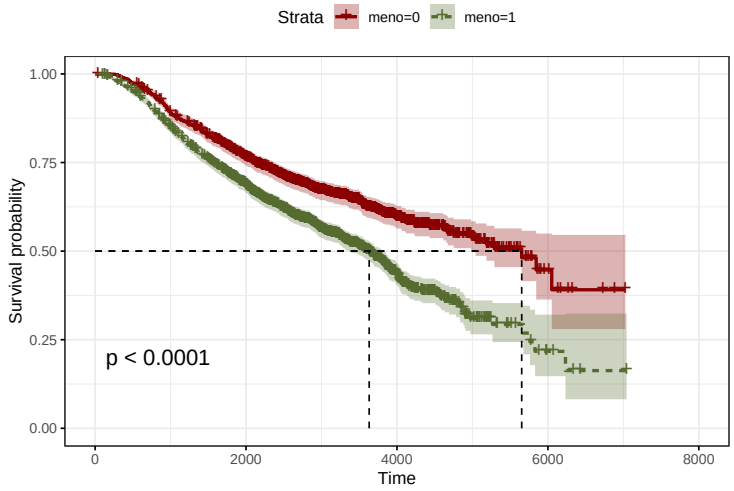


Table 3: Median survival times for each group.

	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

Table 3 shows that the median survival time is 5653 for non-menopausal patients, and 3632 for menopausal, suggesting worse survival for patients that have gone through menopause.

3.1.3 Hormonal Treatment

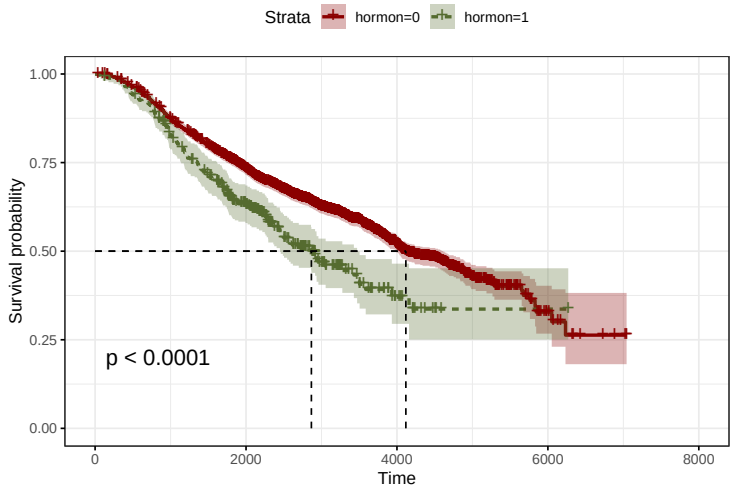


Table 4 shows that the median survival time is 4118 for patients that went through hormonal

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

therapy, and 2866 for those who did not, suggesting slightly worse survival for patients that have gone through the hormonal therapy.

3.1.4 Chemotherapy

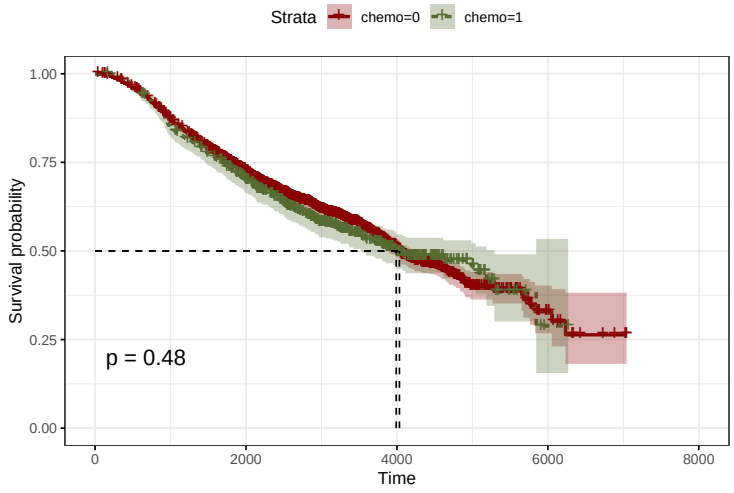


Table 5: Median survival times for each group.

	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

Table 5 shows that the median survival time is 4118 for patients that went through chemotherapy, and 2866 for those who did not, suggesting slightly better survival for patients that have gone through chemotherapy. The reason for this could be that chemotherapy is a very aggressive treatment that is detrimental to a patient's immune system. Due to aggressive nature of chemotherapy it is typically reserved for patients with severe cancer. This is a plausible reason why the patients that have gone through chemotherapy tend to have shorter lives overall.

3.2 Confidence Intervals and Estimators by Levels of *Size*, *Meno*, *Hormon*, and *Chemo*

For each level we will obtain an appropriate estimator and confidence interval for the 3 quartiles of the survival curves and interpret the results.

We first evaluate *size* and see similar results to what was observed in the EDA bar charts. 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.

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$20)	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*

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, and 50th percentiles, there is no overlap in the quantiles' confidence intervals. This means that up until 75th percentile, 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. After the 3rd quantile, we lose information about patients in the non-menopausal group.

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 up until the 75th percentile, when we loose information on the group that did not go through the treatment.

Finally, we evaluate the effect of chemotherapy. Here we see that generally chemotherapy patients live shorter. 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.

Note that if one of the groups has not yet dropped to 50% survival at the end of the available data, we cannot compute a median survival and there will be NA values for median survival in such cases. The same holds for other quantiles.

3.3 Test of Differences Between the Survival Curves

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. 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 subjects, and perform stratified log-rank test in order to obtain an overall assessment of the difference between survival curves, by combining information over the different strata to gain power.

In Section 3.1 we observed differences between survival curves by levels of different categorical variables. In this section, we will also evaluate whether this difference is statistically significant using a formal statistical test - the stratified log-rank test.

We focus on covariate *meno*, and stratify by *chemo*, *size*, and *hormon.*

We observe significant differences between survival curves of each level of variable *meno* after conducting the stratified test (see Table 7), indicating that menopausal status is an important factor in survival time for breast cancer patients.

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

	χ^2	df	p-value
Stratified by Chemo	69.590	1	0
Stratified by Size	42.530	1	0
Stratified by Hormon	44.056	1	0

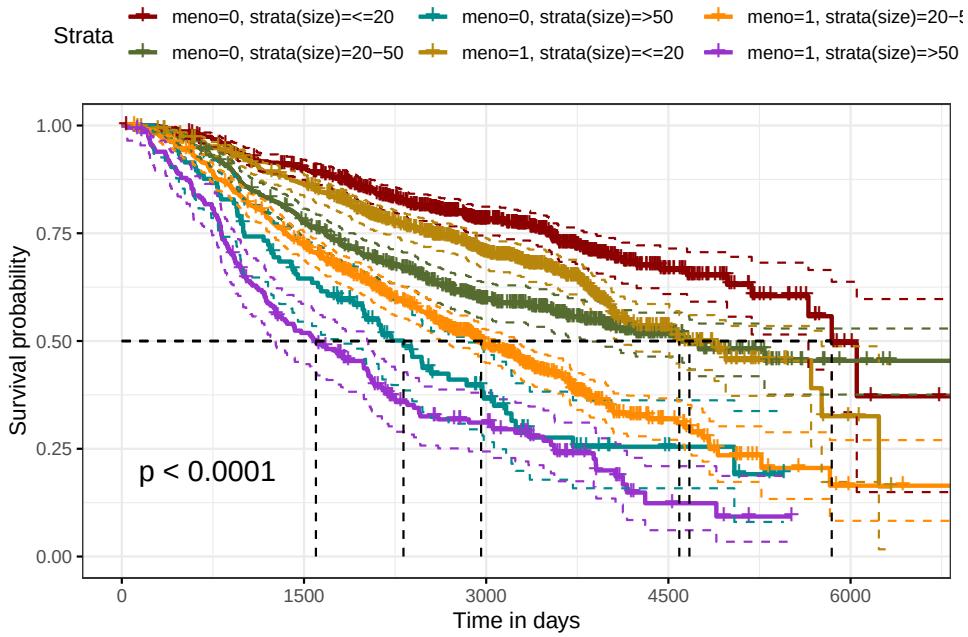


Figure 1: Independent variable meno stratified by size of the tumor.

What we see in Figure 1 when we perform this stratification by the size of the breast cancer, is a very large difference between someone who has not gone through menopause and has small cancer cells, and someone who is menopausal and has very large cancer cells.

4 Modeling

4.1 Semiparametric Proportional Hazards Model

For semiparametric proportional hazards model, we proceed as follows:

The variable *size* is a categorical variable with 3 levels. We recode the levels Size $(20-50)$ and Size (>50) as binary variables, and take the level Size (≤ 20) as reference level. The same holds for the variable *grade*, which has 2 levels. The level Grade = 3 is coded as a new binary variable, with Grade = 2 being the 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, when estimating the full model, variables *rtime* and *recur* are excluded, since they are not the chosen event of interest. We then estimate the full semiparametric proportional hazard model.

Table 8: AIC Values for Models Selected by Stepwise AIC Procedure, and Backward Selection Procedure

	AIC Selection	Backward Selection
AIC	18543.63	18546.82

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 χ^2 statistics in each iteration, we ended up with the model where significant covariates are *size, age, grade, nodes, pgr, meno, hormon*.

As an alternative model selection procedure, we run stepwise AIC selection in both directions, which compares models consisted of all possible combinations of covariates, and selects the one with the lowest AIC value. This resulted with the model where significant covariates are *size, age, grade, nodes, pgr*.

To decide which procedure should be used to choose the model, we approach from 2 different points. From a logical point of view, the model including variable *hormon* seems reasonable, since hormonal treatment has an influence on both death time and censoring time of subjects, as we have seen in Section 2. On the other side, if we compare AIC of both models, from Table 8 we see that there is no big difference between the two models (1.8544×10^4 and 1.8547×10^4, respectively). That is why we decide to use the model selected by the backward selection procedure, as it includes variable *hormon* in exchange of acceptable increase in AIC.

Estimation and CI of relative risks for every pair of levels of the categorical variables are calculated and shown in the Table 10.

With the purpose 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 Table 11, 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

Table 9: Estimated Coefficients of Covariates Under the Selected Model

	$\hat{\beta}$
Meno = 1	0.0540760
Hormon = 1	-0.0674008
Size (20-50)	0.4429826
Size (>50)	0.8268926
Age	0.0134582
Grade = 3	0.3151770
Nodes	0.0735008
Pgr	-0.0003815

Table 10: Relative risks and CI for each covariate

	$exp(\hat{\beta})$	5%	95%
Meno = 1 vs Meno = 0	1.056000	0.870000	1.281000
Hormon = 1 vs Hormon = 0	0.935000	0.786000	1.111000
Size $20-50$ vs Size ≤ 20	1.557000	1.370000	1.770000
Size > 50 vs Size ≤ 20	2.286000	1.912000	2.734000
Age = i vs Age = i - 1	1.014000	1.006000	1.021000
Grade = 3 vs Grade = 2	1.371000	1.193000	1.574000
Nodes = i vs Nodes = i - 1	1.076000	1.066000	1.086000
Pgr = i vs Pgr = i - 1	1.000000	0.999000	1.000000
Size >50 vs Size $20-50$	1.468013	1.252123	1.721128

Table 11: Proportional Hazard Test

	χ^2	df	p-value
Meno	4.3296	1	0.0375
Hormon	0.7301	1	0.3928
Size	5.2191	2	0.0736
Age	13.9604	1	0.0002
Grade	3.1604	1	0.0754
Nodes	3.9616	1	0.0466
Pgr	41.7855	1	0.0000
GLOBAL	60.9142	8	0.0000

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?**

4.2 Parametric Regression Models

We first evaluate parametric models based on variables already selected in the Section 4.1. Here we are comparing log-normal, Weibull, exponential, and log-logistic AFT models. This is done by

Table 12: Parametric Model Evaluation

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

using AIC selection, and identifying if the additional parameters are worth it. What we see is that the log-normal model has the best AIC at 2.4031876×10^4. It should be noted that these AIC values can not be directly compared to the semiparametric values.

The parametric models considered here have two representations:

- Accelerated failure time model (AFT)
- Linear model

Table 12 shows point and interval estimates of the coefficient of variables both in the log-normal AFT model, and in the linear representation of the log-normal model. We are able to estimate the linear coefficients of the log-normal model due to the normality of it's error distribution.

Table 13: 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
Meno = 1	-0.07	0.08	-0.87	0.3856042	-0.22	0.09	0.07
Hormon = 1	0.16	0.07	2.08	0.0379496	0.01	0.30	-0.16
Size (20-50)	-0.34	0.05	-6.76	0.0000000	-0.44	-0.24	0.34
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
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*, we can see that it is significant. However, the significance is borderline.