Using R Language for Stratification, Parametric Modeling, and Coxph Model in the Analysis of Primary Breast Cancer Data

Horngjiun Chao, Jessica Grant, Yutong Xue, Hao Nan Wang

2023-04-14

Section 1

Overview and Introduction

Introduction

- **Introduction:** The effectiveness of primary breast cancer treatment varies based on patient characteristics, including the use of hormonal therapy, chemotherapy, or both. Studying the outcomes of patients who receive these treatments can help optimize breast cancer treatment for improved results.
- **Research Question:** How does the effect of hormonal therapy and chemotherapy on breast cancer outcomes vary based on patient characteristics

Dataset

- This dataset named "rotterdam", comprised 2982 breast cancer patients whose records were included in the Rotterdam tumour bank. The follow-up time ranged from 1 to 231 months.
  - 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 20-50 >50
  - grade: differentiation grade (grade=2, grade=3)
  - 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 (0=no, 1=yes)
  - 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

Literature Review

- One should know that node-negative patients are often excluded from breast cancer studies because lymph node status is an important prognostic factor for breast cancer, and excluding node-negative patients allows researchers to focus on the prognostic factors most relevant to patients, the researchers Royston and Altman (2013) suggested to omit the node-negative patients for dataset creation.
  - There are 1436 node-negative patients and 1546 node-positive patients in this dataset

Data Manipulation

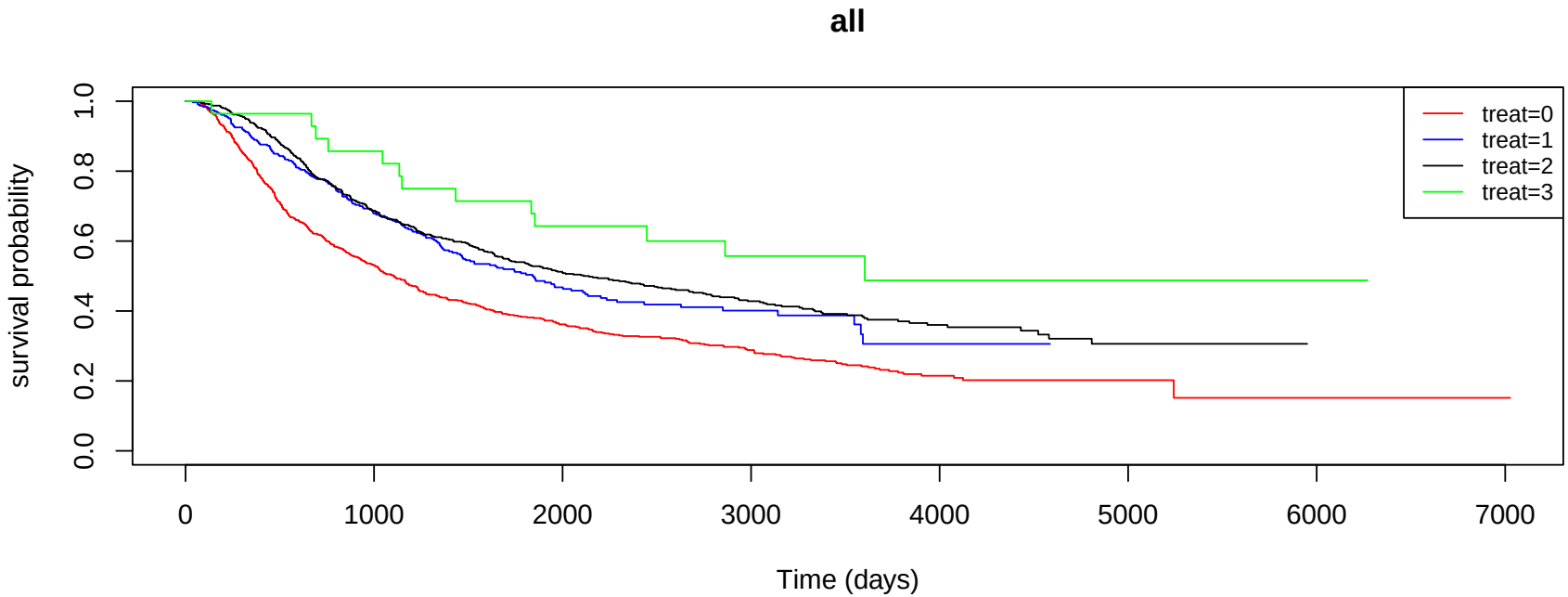
- In this presentation, the variable *treatment* is defined as the four possible combinations of hormonal treatment and chemotherapy by creating a cross-tabulation of *hormon* and *chemo*
  - Treatment 0: no treatments
  - Treatment 1: hormonal treatment only
  - Treatment 2: chemotherapy only
  - Treatment 3: both treatments

Section 2

Stratification

Logrank Test for K=4 Samples

- Consider K=4 groups with survival probability $S_j(t)$ for $j = 1, 2, 3, 4$. The null hypothesis is $H_0 : S_1(t) = S_2(t) = S_3(t) = S_4(t)$



By using logrank test, the result shows that there is significant evidence ($p<0.05$) to reject the null hypothesis that the relapse time distributions in the fours groups are the same. Here we see that there is less difference in the estimated survival functions early on than later.

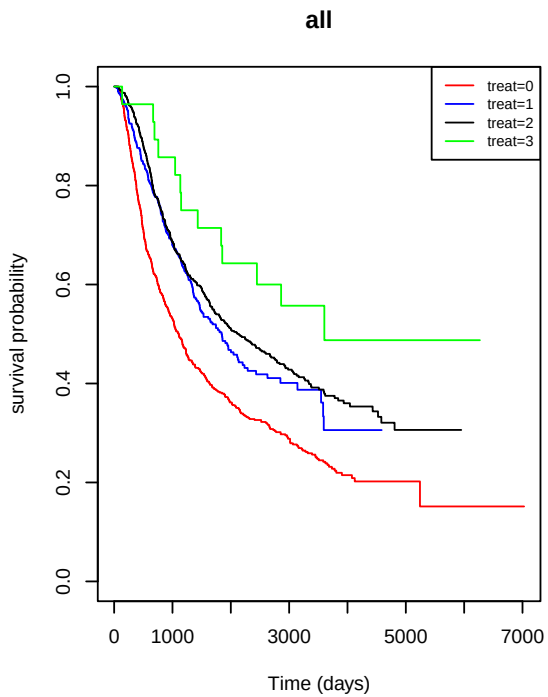
Stratified Logrank Test

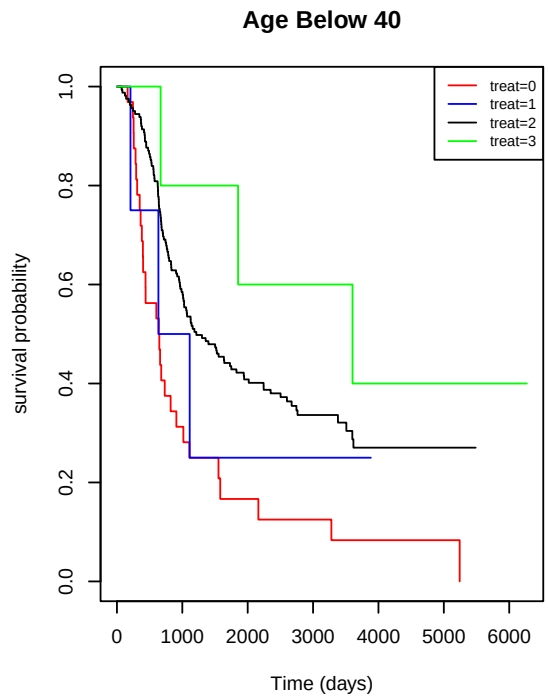
- One may be concerned about the confounding effects that other factors may have on the interpretation of the relationship between survival and groups
- Comparisons of survival distribution between groups are made within each stratum and then these each results are combined across strata
- In this section, stratification is based on *age*, *menopausal status (meno)*, *size*, *grade*
- The resulting stratified logrank test has a standard normal distribution (asymptotically) under the null hypothesis. For K samples, the degree of freedom should be K-1

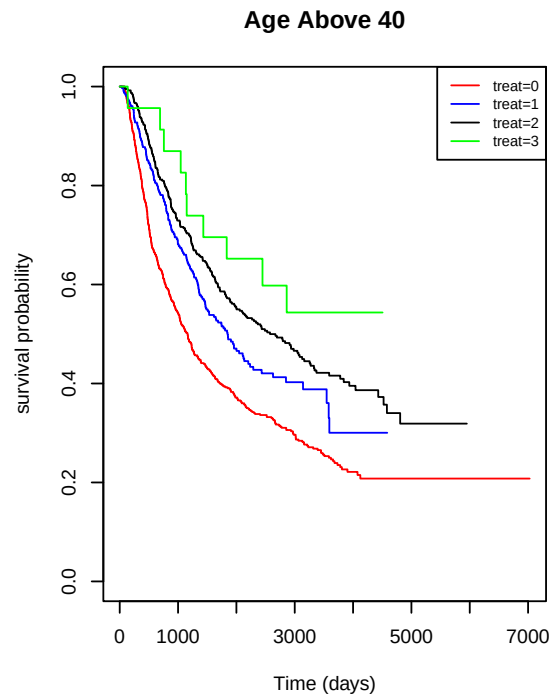
$$[T(w)]^2 \overset{a}{\sim} \chi^2_{K-1}$$

Stratification on Age

- According to the World Health Organization (WHO), "approximately half of breast cancers develop in women who have no identifiable breast cancer risk factor other than gender (female) and age (over 40 years)" (WHO, 2021). In the dataset, *age* is a continuous variable, it is reasonable to stratify it using a cut-off point of 40 years old.







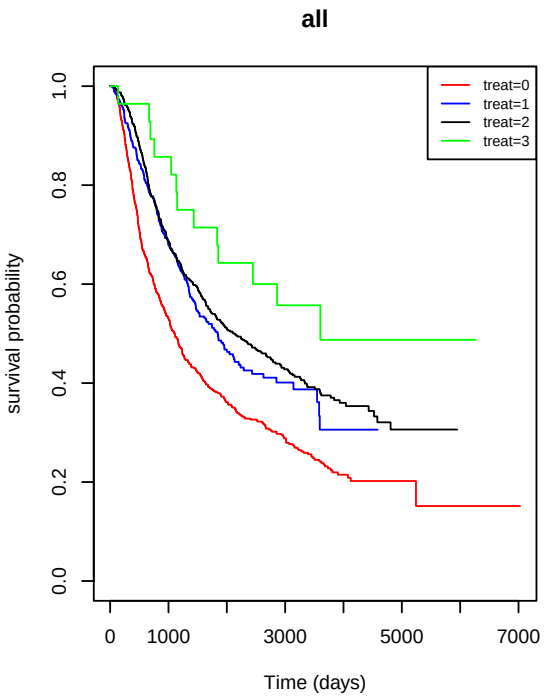
Stratification on Age

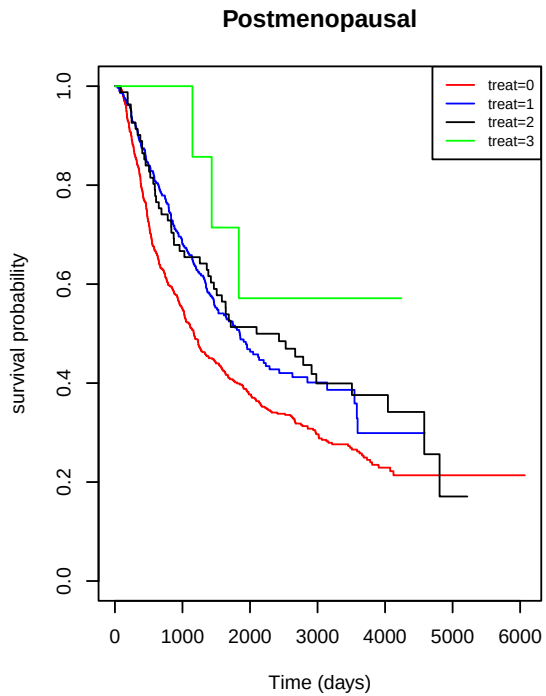
	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Treatment = 0	655	468	355.5	35.57	59.31
Treatment = 1	311	169	184.2	1.25	1.62
Treatment = 2	552	324	409.5	17.84	35.49
Treatment = 3	28	13	24.8	5.62	5.79

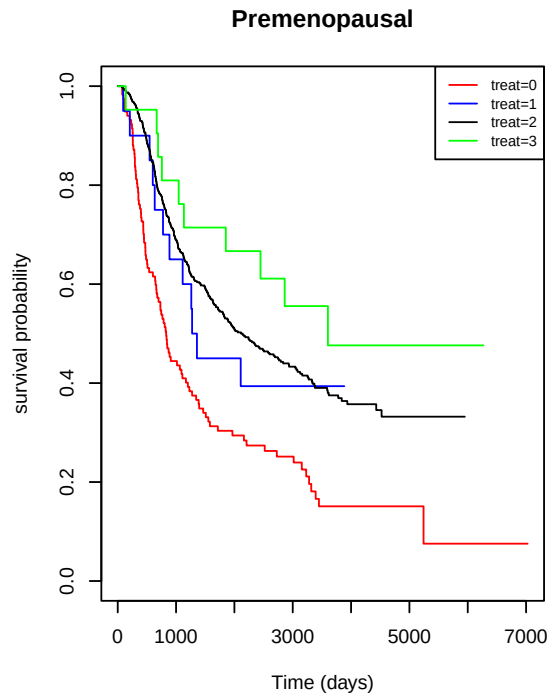
```
## Chisq= 65.8  on 3 degrees of freedom, p= 3e-14
```

Stratification on menopausal status (meno)

- Researchers such as Surakasula, Nagarjunapu, and Raghavaiah study pre- and post-menopausal breast cancer due to the impact of menopausal status on treatment response. Menopause increases the prevalence of hormone receptor-positive breast cancer cells, which can be targeted with hormone therapy (Surakasula et al., 2014).







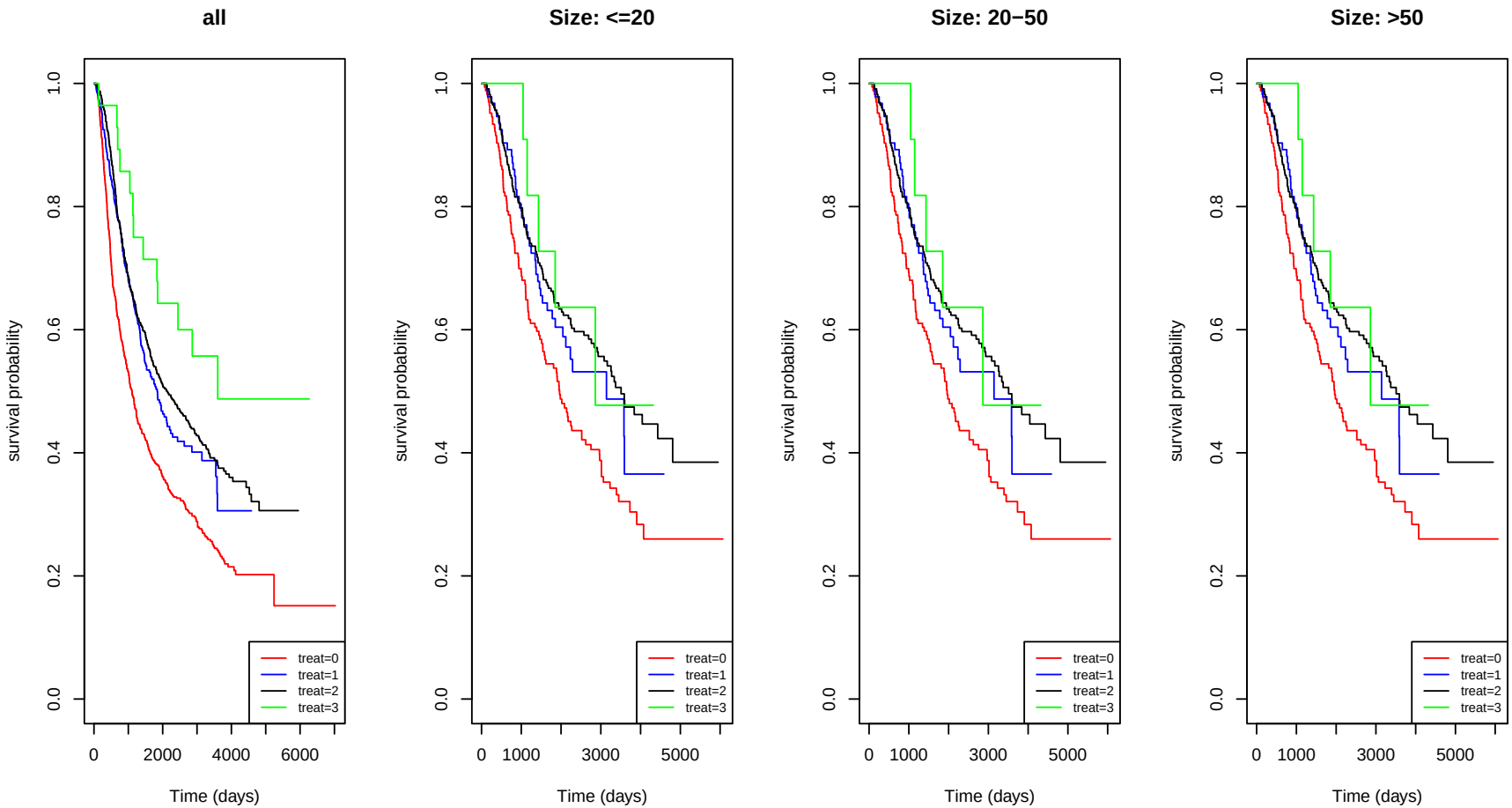
Stratification on menopausal status (meno)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Treatment = 0	655	468	379.1	20.84	41.96
Treatment = 1	311	169	200.1	4.85	7.09
Treatment = 2	552	324	371.4	6.05	18.85
Treatment = 3	28	13	23.3	4.56	4.73

```
## Chisq= 48.2  on 3 degrees of freedom, p= 2e-10
```

Stratification on tumor size (size)

- size $=$ 1: size $\leq$ 20
- size $=$ 2: size $=$ 20-50
- size $=$ 3: size $\geq$ 50

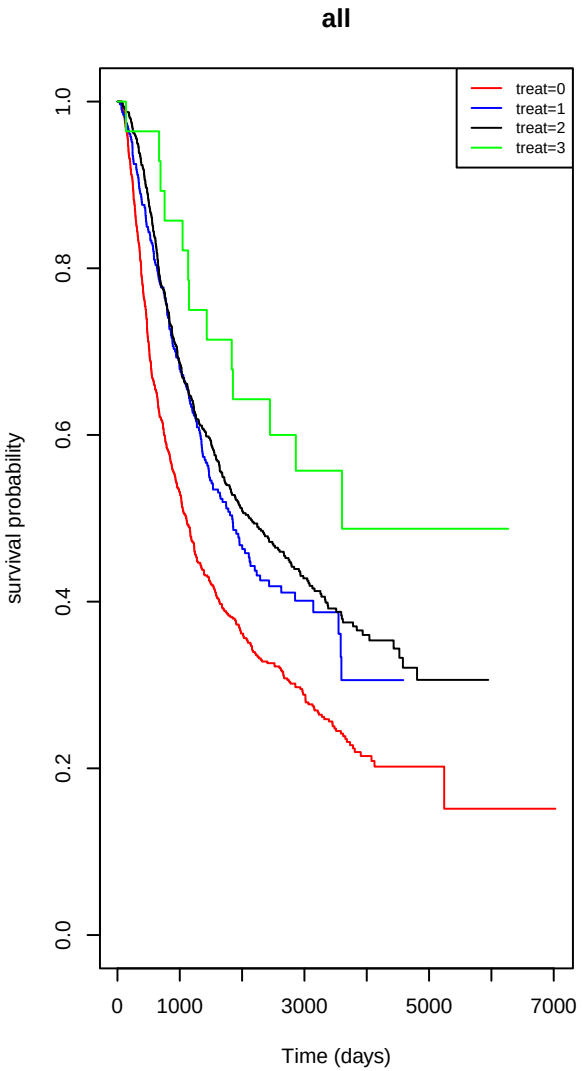


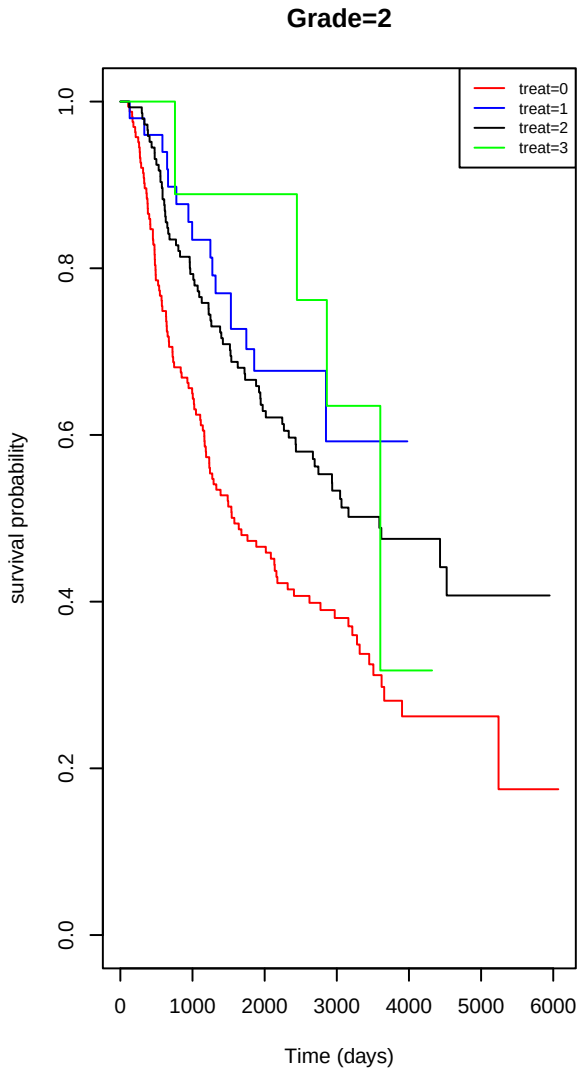
Stratification on tumor size (size)

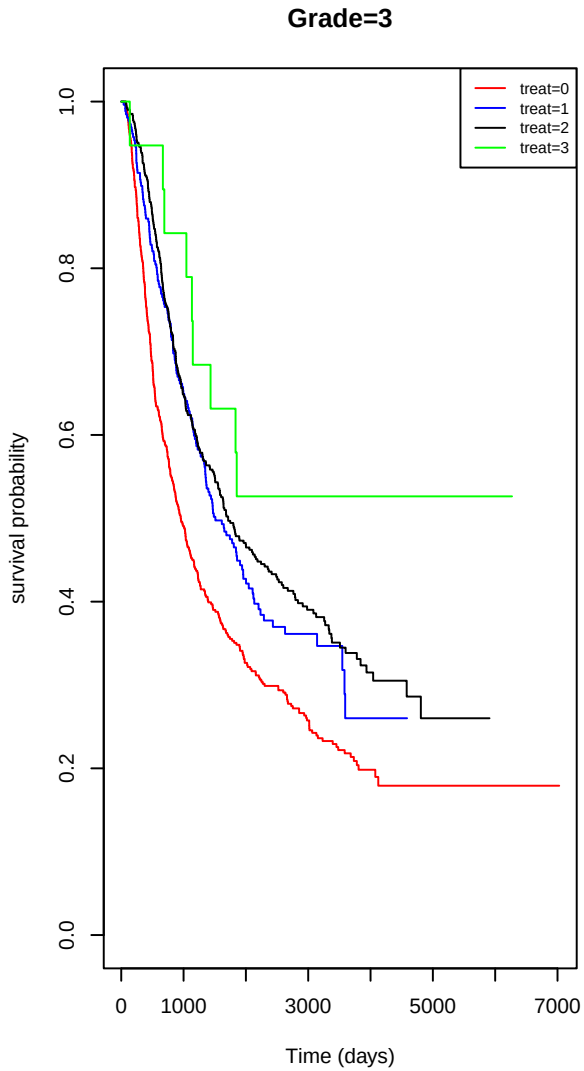
	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Treatment = 0	655	468	373.9	23.69	39.01
Treatment = 1	311	169	194.1	3.25	4.12
Treatment = 2	552	324	383.2	9.15	15.43
Treatment = 3	28	13	22.8	4.21	4.35

```
## Chisq= 40.9  on 3 degrees of freedom, p= 7e-09
```

Stratification on differentiation grade (grade)







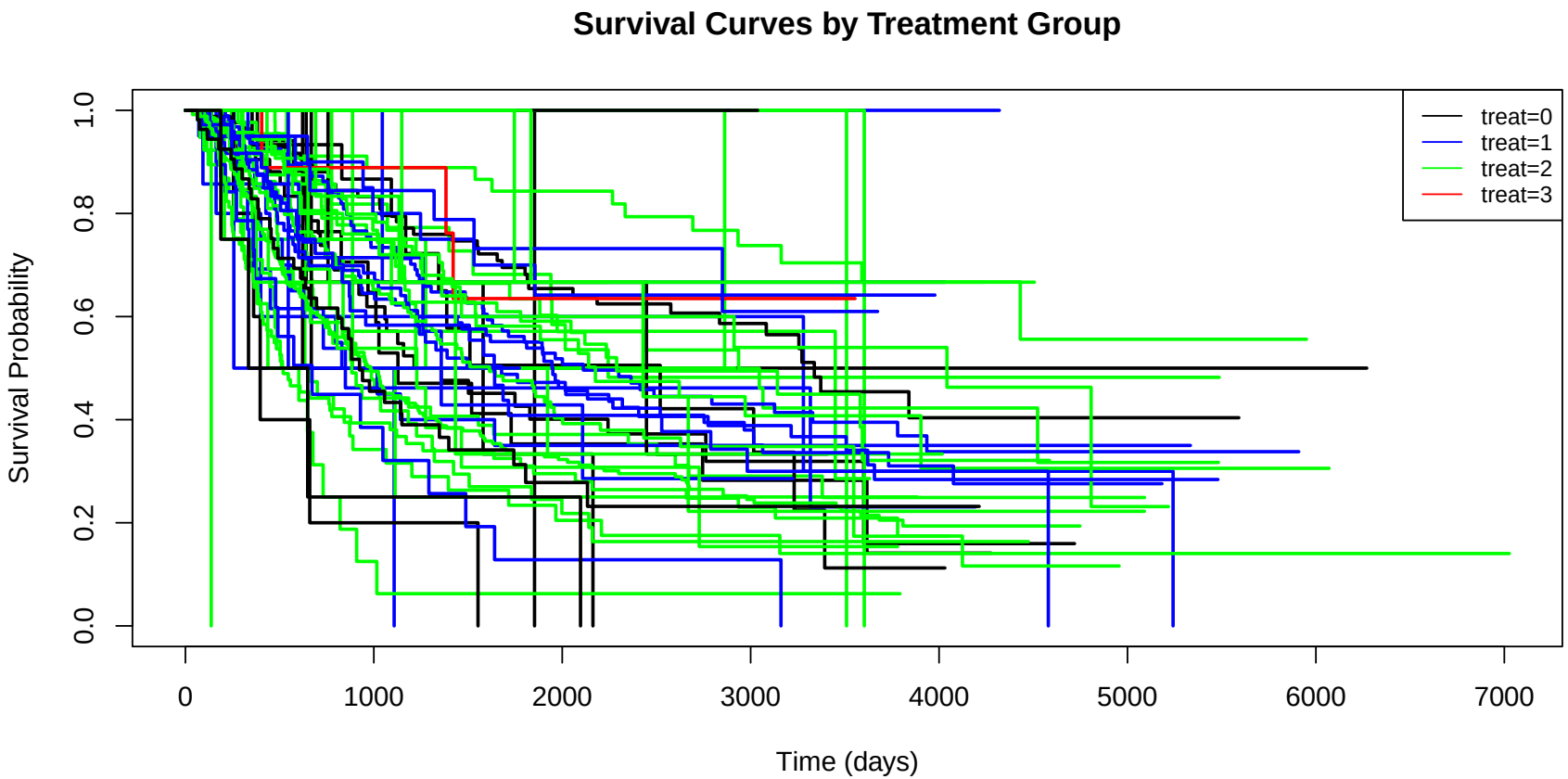
Stratification on differentiation grade (grade)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Treatment = 0	655	468	359.3	32.87	52.30
Treatment = 1	311	169	195.2	3.52	4.48
Treatment = 2	552	324	395.7	12.98	22.03
Treatment = 3	28	13	23.8	4.89	5.02

```
## Chisq= 54.5  on 3 degrees of freedom, p= 9e-12
```

Stratification on all strata

- Not informative because there are 24 strata and 4 groups
- In such cases, one may want to adjust for the effect of these factors through regression modelling

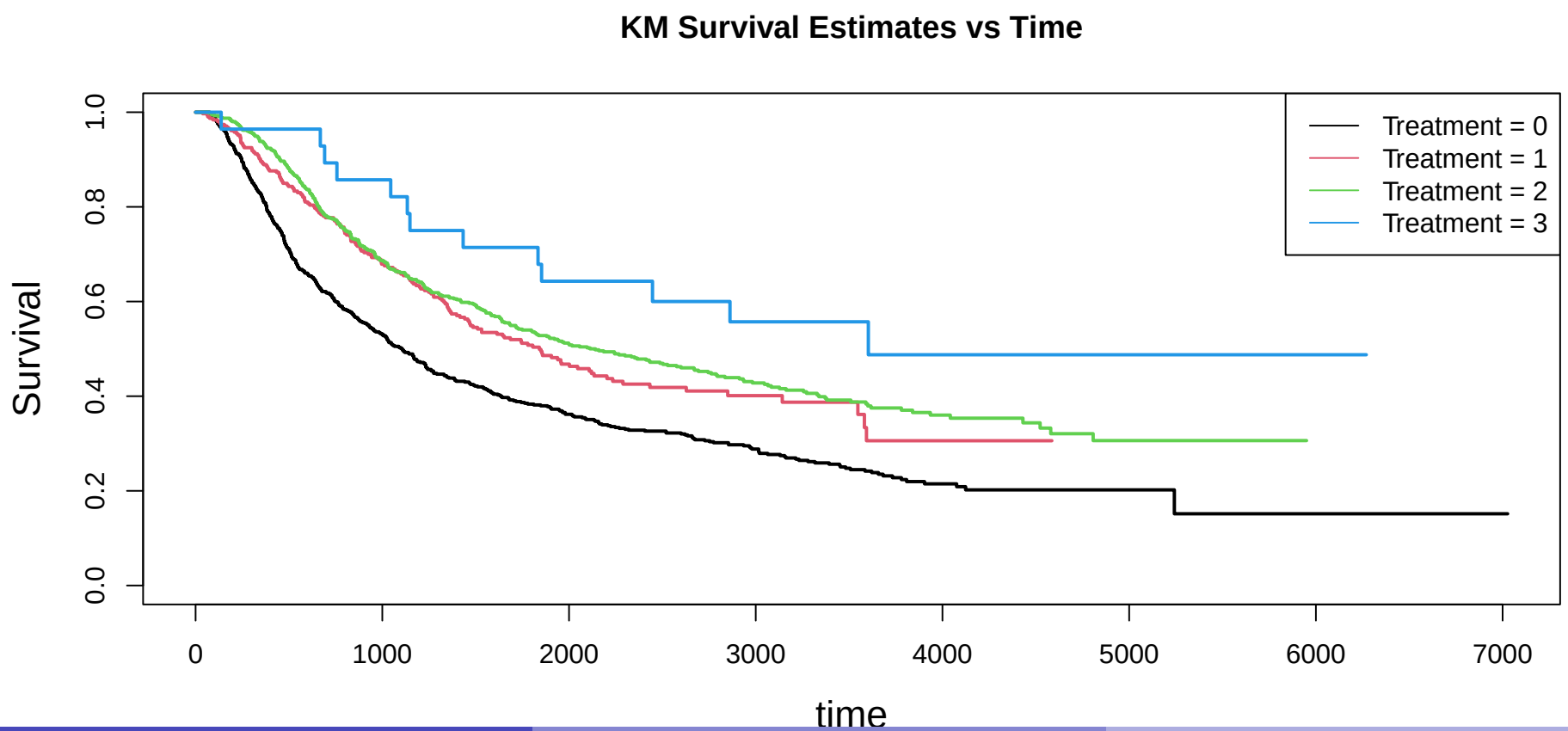


Section 3

Parametic Models

Kaplan Meier Plot

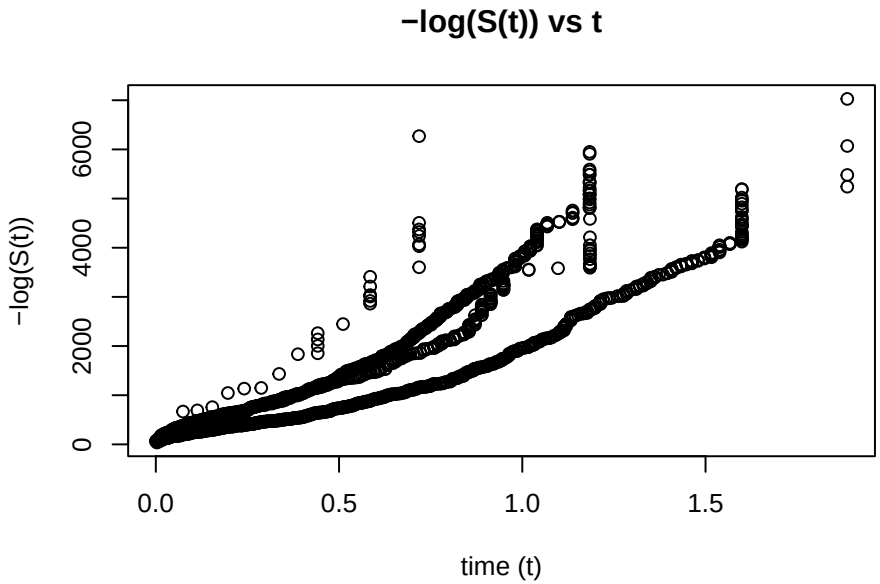
In order to determine whether the data follows a known distribution and a parametric model can be applied, we compute the Kaplan Meier estimates of the survival function and plot them as well as transformations of survival curves.

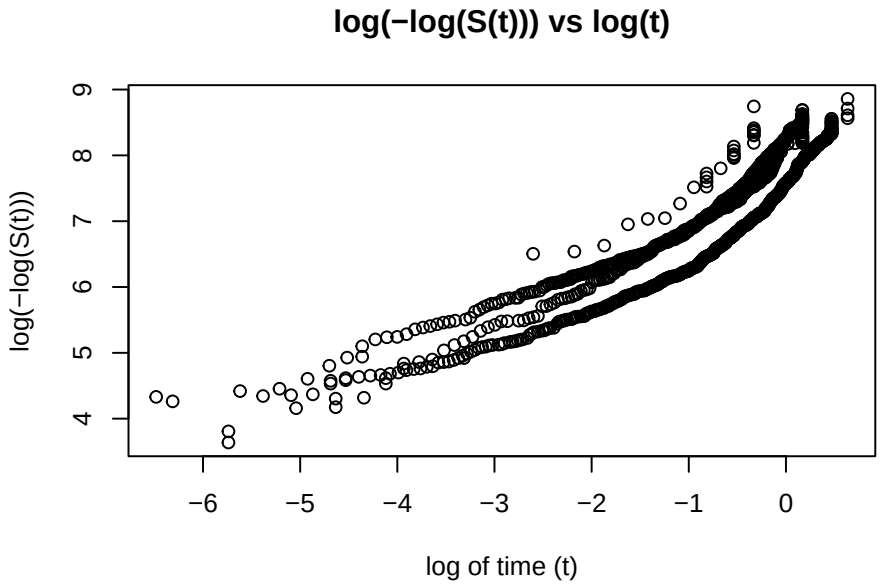


Transformation of Kaplan Meier Plots

- Exponential model is appropriate if points are linear for $\log S(t)$ versus $-\lambda t$. - Weibull model is appropriate if points are linear for $\log(-\log(S(t)))$ versus $\log t$.

Considering the plots below, it appears that the exponential model may be most appropriate for our data, however to be confident, we will apply the the Likelihood Ratio test.





Exponential vs Weibull Underlying Distribution

Recalling that the survival and hazard functions of the exponential model assume the following form.

$$S(x) = \exp(-\lambda x), \quad x \geq 0, \quad \lambda > 0$$
$$h(x) = \lambda$$

And the the survival functions of the Weibull model assume the following form.

$$S(x) = \exp(-\lambda x^{\alpha}), \quad x \geq 0, \quad \lambda > 0, \quad \alpha > 0$$
$$h(x) = \lambda \alpha x^{\alpha - 1}$$

To test which model is more appropriate, we are testing the following.

$$H_0 : \alpha = 1 \quad \text{vs} \quad H_a : \alpha \neq 1, \quad \alpha > 0$$

Likelihood Ratio Test

In order to conduct this test we will use the Likelihood Ratio test, which is of the following form;

$$\text{Likelihood Ratio test: } \chi^2_{LR} = -2(\ell(\theta_0) - \ell(\hat{\theta})) \sim \chi^2$$

Based on the R code found in the Appendix, the likelihood ratio test generates the following results.

- LR Test Statistic: 15.91645
- LR Test Critical Value: 3.841459
- LR p-vlaue: 6.620096e-05

Therefore, we reject the null hypothesis that $\alpha = 1$ at the 0.05 significance level, and concludes that there is sufficient evidence to assume a Weibull model over an exponential model.

Weibull Model

The Weibull model is unique in that it can be parameterised as either a proportional hazards (PH) model or an accelerated failure time (AFT) model, and is the only family of distributions to have this property. Therefore if we consider the AFT model with an underlying Weibull distribution, the PH assumption will hold and vice versa.

Weibull Under AFT Model

Under AFT the model is assumed to have the following forms;

$$S(x|Z) = S_0(x \exp(-\gamma^\top Z))$$
$$h(x|Z) = \exp(\gamma^\top Z) h_0\{x \exp(-\gamma^\top Z)\} = \exp(\gamma^\top Z)\lambda\alpha x^{\alpha-1}$$

where $h_0(x)$ is the baseline hazard function which here is the parametric Weibull hazard function, Z is the vector of covariates, γ is the vector of regression coefficients that satisfy $Y = ln(X) = \mu + \gamma^\top Z + \sigma W$ where W is the error distribution.

Weibull Under PH Model

Under PH the model is assumed to have the following forms;

$$S(x|Z) = S_0(x)^{c(\beta^\top Z)}$$
$$h(x|Z) = h_0(x)c(\beta^\top Z) = \lambda\alpha x^{\alpha-1}c(\beta^\top Z)$$

where $h_0(x)$ is the baseline hazard function which here is the parametric Weibull hazard function, Z is the vector of covariates and β is the vector of regression coefficients. When $c(\beta^\top Z)$ is defined to be equal to $\exp(\beta^\top Z)$, this is the Cox PH model.

- For the purpose of this analysis, we will consider the Cox PH model form.

Model Construction

Moving forward with a Cox PH model with the underlying Weibull distribution, we perform feature selection using forward selection.

To begin we create the following set of factors for each time-fixed covariate that may be associated with the timing of staphylococcus infection.

- Cov.Age is a variable containing patients age.
- Cov.Meno is a binary variable indicating menopausal status .
- Cov.Size is a variable containing tumor size with three factor levels.
- Cov.Grade is a binary categorical variable containing cancer grade.
- Cov.Nodes is a variable containing the number of positive lymph nodes.
- Cov.PGR is a variable containing the number of progesterone receptors.
- Cov.ER is a variable containing the number of estrogen receptors.

Forward Selection

Then we used a self-written likelihood ratio local test function to test the hypothesis that the times to relapse are the same for the four different treatment groups using a model which adjusts for each of the factors.

A summary of the results is as follows.

Local Tests

The first set of local tests where only treatment was adjusted for produced the following results.

Table 1: Table 1.1: Local test for possible confounders, adjusted for treatment groups

	df	LR Test Stat	p-value
Age	1	17.089228	0.0000357
Meno	1	3.073005	0.0796022
Size	2	72.673872	0.0000000
Grade	1	34.145773	0.0000000
Nodes	1	142.670530	0.0000000
PGR	1	13.949103	0.0001878
ER	1	12.917903	0.0003255

Based on the p-values and test statistics above, the nodes covariate is

Local Tests Cont'd

This process was repeated until the covariates treatment groups, nodes, size, grade and age were all added to the model and the final set of local tests which adjusted for each of these factors generated the following results.

Table 2: Table 1.5: Local test for possible confounders, adjusted for treatment groups, nodes, size, grade and age

	df	LR Test Stat	p-value
Meno	1	1.096197	0.2951023
PGR	1	3.420193	0.0644035
ER	1	7.305699	0.0068736

Because all of the local tests are significant at the 0.05 significance level, we stop here. The full forward selection resutls code is in the Appendix.

Final Cox PH model with underlying Weibull distribution

From the forward seleciton, our final Cox PH model with underlying Weibull distribution is as follows;

$$h(t|Z) = \lambda\alpha x^{\alpha-1}\exp(\beta^\top Z)$$

$$\begin{aligned} h(t|Z) = \lambda\alpha x^{\alpha-1}\exp(&\beta_1 Treatment_1 + \beta_2 Treatment_2 + \beta_3 Treatment_3 \\ &+ \beta_4 Cov.Nodes + \beta_5 Cov.Size_1 + \beta_6 Cov.Size_2 + \beta_7 Cov.Grade \\ &+ \beta_8 Cov.Age) \end{aligned}$$

Global Hypothesis Test

Our hypothesis test of interest is as follows;

$$H_0: \quad \beta = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0 \quad \text{vs}$$
$$H_a: \quad \beta_i \neq 0 \quad \text{for some i} \in [1, 8]$$

Using the built in R function, we test the above global hypothesis.

Based on the R output for the global likelihood ratio test, the p-value is foudn to be 3e-102.

Therefore, we reject the null hypothesis at the 0.05 significance level and conclude that the distributions of the times to relapse are not the same amoung the treatment groups and covariates.

Local Hypothesis Test

Using the self-written likelihood ratio local test function, we test the following hypothesis;

$$H_0: \quad \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs}$$
$$H_a: \quad \beta_i \neq 0 \quad \text{for some i} \in [1, 3]$$

The R output for the local likelihood ratio test generates the following restults

- Local LR Test Statistic: 53.60846
- Local LR p-vlaue: 1.359923e-11

Therefore, we fail to reject the null hypothesis at the 0.05 significance level and conclude that the distributions of the times to relapse are the same in the four treatment groups when adjusting for tumor nodes, size, grade and patient age.

Discussion

- Royston and Altman suggested to omit the node-negative patients, since nodal status is an important prognostic factor.

Section 4

Model Results after Adjusting the Data

Parametric Model Results after Adjusting the Data

After re-conducting the analysis with the updated dataset we have the following results:

- Transfored KM curves still suggest underlying exponential or Weibull distribution. LR test generates the following results, therefore we conclude that the distribution is still Weibull
  - LR Test Statistic = 4.782031
  - LR Test p-value = 0.02875819

Forward selection ressulted in the following model:

$$\begin{aligned} h(t|Z) = \lambda\alpha x^{\alpha-1} \exp(&\beta_1 Treatment_1 + \beta_2 Treatment_2 + \beta_3 Treatment_3 \\ &+ \beta_4 Cov.Nodes + \beta_5 Cov.Size_1 + \beta_6 Cov.Size_2 + \beta_7 Cov.Grade \\ &+ \beta_8 Cov.Age + \beta_9 Cov.ER + \beta_{10} Cov.PGR) \end{aligned}$$

All related code found in Appendix.

Parametric Model Results after Adjusting the Data Cont'd

Global hypothsis test generated a p-value of 6.6e-52 concluding that the distributions of the times to relapse are not the same amoung the treatment groups and covariates.

Finally, local test on treatment groups generated the following results:

- Local Test LR Test Statistic: 57.69702
- Local Test LR p-value : 1.824353e-12

Therefore we rejected the local null hypothesis at the 0.05 significance level and concluded that the distributions of the times to relapse are not the same in the four treatment groups when adjusting for tumor nodes, size, grade, patient age, the number of progesterone receptors and the number of estrogen receptors.

All related code found in Appendix.

Parametric Model Results after Adjusting the Data - Final Model

The final Cox PH Model with Weibull underlying distribution was found to be as follows;

$$\begin{aligned} h(t|Z) = \lambda\alpha x^{\alpha-1} \exp(&\beta_1 Treatment_1 + \beta_2 Treatment_2 + \beta_3 Treatment_3 \\ &+ \beta_4 Cov.Nodes + \beta_5 Cov.Size_1 + \beta_6 Cov.Size_2 + \beta_7 Cov.Grade \\ &+ \beta_8 Cov.Age + \beta_9 Cov.ER + \beta_{10} Cov.PGR) \end{aligned}$$

$$\begin{aligned} h(t|Z) = 0.00024(1.14250)x^{0.14250} \exp(&-0.53051 Treatment_1 - 0.55324 Treatment_2 \\ &- 0.67710 Treatment_3 + 0.05787 Cov.Nodes + 0.30413 Cov.Size_1 + 0.55124 Cov.Size_2 \\ &+ 0.29166 Cov.Grade - 0.01280 Cov.Age - 0.00030 Cov.ER - 0.00030 Cov.PGR) \end{aligned}$$

All related code found in Appendix.

Section 5

Cox PH Model

Intro of Cox Model

- It used to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time.
- It can be estimated as follow:

$$h(t) = h_0(t) \cdot exp(\beta^\top Z)$$

where

- t represents the survival time
- $h(t)$ is the hazard function determined by a set of p covariates(Z_i)
- the coefficients $\beta^\top$ measure the impact of covariates
- the term $h_0(t)$ is called the baseline hazard.

Analysis of Cox Model

- Interested variables: Age, Meno, Size, Grade, Nodes, Pgr, Er, Treatment
- Forward selection of likelihood tests for possible confounders adjusting for Age, Meno, Size, Grade, Nodes, Treatment(AIC=22396.42).

Variable	DF	Chi_square	P_value	AIC
Pgr	1	1.967	0.161	22396.45
Er	1	0.694	0.404	22397.73

Final Model

- Final Cox model:

$$\lambda(t|Z_i) = \lambda_0(t) \cdot exp(-0.014 \cdot Age + 0.188 \cdot Meno_1 + 0.373 \cdot Size_{20-50} + 0.641 \cdot Size_{>50} + 0.367 \cdot Grade + 0.078 \cdot Nodes - 0.061 \cdot Treatment_1 - 0.106 \cdot Treatment_2 - 0.441 \cdot Treatment_3)$$

- Interpretation: The risk of death increases 20% in the postmenopausal group($Meno_0$) as compared to premenopausal group($Meno_1$).

Variable	Coef	Exp_coef	Se_coef	P_value
Age	-0.014	0.986	0.003	<0.01
Meno_1	0.188	1.207	0.089	0.03
Size_20-50	0.373	1.452	0.058	<0.01
Size_>50	0.642	1.899	0.088	<0.01
Grade	0.367	1.443	0.064	<0.01
Nodes	0.077	1.080	0.005	<0.01
Treatment_1	-0.062	0.940	0.087	0.4791
Treatment_2	-0.106	0.900	0.072	0.144
Treatment_3	-0.441	0.643	0.280	0.1157

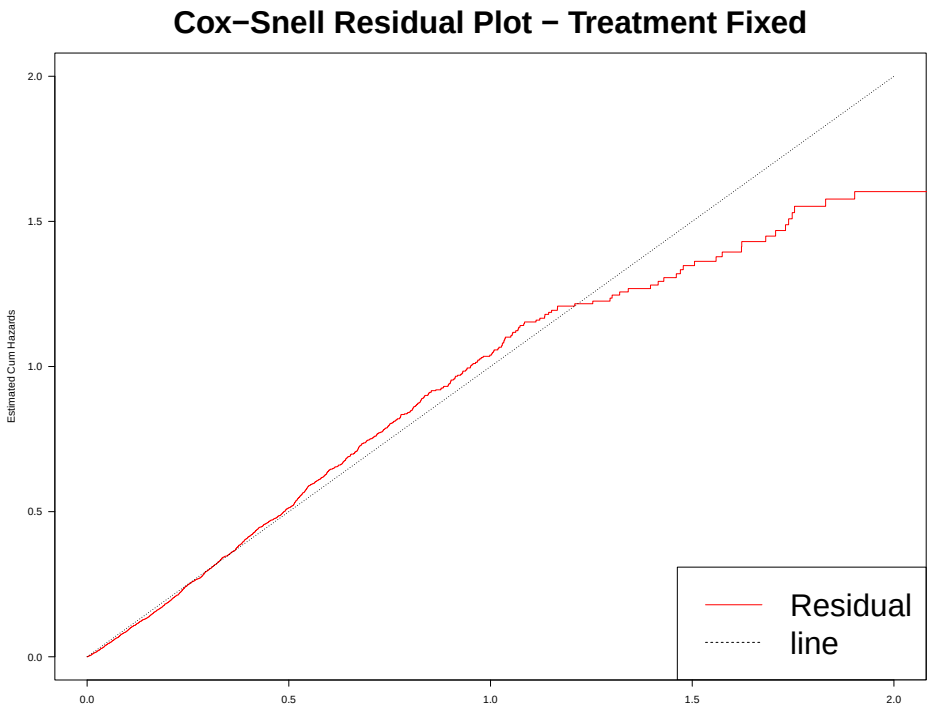
Model Checking

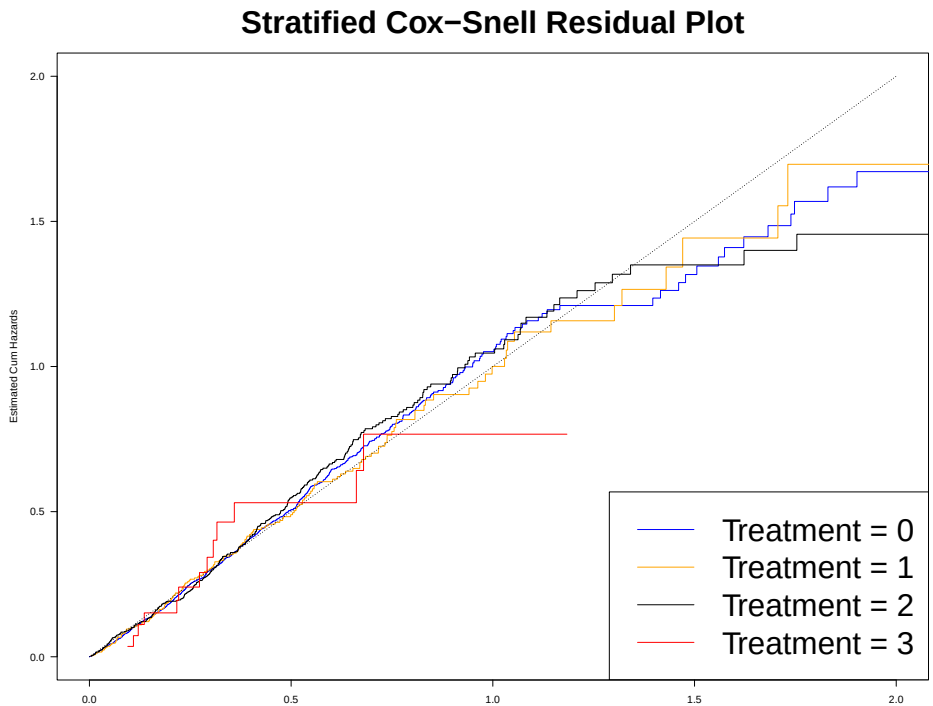
- The LR test of treatment variable:

Variable	Chi_square	P_value	AIC
Treatment	4.8131	0.186	22396

- The AIC of model with treatment is 22396, while the model without treatment is 22390.
- Treatment is not significant.

Cox-Snell Residuals for Overall Fit

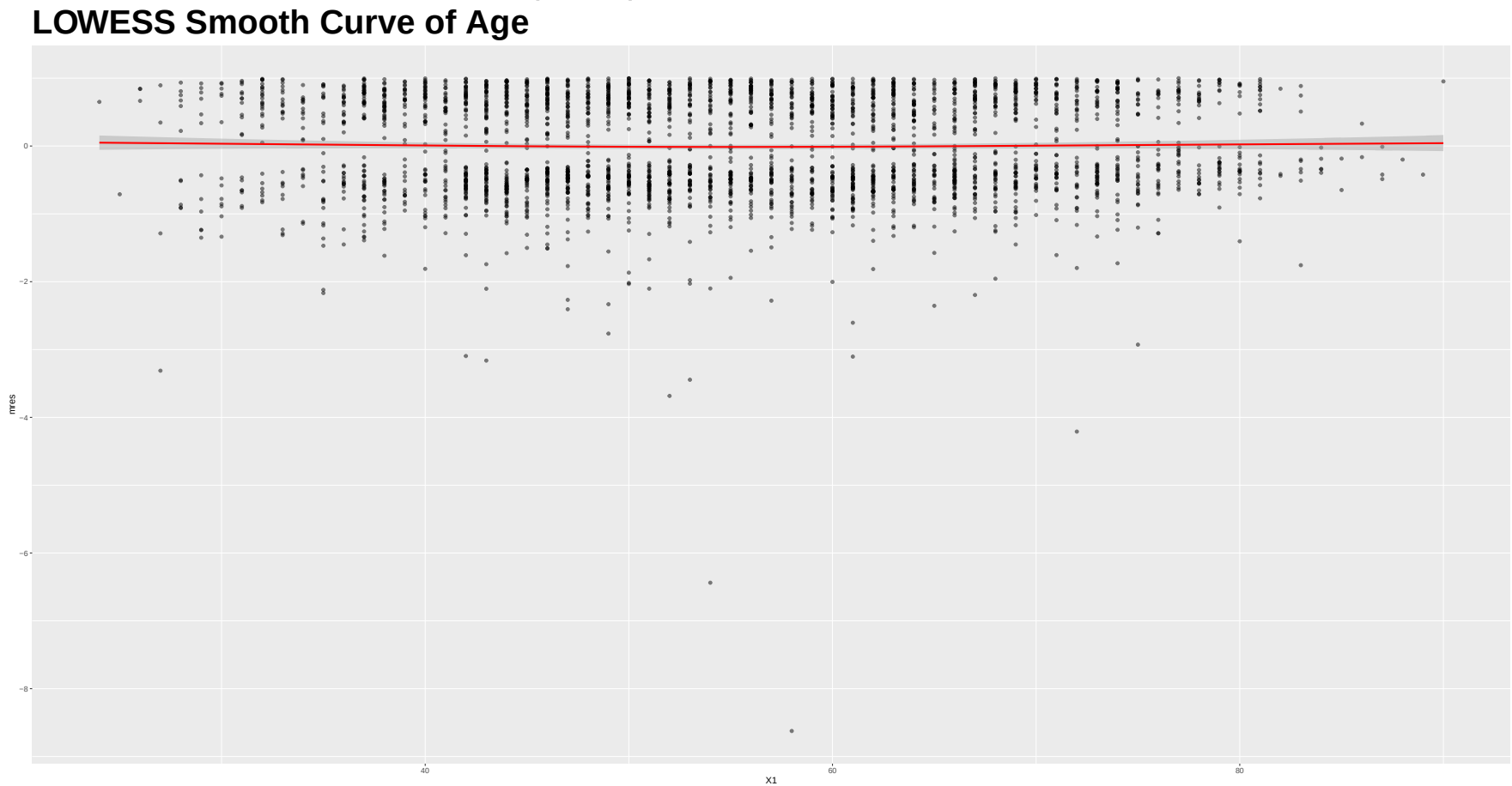




These are the residual plots to check the fit of our final model.The left plot suggests the model does not fit so badly. The right graph plots for 4 different treatment methods groups based on a model straitified on treatment.

Martingel Residual (Check for Age)

- We check the function form of Age. The smooth curve suggests that the function of variable (Age) is linear.



Test of Proportional Assumption

- The Cox model relies on the assumption of proportional hazards (PH) across different covariates.
- Hypothesis of PH:

$$H_0 : The\ model\ meets\ the\ proportional\ hazards\ assumption.$$

$$v.s.$$

$$H_a : The\ model\ does\ not\ meet\ the\ proportional\ hazards\ assumption.$$

Test of Proportional Assumption

- The global test is rejected.

Variable	Chisq	Df	p
Age	0.314	1	0.575
Meno	2.197	1	0.138
Size	27.757	2	<0.01
Grade	4.061	1	0.044
Nodes	6.676	1	0.01
Treatment	7.968	3	0.047
GLOBAL	47.624	9	<0.01

Stratified Cox Model

- Choose the variable Treatment as strata.

The model of stratified cox model for 4 treatment groups:

$$h_i(t, X) = h_{0i}(t)exp(\beta^\top Z), where\ i = 1 \cdots 4$$

The final model of stratified model:

$$h_i(t, X) = h_{0i}(t)exp(-0.014 \cdot Age + 0.19 \cdot Meno_1 + 0.369 \cdot Size_{3\ 20-50} + 0.651 \cdot X_{3\ >50} + 0.368 \cdot X_4 + 0.078 \cdot X_5), where\ i = 1 \cdots 4$$

Variable	Coef	Exp_coef	Se_coef	P_value
Age	-0.014	0.986	0.003	<0.01
Meno_1	0.190	1.210	0.089	<0.01
Size_20-50	0.369	1.446	0.058	<0.01
Size_>50	0.651	1.917	0.088	<0.01
Grade	0.368	1.446	0.064	<0.01
Nodes	0.078	1.080	0.005	<0.01

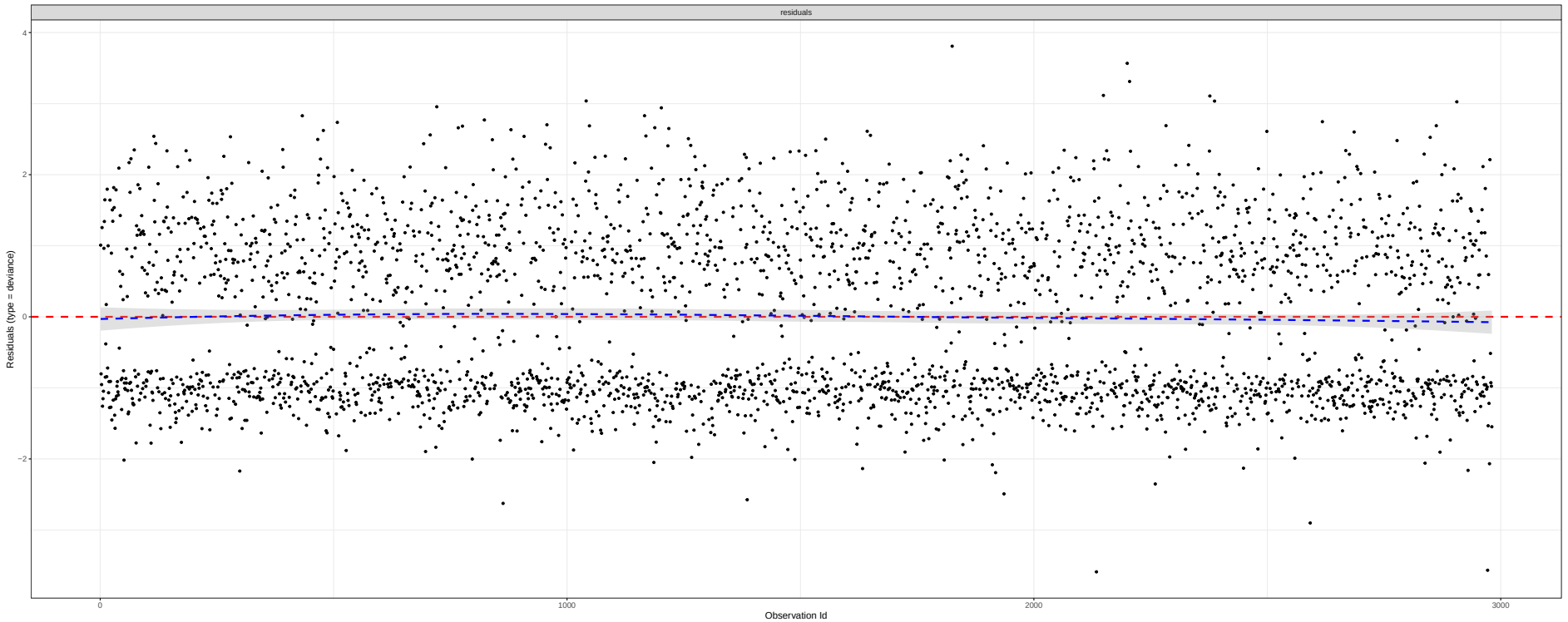
Test of Proportional Assumption

- We decided not to set X_3 as stratification.

Variable	chisq	Df	P_value
Age	0.006	1	0.945
Meno	1.003	1	0.316
Size	32.706	2	<0.01
Grade	4.232	1	0.039
Nodes	11.597	1	<0.01
GLOBAL	40.054	6	<0.01

Deviance Residuals for Outliers

- The deviance residual is a normalized transform of the martingale residual.
- These residuals should be roughly symmetrically distributed about zero with a standard deviation of 1 .



Section 6

Cox-PH Model Results after Adjusting the Data

Discussion

- Royston and Altman suggested to omit the node-negative patients, since nodal status is an important prognostic factor.

Discussion

- Remove negative node patients.
- Final Cox model:

$$\lambda(t|Z_i) = \lambda_0(t) \cdot exp(-0.370 \cdot Treatment_1 - 0.585 \cdot Treatment_2 - 0.932 \cdot Treatment_3 + 0.055 \cdot Nodes + 0.289 \cdot Size_{20-50} + 0.537 \cdot Size_{>50} - 0.012 \cdot Age + 0.299 \cdot Grade - 0.0004 \cdot Er)$$

Discussion

- Final model report:

Variable	coef	exp_coef	se_coef	p_value
Treatment1	-0.3700	0.690	0.0910	<0.01
Treatment2	-0.5850	0.557	0.0920	<0.01
Treatment3	-0.9320	0.394	0.2870	<0.01
Nodes	0.0550	1.057	0.0050	<0.01
Size_20-50	0.2890	1.335	0.0780	<0.01
Size_>50	0.5370	1.711	0.0990	<0.01
Age	-0.0120	0.988	0.0030	<0.01
Grade	0.2990	1.348	0.0810	<0.01
Er	-0.0004	0.999	0.0001	<0.01

- We could dive into the data after right truncation in the future.

Reference

- Mohammed, Shariq. (2019). Introduction to Survival Analysis in R. Retrieved from https://shariq-mohammed.github.io/files/cbsa2019/1-intro-to-survival.html#62_diagnosticsm
- Royston, P., Altman, D.G. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol 13, 33 (2013). https://doi.org/10.1186/1471-2288-13-33
- Shen, H. (W 2023). Functions, Quantities and Models [Lecture Slides]. University of Calgary, MS421.

Reference

- Surakasula A, Nagarjunapu GC, Raghavaiah KV. A comparative study of pre- and post-menopausal breast cancer: Risk factors, presentation, characteristics and management. J Res Pharm Pract. 2014 Jan;3(1):12-8. doi: 10.4103/2279-042X.132704. PMID: 24991630; PMCID: PMC4078652.
- World Health Organization. (2021). Breast cancer: Key facts. Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

Section 7

Apendix

Section 1 R Code

```
library("survival")
attach(rotterdam)
# count how many values in the age variable are under 40
# sum(rotterdam$age < 40)
# count how many values in the age variable are above 50
# sum(rotterdam$age >= 40)
# define new vairable named treatment
rotterdam$treatment = ifelse((rotterdam$hormon==0 & rotterdam$chemo==0), 0,
                             ifelse((rotterdam$hormon==1 & rotterdam$chemo==0), 1,
                                    ifelse((rotterdam$hormon==0 & rotterdam$chemo==1),2,3)))
# table(rotterdam$treatment)
# tail(rotterdam,4)

rotterdam <- rotterdam[rotterdam$nodes > 0, ]

logrank.test <- survdiff(Surv(rtime, recur) ~ treatment, rho=0, rotterdam)
# logrank.test #Reject the null
# plot the KM estimates for survival function in four treatment groups
fit <- survfit(Surv(rtime, recur)~treatment, rotterdam)
plot(fit, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="all")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)
```

Section 1 R Code Cont'd.

```
#########################################################
# age strata (divide age group into "<= 40" and "> 40")
rotterdam$age_strata = ifelse(rotterdam$age<=40, 1, 2)
#########################################################
# treatment
# if hormon=0 and chemo=0, we have treatment=0;
# if hormon=1 and chemo=0, we have treatment=1;
# if hormon=0 and chemo=1, we have treatment=2;
# if hormon=1 and chemo=1, we have treatment=3;
# table(rotterdam$treatment, rotterdam$age_strata) #imbalance
par(mfrow=c(1,3))
fit <- survfit(Surv(rtime, recur) ~ treatment, rotterdam)
plot(fit, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="all")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)

fit.1 <- survfit(Surv(rtime, recur) ~ treatment, subset(rotterdam, age_strata==1))
plot(fit.1, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="Age Below 40")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)
fit.0 <- survfit(Surv(rtime, recur) ~ treatment, subset(rotterdam, age_strata==2))
plot(fit.0, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="Age Above 40")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)
```

Section 1 R Code Cont'd.

```
logrank.test<-survdiff(Surv(rtime, recur) ~ treatment+strata(age_strata), rotterdam)
#print(logrank.test)
table <- data.frame(N = c(655,311,552,28), Observed = c(468,169,324,13), Expected = c(355.5,184.2,409.5,24.8),
rownames(table) = c('Treatment = 0','Treatment = 1','Treatment = 2','Treatment = 3')
colnames(table) = c('N','Observed','Expected','(O-E)^2/E','(O-E)^2/V')
kbl(table) %>%
kable_classic(full_width = F, html_font = "Cambria")
cat("Chisq= 65.8  on 3 degrees of freedom, p= 3e-14")
#########################################################
# menopausal status strata (0= premenopausal, 1= postmenopausal)
#########################################################
# table(rotterdam$treatment, rotterdam$meno) #imbalance
par(mfrow=c(1,3))
fit <- survfit(Surv(rtime, recur) ~ treatment, rotterdam)
plot(fit, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="all")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)
fit.1 <- survfit(Surv(rtime, recur) ~ treatment, subset(rotterdam, meno==1))
plot(fit.1, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="Postmenopausal")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)
fit.0 <- survfit(Surv(rtime, recur) ~ treatment, subset(rotterdam, meno==0))
plot(fit.0, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="Premenopausal")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)
```

Section 1 R Code Cont'd.

```
logrank.test<-survdiff(Surv(rtime, recur) ~ treatment+strata(meno), rotterdam)
# print(logrank.test)
table <- data.frame(N = c(655,311,552,28), Observed = c(468,169,324,13), Expected = c(379.1,200.1,371.4,23.3),
rownames(table) = c('Treatment = 0','Treatment = 1','Treatment = 2','Treatment = 3')
colnames(table) = c('N','Observed','Expected','(O-E)^2/E','(O-E)^2/V')
kbl(table) %>%
kable_classic(full_width = F, html_font = "Cambria")
cat("Chisq= 48.2  on 3 degrees of freedom, p= 2e-10")

#########################################################
# size:  tumor size, a factor with levels <=20 20-50 >50
# We will recode size as 1 for <=20; 2 for 20-50; 3 for >50
#########################################################

rotterdam$size_strata <- ifelse(rotterdam$size == "<=20", 1,
                                ifelse(rotterdam$size == "20-50", 2, 3))

# treatment
# if hormon=0 and chemo=0, we have treatment=0;
# if hormon=1 and chemo=0, we have treatment=1;
# if hormon=0 and chemo=1, we have treatment=2;
# if hormon=1 and chemo=1, we have treatment=3;
# table(rotterdam$treatment, rotterdam$size_strata) #imbalance
```

Section 1 R Code Cont'd.

```
par(mfrow=c(1,4))
fit <- survfit(Surv(rtime, recur) ~ treatment, rotterdam)
plot(fit, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="all")
legend("bottomright", c("treat=0", "treat=1",
                        "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)

fit.1 <- survfit(Surv(rtime, recur) ~ treatment, subset(rotterdam, size_strata==1))
plot(fit.1, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="Size: <=20")
legend("bottomright", c("treat=0", "treat=1",
                        "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)
fit.2 <- survfit(Surv(rtime, recur) ~ treatment, subset(rotterdam, size_strata==2))
plot(fit.1, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="Size: 20-50")
legend("bottomright", c("treat=0", "treat=1",
                        "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)
fit.3 <- survfit(Surv(rtime, recur) ~ treatment, subset(rotterdam, size_strata==3))
plot(fit.1, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="Size: >50")
legend("bottomright", c("treat=0", "treat=1",
                        "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)
```

Section 1 R Code Cont'd.

```
logrank.test<-survdiff(Surv(rtime, recur) ~ treatment+strata(size_strata), rotterdam)
# logrank.test
table <- data.frame(N = c(655,311,552,28), Observed = c(468,169,324,13), Expected = c(373.9,194.1,383.2,22.8),
rownames(table) = c('Treatment = 0','Treatment = 1','Treatment = 2','Treatment = 3')
colnames(table) = c('N','Observed','Expected','(O-E)^2/E','(O-E)^2/V')
kbl(table) %>%
kable_classic(full_width = F, html_font = "Cambria")
cat("Chisq= 40.9  on 3 degrees of freedom, p= 7e-09")

#########################################################
# grade: differentiation grade strata (grade = 2, grade = 3)
#########################################################
# table(rotterdam$treatment, rotterdam$grade) #imbalance
par(mfrow=c(1,3))
fit <- survfit(Surv(rtime, recur) ~ treatment, rotterdam)
plot(fit, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="all")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)

fit.1 <- survfit(Surv(rtime, recur) ~ treatment, subset(rotterdam, grade==2))
plot(fit.1, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="Grade=2")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)
```

Section 1 R Code Cont'd.

```
fit.0 <- survfit(Surv(rtime, recur) ~ treatment, subset(rotterdam, grade==3))
plot(fit.0, xlab="Time (days)", ylab="survival probability",
     col=c("red", "blue", "black", "green"), main="Grade=3")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("red", "blue", "black", "green"), cex=0.8, lty=1)

logrank.test<-survdiff(Surv(rtime, recur) ~ treatment+strata(grade), rotterdam)
# print(logrank.test)
table <- data.frame(N = c(655,311,552,28), Observed = c(468,169,324,13), Expected = c(359.3,195.2,395.7,23.8),
rownames(table) = c('Treatment = 0','Treatment = 1','Treatment = 2','Treatment = 3')
colnames(table) = c('N','Observed','Expected','(O-E)^2/E','(O-E)^2/V')
kbl(table) %>%
kable_classic(full_width = F, html_font = "Cambria")
cat("Chisq= 54.5  on 3 degrees of freedom, p= 9e-12")

fit <- survfit(Surv(rtime, recur) ~ strata(age_strata, meno, size_strata, grade) + treatment,
               data = rotterdam)
plot(fit, col = ifelse(rotterdam$treatment == 1, "blue",
                       ifelse(rotterdam$treatment == 2, "green",
                              ifelse(rotterdam$treatment == 3, "red","black"))),
     lwd=2,
     xlab = "Time (days)", ylab = "Survival Probability",
     main = "Survival Curves by Treatment Group")
legend("topright", c("treat=0", "treat=1",
                     "treat=2", "treat=3"),
       col=c("black", "blue", "green", "red"), cex=0.8, lty=1)
```

Section 2 R Code

```r
#Import dataset rotterdam from survival package.
library(survival)
library(KMsurv)
attach(rotterdam)

#Combine variables "hormon" and "chemo" into one factor variable "treatment".
treatment <- c()
for (i in 1:nrow(rotterdam)){
  if (rotterdam$hormon[i] == 0 & rotterdam$chemo[i] == 0) {
    treatment[i] = 0
  }
  if (rotterdam$hormon[i] == 1 & rotterdam$chemo[i] == 0) {
    treatment[i] <- 1
  }
  if (rotterdam$hormon[i] == 0 & rotterdam$chemo[i] == 1) {
    treatment[i] <- 2
  }
  if (rotterdam$hormon[i] == 1 & rotterdam$chemo[i] == 1) {
    treatment[i] <- 3
  }
}

#plot Kaplan Meier Curves
survfunc <- survfit(Surv(rtime,recur) ~ treatment, data = rotterdam)
survfunc.fit <- survfit(Surv(rtime,recur) ~ treatment, data = rotterdam)

plot(survfunc.fit, lwd = 2, lty = 1, conf.int = F,  mark.time = F, cex = 2, cex.lab = 1.4 , cumhaz = FALSE, xla
legend("topright", c("Treatment = 0", "Treatment = 1", "Treatment = 2", "Treatment = 3"), lty = 1, col = c(1,2,
```

Section 2 R Code Cont'd.

```
#Plot transformations of KM Curves
par(mfrow = c(2,2))
plot(-log(survfunc.fit$surv),survfunc.fit$time, main = "-log(S(t)) vs t", xlab = "time (t)", ylab = "-log(S(t))
plot(log(-log(survfunc.fit$surv)),log(survfunc.fit$time), main = "log(-log(S(t))) vs log(t)" , xlab = "log of t

rotterdam['treatment'] <- as.factor(treatment)
# fit Exponential model
fit.exp <- survreg(Surv(rtime,recur) ~ treatment, dist  = "exponential", data = rotterdam)
# fit Weibull model
fit.web <- survreg(Surv(rtime,recur) ~ treatment, dist = "weibull",data = rotterdam)

# LR test
LR.ts <- (-2)*(fit.exp$loglik[1]-fit.web$loglik[1])
cat("LR Test Statistic =", LR.ts, "\n")
# rejection region and critical value
cat("LR Test Rejection Region and Critical Value  =", qchisq(p=0.05, df=1,lower.tail=FALSE), "\n")
# p-value
cat("LR Test p-value  =", pchisq(q=LR.ts, df=1,lower.tail=FALSE), "\n")

#define covariates
Cov.Age <- rotterdam$age
Cov.Meno <- rotterdam$meno
Cov.Size <- as.factor(rotterdam$size)
Cov.Grade <- as.factor(rotterdam$grade)
Cov.Nodes <- rotterdam$nodes
Cov.PGR <- rotterdam$pgr
Cov.ER <- rotterdam$er
```

Section 2 R Code Cont'd

```
#This local test is based on the LR test;
exp.localtest<-function(web.fit, web.fit1,df){
LR.ts <- (-2)*(web.fit$loglik[2]- web.fit1$loglik[2])
pval <- pchisq(q=LR.ts, df=df,lower.tail=FALSE)
results<-c(df,LR.ts, pval)
list(results)
}

#The following are the models that adjusts for the treatment groups.
expTable1 <- survreg(Surv(rtime,recur) ~ treatment, dist  = "weibull", data = rotterdam)
expTable.Age1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Age, dist  = "weibull", data = rotterdam)
expTable.Meno1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Meno, dist  = "weibull", data = rotterdam)
expTable.Size1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Size, dist  = "weibull", data = rotterdam)
expTable.Grade1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Grade, dist  = "weibull", data = rotterdam)
expTable.Nodes1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes, dist  = "weibull", data = rotterdam)
expTable.PGR1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.PGR, dist  = "weibull", data = rotterdam)
expTable.ER1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.ER, dist  = "weibull", data = rotterdam)
Table1.1<-matrix(0,7,3)
Table1.1[1,]<-c(exp.localtest(expTable1, expTable.Age1.1,1)[[1]])
Table1.1[2,]<-c(exp.localtest(expTable1,expTable.Meno1.1,1)[[1]])
Table1.1[3,]<-c(exp.localtest(expTable1,expTable.Size1.1,2)[[1]])
Table1.1[4,]<-c(exp.localtest(expTable1,expTable.Grade1.1,1)[[1]])
Table1.1[5,]<-c(exp.localtest(expTable1,expTable.Nodes1.1,1)[[1]])
Table1.1[6,]<-c(exp.localtest(expTable1,expTable.PGR1.1,1)[[1]])
Table1.1[7,]<-c(exp.localtest(expTable1,expTable.ER1.1,1)[[1]])
cat("Table 1.1: Local test for possible confounders, adjusted for treatment groups", "\n")
colnames(Table1.1) <- c("df", "LR Test Stat", "p-value")
rownames(Table1.1) <- c("Age", "Meno","Size", "Grade", "Nodes", "PGR", "ER")
Table1.1
```

Section 2 R Code Cont'd

```
#The following are the models that adjusts for the treatment groups and nodes.
expTable2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes, dist  = "weibull", data = rotterdam)
expTable.Age1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Age, dist  = "weibull", data = rotte
expTable.Meno1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Meno, dist  = "weibull", data = rot
expTable.Size1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size, dist  = "weibull", data = rot
expTable.Grade1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Grade, dist  = "weibull", data = r
expTable.PGR1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.PGR, dist  = "weibull", data = rotte
expTable.ER1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.ER, dist  = "weibull", data = rotterd
Table1.2<-matrix(0,6,3)
Table1.2[1,]<-c(exp.localtest(expTable2,expTable.Age1.2 , df=1)[[1]])
Table1.2[2,]<-c(exp.localtest(expTable2,expTable.Meno1.2, df=1)[[1]])
Table1.2[3,]<-c(exp.localtest(expTable2,expTable.Size1.2, df=2)[[1]])
Table1.2[4,]<-c(exp.localtest(expTable2,expTable.Grade1.2, df=1)[[1]])
Table1.2[5,]<-c(exp.localtest(expTable2,expTable.PGR1.2, df=1)[[1]])
Table1.2[6,]<-c(exp.localtest(expTable2,expTable.ER1.2, df=1)[[1]])
cat("Table 1.2: Local test for possible confounders, adjusted for treatment groups and nodes", "\n")
colnames(Table1.2) <- c("df", "LR Test Stat", "p-value")
rownames(Table1.2) <- c("Age", "Meno", "Size", "Grade", "PGR", "ER")
Table1.2
```

Section 2 R Code Cont'd

```
#The following are the models that adjusts for the treatment groups, nodes and size.
expTable3<- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size, dist  = "weibull", data = rotterdam)
expTable.Age1.3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Age, dist  = "weibull", d
expTable.Meno1.3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Meno, dist  = "weibull",
expTable.Grade1.3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes+ Cov.Size + Cov.Grade, dist  = "weibull"
expTable.PGR1.3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.PGR, dist  = "weibull", d
expTable.ER1.3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.ER, dist  = "weibull", dat
Table1.3<-matrix(0,5,3)
Table1.3[1,]<-c(exp.localtest(expTable3,expTable.Age1.3 , df=1)[[1]])
Table1.3[2,]<-c(exp.localtest(expTable3,expTable.Meno1.3, df=1)[[1]])
Table1.3[3,]<-c(exp.localtest(expTable3,expTable.Grade1.3, df=1)[[1]])
Table1.3[4,]<-c(exp.localtest(expTable3,expTable.PGR1.3, df=1)[[1]])
Table1.3[5,]<-c(exp.localtest(expTable3,expTable.ER1.3, df=1)[[1]])
cat("Table 1.3: Local test for possible confounders, adjusted for treatment groups, nodes and size", "\n")
colnames(Table1.3) <- c("df", "Wald's Test Stat", "p-value")
rownames(Table1.3) <- c("Age", "Meno", "Grade", "PGR", "ER")
Table1.3
```

Section 2 R Code Cont'd

```
#The following are the models that adjusts for the treatment groups, nodes, size and grade.
expTable4 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade, dist  = "weibull", data
expTable.Age1.4 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age, dist  =
expTable.Meno1.4 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.Meno, dist
expTable.PGR1.4 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.PGR, dist  =
expTable.ER1.4 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.ER, dist  = "w
Table1.4<-matrix(0,4,3)
Table1.4[1,]<-c(exp.localtest(expTable4,expTable.Age1.4 , df=1)[[1]])
Table1.4[2,]<-c(exp.localtest(expTable4,expTable.Meno1.4, df=1)[[1]])
Table1.4[3,]<-c(exp.localtest(expTable4,expTable.PGR1.4, df=1)[[1]])
Table1.4[4,]<-c(exp.localtest(expTable4,expTable.ER1.4, df=1)[[1]])
cat("Table 1.3: Local test for possible confounders, adjusted for treatment groups, nodes, size and grade", "\n
colnames(Table1.4) <- c("df", "LR Test Stat", "p-value")
rownames(Table1.4) <- c("Age", "Meno", "PGR", "ER")
Table1.4

#The following are the models that adjusts for the treatment groups, nodes, size, grade and age.
expTable5 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade  + Cov.Age, dist  = "weib
expTable.Meno1.5 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade  + Cov.Age + Cov.M
expTable.PGR1.5 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age + Cov.PGR
expTable.ER1.5 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age + Cov.ER,
Table1.5<-matrix(0,3,3)
Table1.5[1,]<-c(exp.localtest(expTable5,expTable.Meno1.5, df=1)[[1]])
Table1.5[2,]<-c(exp.localtest(expTable5,expTable.PGR1.5, df=1)[[1]])
Table1.5[3,]<-c(exp.localtest(expTable5,expTable.ER1.5, df=1)[[1]])
cat("Table 1.3: Local test for possible confounders, adjusted for treatment groups, nodes, size, grade and age"
colnames(Table1.5) <- c("df", "LR Test Stat", "p-value")
rownames(Table1.5) <- c("Meno", "PGR", "ER")
Table1.5
```

Section 2 R Code Cont'd

```
expModel <- survreg(Surv(rtime,recur) ~ Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age, dist  = "weibull", data = r
expModel.treatment <- survreg(Surv(rtime,recur) ~ Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age + treatment, dist
summary(expModel.treatment)
cat('\n',"Local Test LR Test Statistic: ", c(exp.localtest(expModel ,expModel.treatment, df=3)[[1]])[2]
    ,'\n',"Local Test LR p-value :", c(exp.localtest(expModel ,expModel.treatment, df=3)[[1]])[3])
```

Section 3 R Code

```
#add right truncation and consider positive nodes only
rotterdam <- rotterdam[rotterdam$nodes > 0, ] # positive-node only
# fit Exponential model
fit.exp <- survreg(Surv(rtime,recur) ~ treatment, dist  = "exponential", data = rotterdam)
# fit Weibull model
fit.web <- survreg(Surv(rtime,recur) ~ treatment, dist = "weibull",data = rotterdam)

# LR test
LR.ts <- (-2)*(fit.exp$loglik[1]-fit.web$loglik[1])
cat("LR Test Statistic =", LR.ts, "\n")
# rejection region and critical value
cat("LR Test Rejection Region and Critical Value  =", qchisq(p=0.05, df=1,lower.tail=FALSE), "\n")
# p-value
cat("LR Test p-value  =", pchisq(q=LR.ts, df=1,lower.tail=FALSE), "\n")
#reject null and conclude there is sufficient evidence to assume a Weibull model over an exponential model
Cov.Age <- rotterdam$age
Cov.Meno <- rotterdam$meno
Cov.Size <- as.factor(rotterdam$size)
Cov.Grade <- as.factor(rotterdam$grade)
Cov.Nodes <- rotterdam$nodes
Cov.PGR <- rotterdam$pgr
Cov.ER <- rotterdam$er
#This local test is based on the LR test;
exp.localtest<-function(web.fit, web.fit1,df){
LR.ts <- (-2)*(web.fit$loglik[2]- web.fit1$loglik[2])
pval <- pchisq(q=LR.ts, df=df,lower.tail=FALSE)
results<-c(df,LR.ts, pval)
list(results)
}
```

Section 3 R Code Cont'd

```
#The following are the models that adjusts for the treatment groups.
expTable1 <- survreg(Surv(rtime,recur) ~ treatment, dist  = "weibull", data = rotterdam)
expTable.Age1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Age, dist  = "weibull", data = rotterdam)
expTable.Meno1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Meno, dist  = "weibull", data = rotterdam)
expTable.Size1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Size, dist  = "weibull", data = rotterdam)
expTable.Grade1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Grade, dist  = "weibull", data = rotterdam)
expTable.Nodes1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes, dist  = "weibull", data = rotterdam)
expTable.PGR1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.PGR, dist  = "weibull", data = rotterdam)
expTable.ER1.1 <- survreg(Surv(rtime,recur) ~ treatment + Cov.ER, dist  = "weibull", data = rotterdam)
Table1.1<-matrix(0,7,3)
Table1.1[1,]<-c(exp.localtest(expTable1, expTable.Age1.1,1)[[1]])
Table1.1[2,]<-c(exp.localtest(expTable1,expTable.Meno1.1,1)[[1]])
Table1.1[3,]<-c(exp.localtest(expTable1,expTable.Size1.1,2)[[1]])
Table1.1[4,]<-c(exp.localtest(expTable1,expTable.Grade1.1,1)[[1]])
Table1.1[5,]<-c(exp.localtest(expTable1,expTable.Nodes1.1,1)[[1]])
Table1.1[6,]<-c(exp.localtest(expTable1,expTable.PGR1.1,1)[[1]])
Table1.1[7,]<-c(exp.localtest(expTable1,expTable.ER1.1,1)[[1]])
cat("Table 1.1: Local test for possible confounders, adjusted for treatment groups", "\n")
colnames(Table1.1) <- c("df", "LR Test Stat", "p-value")
rownames(Table1.1) <- c("Age", "Meno","Size", "Grade", "Nodes", "PGR", "ER")
Table1.1
```

Section 3 R Code Cont'd

```
#The following are the models that adjusts for the treatment groups and nodes
expTable2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes, dist  = "weibull", data = rotterdam)
expTable.Age1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Age, dist  = "weibull", data = rotte
expTable.Meno1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Meno, dist  = "weibull", data = rot
expTable.Size1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size, dist  = "weibull", data = rot
expTable.Grade1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Grade, dist  = "weibull", data = r
expTable.PGR1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.PGR, dist  = "weibull", data = rotte
expTable.ER1.2 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.ER, dist  = "weibull", data = rotterd
Table1.2<-matrix(0,6,3)
Table1.2[1,]<-c(exp.localtest(expTable2,expTable.Age1.2 , df=1)[[1]])
Table1.2[2,]<-c(exp.localtest(expTable2,expTable.Meno1.2, df=1)[[1]])
Table1.2[3,]<-c(exp.localtest(expTable2,expTable.Size1.2, df=2)[[1]])
Table1.2[4,]<-c(exp.localtest(expTable2,expTable.Grade1.2, df=1)[[1]])
Table1.2[5,]<-c(exp.localtest(expTable2,expTable.PGR1.2, df=1)[[1]])
Table1.2[6,]<-c(exp.localtest(expTable2,expTable.ER1.2, df=1)[[1]])
cat("Table 1.2: Local test for possible confounders, adjusted for treatment groups and nodes", "\n")
colnames(Table1.2) <- c("df", "LR Test Stat", "p-value")
rownames(Table1.2) <- c("Age", "Meno", "Size", "Grade", "PGR", "ER")
Table1.2
```

Section 3 R Code Cont'd

```
#The following are the models that adjusts for the treatment groups, nodes and size
expTable3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size, dist  = "weibull", data = rotterdam)
expTable.Age1.3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Age, dist  = "weibull", d
expTable.Meno1.3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Meno, dist  = "weibull",
expTable.Grade1.3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes+ Cov.Size + Cov.Grade, dist  = "weibull"
expTable.PGR1.3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.PGR, dist  = "weibull", d
expTable.ER1.3 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.ER, dist  = "weibull", dat
Table1.3<-matrix(0,5,3)
Table1.3[1,]<-c(exp.localtest(expTable3,expTable.Age1.3 , df=1)[[1]])
Table1.3[2,]<-c(exp.localtest(expTable3,expTable.Meno1.3, df=1)[[1]])
Table1.3[3,]<-c(exp.localtest(expTable3,expTable.Grade1.3, df=1)[[1]])
Table1.3[4,]<-c(exp.localtest(expTable3,expTable.PGR1.3, df=1)[[1]])
Table1.3[5,]<-c(exp.localtest(expTable3,expTable.ER1.3, df=1)[[1]])
cat("Table 1.3: Local test for possible confounders, adjusted for treatment groups, nodes and size", "\n")
colnames(Table1.3) <- c("df", "Wald's Test Stat", "p-value")
rownames(Table1.3) <- c("Age", "Meno", "Grade", "PGR", "ER")
Table1.3
```

Section 3 R Code Cont'd

```
#The following are the models that adjusts for the treatment groups, nodes, size and age
expTable4 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Age, dist  = "weibull", data =
expTable.Grade1.4 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Age + Cov.Grade, dist
expTable.Meno1.4 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.Meno, dist
expTable.PGR1.4 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.PGR, dist  =
expTable.ER1.4 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.ER, dist  = "w
Table1.4<-matrix(0,4,3)
Table1.4[1,]<-c(exp.localtest(expTable4,expTable.Grade1.4 , df=1)[[1]])
Table1.4[2,]<-c(exp.localtest(expTable4,expTable.Meno1.4, df=1)[[1]])
Table1.4[3,]<-c(exp.localtest(expTable4,expTable.PGR1.4, df=1)[[1]])
Table1.4[4,]<-c(exp.localtest(expTable4,expTable.ER1.4, df=1)[[1]])
cat("Table 1.3: Local test for possible confounders, adjusted for treatment groups, nodes and size", "\n")
colnames(Table1.4) <- c("df", "LR Test Stat", "p-value")
rownames(Table1.4) <- c("Grade", "Meno", "PGR", "ER")
Table1.4
```

Section 3 R Code Cont'd

```
#The following are the models that adjusts for the treatment groups, nodes, size, age and grade
expTable5 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade  + Cov.Age, dist  = "weib
expTable.Meno1.5 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade  + Cov.Age + Cov.M
expTable.PGR1.5 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age + Cov.PGR
expTable.ER1.5 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age + Cov.ER,
Table1.5<-matrix(0,3,3)
Table1.5[1,]<-c(exp.localtest(expTable5,expTable.Meno1.5, df=1)[[1]])
Table1.5[2,]<-c(exp.localtest(expTable5,expTable.PGR1.5, df=1)[[1]])
Table1.5[3,]<-c(exp.localtest(expTable5,expTable.ER1.5, df=1)[[1]])
cat("Table 1.%: Local test for possible confounders, adjusted for treatment groups, nodes, size, grade and age"
colnames(Table1.5) <- c("df", "LR Test Stat", "p-value")
rownames(Table1.5) <- c("Meno", "PGR", "ER")
Table1.5
#The following are the models that adjusts for the treatment groups, nodes, size, age, grade and ER
expTable6 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade  + Cov.Age + Cov.ER, dist
expTable.Meno1.6 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade  + Cov.Age + Cov.E
expTable.PGR1.6 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age + Cov.ER
Table1.6<-matrix(0,2,3)
Table1.6[1,]<-c(exp.localtest(expTable6,expTable.Meno1.6, df=1)[[1]])
Table1.6[2,]<-c(exp.localtest(expTable6,expTable.PGR1.6, df=1)[[1]])
cat("Table 1.6: Local test for possible confounders, adjusted for treatment groups, nodes, size, grade, age and
colnames(Table1.6) <- c("df", "LR Test Stat", "p-value")
rownames(Table1.6) <- c("Meno", "PGR")
Table1.6
```

Section 3 R Code Cont'd

```
#The following are the models that adjusts for the treatment groups, nodes, size, age, grade, ER and PGR
expTable7 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade  + Cov.Age + Cov.ER+ Cov.
expTable.Meno1.7 <- survreg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade  + Cov.Age + Cov.E
Table1.7<-matrix(0,1,3)
Table1.7[1,]<-c(exp.localtest(expTable7,expTable.Meno1.7, df=1)[[1]])
cat("Table 1.7: Local test for possible confounders, adjusted for treatment groups, nodes, size, grade, age, ER
colnames(Table1.7) <- c("df", "LR Test Stat", "p-value")
rownames(Table1.7) <- c("Meno")
Table1.7
expModel <- survreg(Surv(rtime,recur) ~ Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age +Cov.ER+ Cov.PGR, dist  = "w
expModel.treatment <- survreg(Surv(rtime,recur) ~ Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age + Cov.ER+ Cov.PGR
summary(expModel.treatment)
cat('\n',"Local Test LR Test Statistic: ", c(exp.localtest(expModel ,expModel.treatment, df=3)[[1]])[2]
    ,'\n',"Local Test LR p-value :", c(exp.localtest(expModel ,expModel.treatment, df=3)[[1]])[3])
 #install.packages("SurvRegCensCov")
library(SurvRegCensCov)
weibullmodel <- WeibullReg(Surv(rtime,recur) ~ treatment + Cov.Nodes + Cov.Size + Cov.Grade + Cov.Age +Cov.ER+
weibullmodel
```

Section 4 R Code

```
library("survival")
library(MASS)
library(formatR)
data(cancer)
attach(rotterdam)
rotterdam$treatment <- ifelse(rotterdam$hormon==0 & rotterdam$chemo==0,0,
                      ifelse(rotterdam$hormon==1 & rotterdam$chemo==0,1,
                      ifelse(rotterdam$hormon==0 & rotterdam$chemo==1,2,3)))
X1 <- rotterdam$age
X2 <- rotterdam$meno
X3 <- rotterdam$size
X4 <- rotterdam$grade
X5 <- rotterdam$nodes
X6 <- rotterdam$pgr
X7 <- rotterdam$er
X8 <- as.factor(rotterdam$treatment)
rtime <- rotterdam$rtime
recur <- rotterdam$recur

rot_coxnull <- coxph(Surv(rtime,recur)~X8,method=c("breslow"))
rot_coxfor<- stepAIC(rot_coxnull, ~., ditrection="foreward",scope=list(lower=rot_coxnull, upper=~X1+X2+X3+X4+X5

rot_final <- coxph(Surv(rtime,recur)~X1+X2+X3+X4+X5+strata(X8),method=c("breslow"),data = rotterdam)
mres = resid(rot_final, type="martingale")
csres = recur-mres
r.surv_1 = survfit(Surv(csres,recur)~1,type="fleming-harrington")
par(las=1, mfrow=c(1,2), mai = c(0.5,1.0,1,0.1), omi = c(1.0,0,0.5,0))
plot(0,0,lty=1,type='n',xlim = c(0,2),ylim = c(0,2),xlab="Residual",
     ylab="Estimated Cum Hazards", main="Cox-Snell Residual Plot - Treatment Fixed",cex.main = 3)
```

Section 4 R Code Cont'd

```
lines(r.surv_1$time, -log(r.surv_1$surv), type='s',col = "red",lty =1)
lines(c(0,2),c(0,2),lty = 3)
legend("bottomright", legend=c("Residual", "line"),
       col = c("red","black"),lty=c(1,2),cex=3)
rot_stratified <- coxph(Surv(rtime,recur)~X1+X2+X3+X4+X5+strata(X8),method=c("breslow"),data = rotterdam)
X8_0 <- (X8 == 0)
X8_1 <- (X8 == 1)
X8_2 <- (X8 == 2)
X8_3 <- (X8 == 3)
mres2 = resid(rot_stratified, type="martingale")
csres2 = recur-mres2
r.surv20 = survfit(Surv(csres2[X8_0],recur[X8_0])~1,type="fleming-harrington")
r.surv21 = survfit(Surv(csres2[X8_1],recur[X8_1])~1,type="fleming-harrington")
r.surv22 = survfit(Surv(csres2[X8_2],recur[X8_2])~1,type="fleming-harrington")
r.surv23 = survfit(Surv(csres2[X8_3],recur[X8_3])~1,type="fleming-harrington")

plot(0,0,lty=1,type='n',xlim = c(0,2),ylim = c(0,2),xlab="Residual",
     ylab="Estimated Cum Hazards", main="Stratified Cox-Snell Residual Plot",cex.main = 3)
lines(r.surv20$time, -log(r.surv20$surv), type='s', lty=1,col = "blue")
lines(r.surv21$time, -log(r.surv21$surv), type='s', lty=1, col = "orange")
lines(r.surv22$time, -log(r.surv22$surv), type='s', lty=1, col = "black")
lines(r.surv23$time, -log(r.surv23$surv), type='s', lty=1, col = "red")
lines(c(0,2),c(0,2),lty = 3)
legend("bottomright",legend=c("Treatment = 0","Treatment = 1","Treatment = 2","Treatment = 3"),
       lty=c(1,1,1,1),col=c("blue","orange","black","red"),cex = 3)
```

Section 4 R Code

```
rot_stratified <- coxph(Surv(rtime,recur)~X1+X2+X3+X4+X5+strata(X8),method=c("breslow"),data = rotterdam)
mres = resid(rot_stratified, type="martingale")
library(ggplot2)
resid <- as.data.frame(cbind(X1,mres))
ggplot(aes(x=X1,y=mres),data=resid)+geom_point(alpha=0.5)+geom_smooth(col="red")+ggtitle("LOWESS Smooth Curve o
library(survminer)
library(ggpubr)
library(ggplot2)
try <- rotterdam[sample(nrow(rotterdam),2981,replace=F),]
tX1 <- try$age
tX2 <- as.factor(try$meno)
tX3 <- try$size
tX4 <- try$grade
tX5 <- try$nodes
tX6 <- try$pgr
tX7 <- try$er
tX8 <- as.factor(try$treatment)
rot_stratified_try <- coxph(Surv(rtime,recur)~tX1+tX2+tX3+tX4+tX5+strata(tX8),method=c("breslow"),data = try)
ggcoxdiagnostics(rot_stratified_try, type = "deviance",
                 linear.predictions = FALSE, gather = theme_bw())
```

Section 5 R Code

```
# remove negative node
data(cancer)
attach(rotterdam)

rotterdam <- rotterdam[rotterdam$nodes > 0, ] # positive-node only

rotterdam$treatment <- ifelse(rotterdam$hormon==0 & rotterdam$chemo==0,0,
                              ifelse(rotterdam$hormon==1 & rotterdam$chemo==0,1,
                                     ifelse(rotterdam$hormon==0 & rotterdam$chemo==1,2,3)))
Age <- rotterdam$age
Meno <- as.factor(rotterdam$meno)
Size <- rotterdam$size
Grade <- rotterdam$grade
Nodes <- rotterdam$nodes
Pgr <- rotterdam$pgr
Er <- rotterdam$er
Treatment <- as.factor(rotterdam$treatment)
rtime <- rotterdam$rtime
recur <- rotterdam$recur
rot_coxnull.trun <- coxph(Surv(rtime,recur)~Treatment,method=c("breslow"))
# forward selection
rot_coxfor.trun<- stepAIC(rot_coxnull.trun, ~., ditrection="forward",scope=list(lower=rot_coxnull, upper=~Age+M
rot_final.trun <- coxph(Surv(rtime,recur)~Treatment + Nodes + Size + Age + Grade +
                          Er + Meno)
print(summary(rot_final.trun))
# meno is not significant, so remove it
rot_final.trun1 <- coxph(Surv(rtime,recur)~Treatment + Nodes + Size + Age + Grade +
                          Er )
print(summary(rot_final.trun1))
```