Final_1

Wenjuan Bian

2023-04-18

A clinical trial of chemotherapy against chemotherapy combined with radio-therapy in the treatment of locally unresectable gastric cancer was conducted by the Gastrointestinal Tumor Study Group in 1982. In this trial, forty-five patients were randomized to each of the two arms and followed for about eight years. The data (in days) are as follows (the right-censored times are indicated by an asterisk). The clinical trial researchers wished to see whether that the survival rate of the two groups is the same.

Chemotherapy Only											
1	63	105	129	182	216	250	262	301	301	342	354
356	358	380	383	383	388	394	408	460	489	499	523
524	535	562	569	675	676	748	778	786	797	955	968
1000	1245	1271	1420	1551	1694	2363	2754*	2950*			
Chemotherapy Plus Radiotherapy											
17	42	44	48	60	72	74	95	103	108	122	144
167	170	183	185	193	195	197	208	234	235	254	307
315	401	445	464	484	528	542	547	577	580	795	855
1366	1577	2060	2412*	2486*	2796*	2802*	2934*	2988*			

There are at three models that can be consid- ered for the research aim: nonparametric model (Chapter 4), proportional hazard/Cox model (Chapter 5), parametric model (Chapter 10). Test the hypothesis at level _ = 5% that the survival rate of the two groups is the same, in each of the three models. For each approach, you should outline the model and the method/algorithm employed to carry out the test and state your conclusion.

1. (30 points) Test in a nonparametric model.

For the nonparametric model, we can perform a log-rank test to see whether the two groups are different.

Let $S_0(t)$ be the survival rate of the group treated by Chemotherapy only, and $S_1(t)$ be the survival rate of the group treated by Chemotherapy plus Radiotherapy.

H0: Survival rate in the two groups is same, $S_1(t) = S_0(t)$

Ha: Survival rate in the two groups is different, $S_1(t) > S_0(t)$

The test statistic is computed in the following:

$$e_{0\mathrm{i}} = E(d_{0i}) = \frac{n_{0i} d_i}{n_i}$$

Where $e_{0\mathrm{i}}$ is the expected survival value for d_{0i}, and d_{0i} is the observations in the group of Chemo at time i, and d_i is the total observations in the two group and n_i is the total number at risk at time i

$$v_{0\mathrm{i}} = var(d_{0i}) = \frac{n_{0i}n_{1i}d_i(n_i - d_i)}{{n_i}^2(n_i - 1)}$$

Where $v_{0\mathrm{i}}$ is the variance of d_{0i}, $n_{1i}$is the
Therefore,

$$\frac{\sum_{i=1}^{N}(d_{0i} - e_{0\mathrm{i}})^{\,2}}{\sum_{i=1}^{N} v_{0\mathrm{i}}} \sim \chi_1^2$$

For example, at day 1:

d_{01}=1	d_{11}=0
n_{01}=45	n_{11}=45

$$e_{01} = \frac{n_{01}d_1}{n_1} = \frac{45 * 1}{90} = 0.5$$

$$v_{01} = \frac{45 * 45 * 1 * (90 - 1)}{90^2(90 - 1)} = 0.25$$

```
###
# 1. (30 points) Test in a nonparametric model.
####

library(survival)
library(asaur)
library(MASS)
library(fitdistrplus)

## Warning: package 'fitdistrplus' was built under R version 4.2.3

library(rms)

## Warning: package 'rms' was built under R version 4.2.3

## Loading required package: Hmisc

## Warning: package 'Hmisc' was built under R version 4.2.3

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
finaldata <- data.frame(tt = c(1, 63, 105, 129, 182, 216, 250, 262, 301, 301, 342,
354, 356, 358, 380, 383, 383, 388, 394, 408, 460, 489, 499, 523, 524, 535, 562, 569
, 675, 676, 748, 778, 786, 797, 955, 968, 1000, 1245, 1271, 1420, 1551, 1694, 2363,
 2754, 2950,17, 42, 44, 48, 60, 72, 74, 95, 103, 108, 122, 144, 167, 170, 183, 185,
 193, 195, 197, 208, 234, 235, 254, 307, 315, 401, 445, 464, 484, 528, 542, 547, 57
7, 580, 795, 855, 1366, 1577, 2060, 2412, 2486, 2796, 2802, 2934, 2988), delta = c(
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1,
 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1,1,0, 0,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 0, 0, 0
, 0, 0, 0), trt = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1,
 1, 1, 1,1, 1, 1, 1, 1, 1,1,1, 1))
finaldata$trt <- ifelse(finaldata$trt == 0, "chemo", "chemo_radio")
head(finaldata)
```

```
##    tt delta   trt
## 1   1     1 chemo
## 2  63     1 chemo
## 3 105     1 chemo
## 4 129     1 chemo
## 5 182     1 chemo
## 6 216     1 chemo
```

```r
tail(finaldata)
```

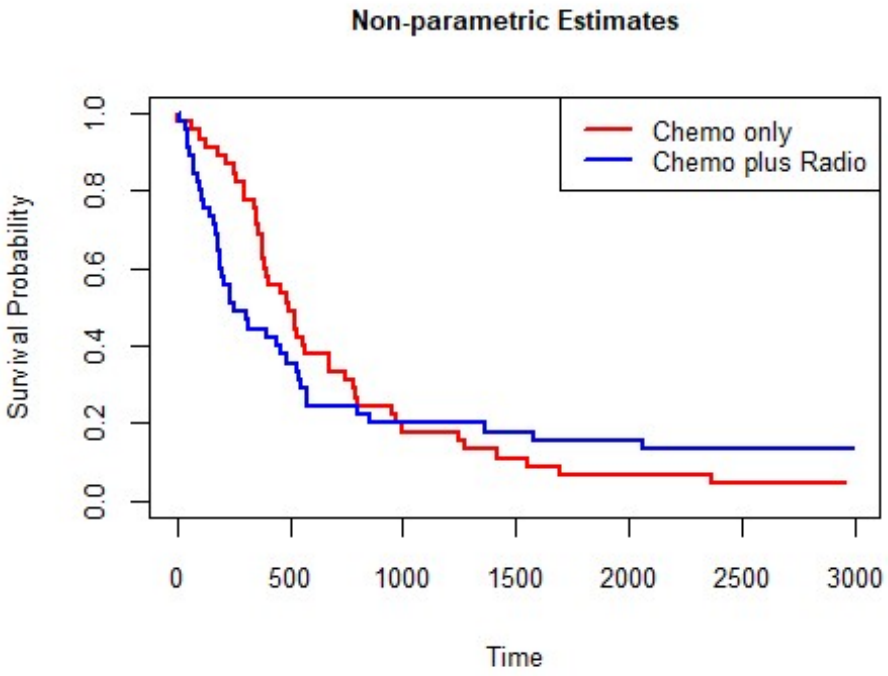
```
##      tt delta         trt
## 85 2412     0 chemo_radio
## 86 2486     0 chemo_radio
## 87 2796     0 chemo_radio
## 88 2802     0 chemo_radio
## 89 2934     0 chemo_radio
## 90 2988     0 chemo_radio
```

```r
result.km <- survfit(Surv(tt, delta)~trt,data=finaldata)

print(result.km)
```

```
## Call: survfit(formula = Surv(tt, delta) ~ trt, data = finaldata)
##
##                   n events median 0.95LCL 0.95UCL
## trt=chemo        45     43    499     383     748
## trt=chemo_radio 45     39    254     193     542
```

```r
plot(result.km, main = "Non-parametric Estimate", xlab = "Time", ylab = "Survival P
robability", col=c("red", "blue"),  lwd=2, cex.axis=0.9, cex.lab=0.9)
legend("topright",
legend=c("Chemo only", "Chemo and Radio"),lty=c(1, 1), col=c("red", "blue"), lwd=2,
 cex=1.0)
```




```
# Perform log-rank test

survdiff(Surv(tt, delta) ~ trt, data = finaldata)

## Call:
## survdiff(formula = Surv(tt, delta) ~ trt, data = finaldata)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=chemo        45       43     45.1     0.102     0.232
## trt=chemo_radio 45       39     36.9     0.125     0.232
##
##  Chisq= 0.2  on 1 degrees of freedom, p= 0.6
```

With p-value =0.6, we fail to reject the null. This means that there is no sufficient evidence to conclude that the two different treatments have made any difference.

2. (30 points) Test in a proportional hazard model.

The proportional hazard model assumes that there is an unspecified baseline survival distribution. The observed hazard function for subject i at time tj is:

$$h_i(t_j) = h_0(t_j)e^{z_i\beta}$$

Where $h_0(t_j)$ is the baseline hazard that is a function of time t.

This model assumes that the hazard proportion between a treatment group and a control group is:

$$\frac{h_i(t_j)}{h_0(t_j)} = constant$$

Which measures the relative risk between groups.

Under this model, for the given data, the hazard ratio between the Chemo group and the Chemo plus Radio group is:

$$\frac{h_i(t_j|trt=1)}{h_i(t_j|trt=0)} = e^{z\beta} = constant$$

The corresponding probability is:

$$p_{ij} = \frac{h_i(t_j)}{\sum h_i(t_j)} = \frac{e^{z_i\beta}}{\sum_i e^{z_i\beta}}$$

The partial likelihood test is:
H_0: Hazard rate in the two groups is same, $\beta = 0$
Ha: Hazard rate in the two groups is different, $\beta \neq 0$
The test statistic is computed in the following:

$$Z_w = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim N(0,1)$$

Or,

$$Z_w{}^2 \sim \chi_1^2$$

```
###
# 2. (30 points) Test in a proportional hazard model.
###

cox_fit <- coxph(Surv(tt, delta) ~ trt, data = finaldata)
summary(cox_fit)

## Call:
## coxph(formula = Surv(tt, delta) ~ trt, data = finaldata)
##
##   n= 90, number of events= 82
##
##                   coef exp(coef) se(coef)     z Pr(>|z|)
## trtchemo_radio 0.1067    1.1126   0.2234 0.478    0.633
##
##                exp(coef) exp(-coef) lower .95 upper .95
## trtchemo_radio     1.113     0.8988    0.7182     1.724
##
## Concordance= 0.562  (se = 0.031 )
## Likelihood ratio test= 0.23  on 1 df,   p=0.6
## Wald test            = 0.23  on 1 df,   p=0.6
## Score (logrank) test = 0.23  on 1 df,   p=0.6
```

The p-value is 0.6 from three tests. We draw the same conclusion as part (a) that there is no sufficient evidence to conclude that the two different treatments have made any difference.

However, for hazard ratio model, the important assumption is that the proportion is constant. From the plot in part (a), we can see the survival curve cross each other at about t=1000. This suggests that the proportion assumption is violated. To check whether the proportion assumption is met, we can perform a Schoenfeld test with command cox.zph in R:

H_0: the hazards are proportional.

Ha: the hazards are not proportional

```
cox.zph(cox_fit)
```

```
##        chisq df       p
## trt     13.2  1 0.00028
## GLOBAL  13.2  1 0.00028
```

This p-value is 0.00028, very small, which suggests that there is sufficient evidence to conclude that the hazard ratio between Chemo group and Chemo plus Radio group are not constant, changing over time.

By fix this violation, one suggestion is using time-depending coefficient model, allowing the variable to interact with time. So I add an trt*tt time interaction covariate to the cox model:

```
cox_fit2 <- coxph(Surv(tt, delta) ~ trt+trt*tt, data = finaldata)
```

```
## Warning in coxph(Surv(tt, delta) ~ trt + trt * tt, data = finaldata): a
## variable appears on both the left and right sides of the formula
```

```
## Warning in coxph.fit(X, Y, istrat, offset, init, control, weights = weights, :
## Ran out of iterations and did not converge
```

```
summary(cox_fit2)
```

```
## Call:
## coxph(formula = Surv(tt, delta) ~ trt + trt * tt, data = finaldata)
##
##   n= 90, number of events= 82
##
##                       coef  exp(coef)   se(coef)      z Pr(>|z|)
## trtchemo_radio    -0.4746309  0.6221147  0.7092301 -0.669    0.503
## tt                -0.4396889  0.6442368  0.0992376 -4.431 9.39e-06 ***
## trtchemo_radio:tt  0.0007444  1.0007447  0.0017724  0.420    0.674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## trtchemo_radio       0.6221     1.6074    0.1549    2.4979
## tt                   0.6442     1.5522    0.5304    0.7826
## trtchemo_radio:tt    1.0007     0.9993    0.9973    1.0042
##
## Concordance= 1  (se = 0 )
## Likelihood ratio test= 594.5  on 3 df,   p=<2e-16
```

```
## Wald test            = 20.26  on 3 df,   p=2e-04
## Score (logrank) test = 92.18  on 3 df,   p=<2e-16
```

By the above output, we can see that time interaction term yields a coefficient of -0.44 with p-value close to 0 for time, which suggests that the hazard ratio is changing over time and the risk in the Chemo plus Radio group is relatively decreasing, compared with the group with treatment of Chemo only.

Perform a Schoenfeld test for the model with trt*time interaction:

```
cox.zph(cox_fit2)
```

```
##        chisq df    p
## trt     1.01  1 0.32
## tt      1.24  1 0.27
## trt:tt  1.21  1 0.27
## GLOBAL  2.50  3 0.48
```

This p-value is much larger than 0.05, which suggests that there is sufficient evidence to conclude that the hazard ratio between Chemo group and Chemo plus Radio group are proportional by adding the time interaction covariate.

Also, I checked the change of goodness fit by AIC for the two models:

```
aic_cox <- AIC(cox_fit)
aic_cox2 <- AIC(cox_fit2)
aic_cox
```

```
## [1] 616.8683
```

```
aic_cox2
```

```
## [1] 26.59404
```

By adding the time interaction covariate, the value of AIC changes dramatically from 616.8683 to 26.59404, suggesting a much better fit by the model with trt*time interaction covariate.

3. (30 points) Test in a parametric model.

For comparing treatment effects, we need an accelerated failure time Weibull model. This model assumes that the survival time for a treated patient is a multiple e^{γ} of what the survival time would have been had the patient received the control treatment.

For Weibull distribution, the hazard function is:

$$h_1(t) = e^{-\gamma} h_0(e^{-\gamma} t) = e^{-\frac{\gamma}{\delta}} * \frac{1}{\sigma} e^{-\frac{\mu_0}{\sigma}} t^{\frac{1}{\sigma}-1}$$

$e^{-\frac{\gamma}{\delta}}$ is a proportional factor equivalent to e^{β}.

The survival function is:

$$S_1(t) = e^{-e^{-\frac{\gamma}{\delta}} * \frac{1}{\sigma} e^{-\frac{\mu_0}{\sigma}} t^{\frac{1}{\sigma}}}$$

Take log(-log($S_1(t)$):

$$\log\left(-\log\left(S_1(t)\right)\right) = -\frac{\gamma}{\sigma} + \frac{1}{\sigma} - \frac{\mu_0}{\sigma} + \left(\frac{1}{\sigma}\right) logt$$

Plot$y_i = \log\left(-\log\left(S_1(t)\right)\right)$ verse $logt_i$ and then fit through these points a straight line to estimate the parameter.

Besides Weibull, I also fitted the data with log normal and log logistic models.

```
###
# 3. (30 points) Test in a parametric model.
###


fit_lognormal <- survreg(Surv(tt, delta) ~ trt, dist="lognormal",data=finaldata)
summary(fit_lognormal)

##
## Call:
## survreg(formula = Surv(tt, delta) ~ trt, data = finaldata, dist = "lognormal")
##                   Value Std. Error     z       p
## (Intercept)      6.1681     0.2040 30.24 < 2e-16
## trtchemo_radio -0.3237     0.2894 -1.12 0.26334
## Log(scale)      0.3097     0.0801  3.87 0.00011
##
## Scale= 1.36
##
## Log Normal distribution
## Loglik(model)= -624   Loglik(intercept only)= -624.6
##  Chisq= 1.24 on 1 degrees of freedom, p= 0.27
## Number of Newton-Raphson Iterations: 3
## n= 90

fit_weibull <- survreg(Surv(tt, delta) ~ trt, data=finaldata, dist="weibull")
summary(fit_weibull)

##
## Call:
## survreg(formula = Surv(tt, delta) ~ trt, data = finaldata, dist = "weibull")
##                  Value Std. Error     z      p
## (Intercept)    6.5811     0.1799 36.59 <2e-16
## trtchemo_radio 0.0351     0.2596  0.14  0.892
## Log(scale)     0.1582     0.0866  1.83  0.068
##
## Scale= 1.17
##
## Weibull distribution
## Loglik(model)= -625.9   Loglik(intercept only)= -626
##  Chisq= 0.02 on 1 degrees of freedom, p= 0.89
## Number of Newton-Raphson Iterations: 5
## n= 90
```

```
fit_loglogistic <- survreg(Surv(tt, delta) ~ trt,data=finaldata, dist="loglogistic"
)
summary(fit_loglogistic)
```

```
##
## Call:
## survreg(formula = Surv(tt, delta) ~ trt, data = finaldata, dist = "loglogistic")
##                   Value Std. Error     z      p
## (Intercept)      6.2452     0.1688 36.99 <2e-16
## trtchemo_radio -0.5046     0.2541 -1.99 0.0471
## Log(scale)     -0.3496     0.0939 -3.72 0.0002
##
## Scale= 0.705
##
## Log logistic distribution
## Loglik(model)= -618   Loglik(intercept only)= -619.9
##  Chisq= 3.79 on 1 degrees of freedom, p= 0.051
## Number of Newton-Raphson Iterations: 3
## n= 90
```

The p-value by log normal model is 0.26 and by Weibull model is 0.89, both suggest that there is no significant difference between the Chemo only group and Chemo plus Radio group. The log logistic model yields a marginal p-value of 0.051, suggesting a potential effect of Chemo plus Radio treatment compared with Chemo treatment only. However, we need to check the goodness of fit for these models.

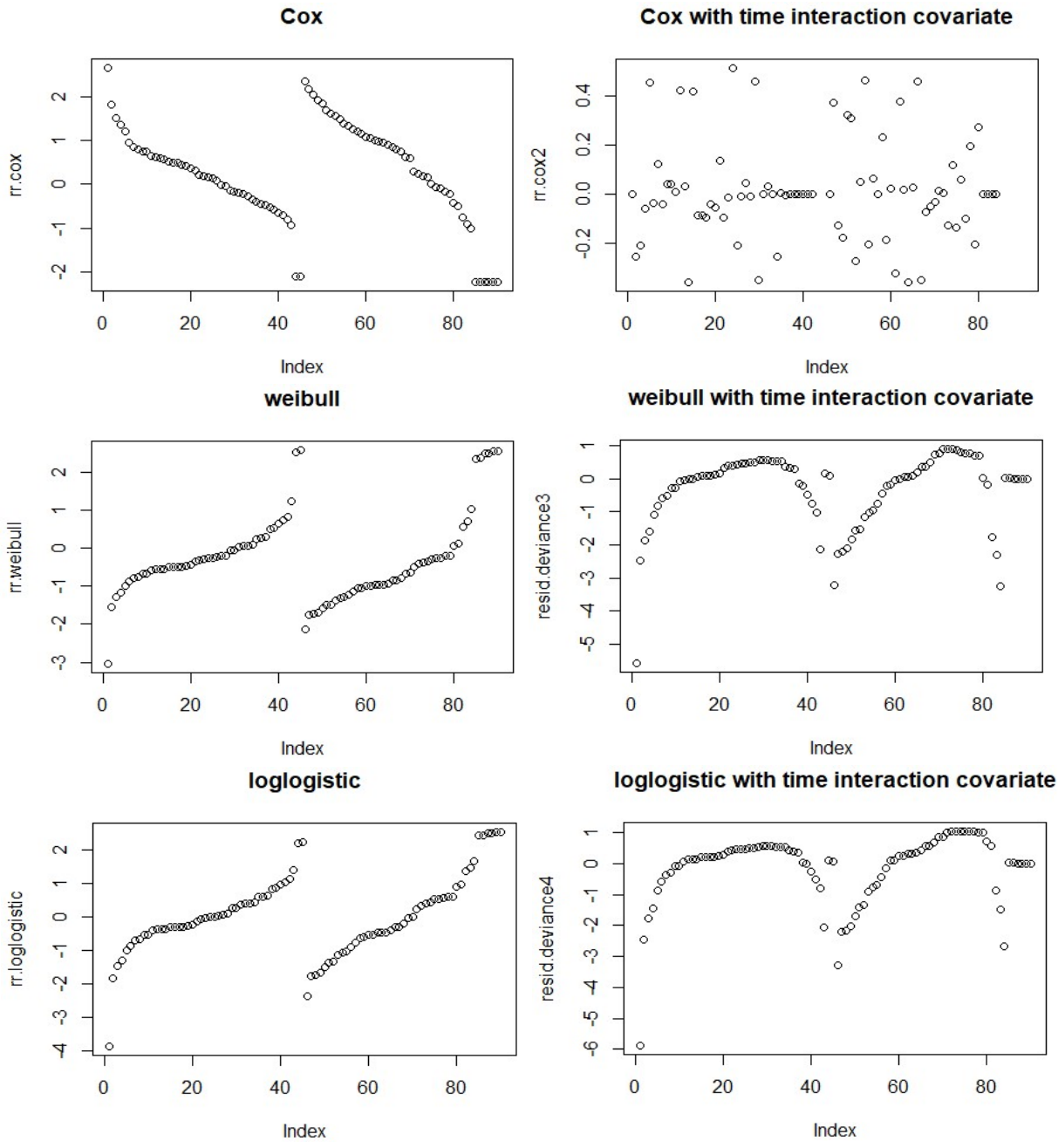
4. (20 points) Which model do you recommend for the survival analysis? Explain.

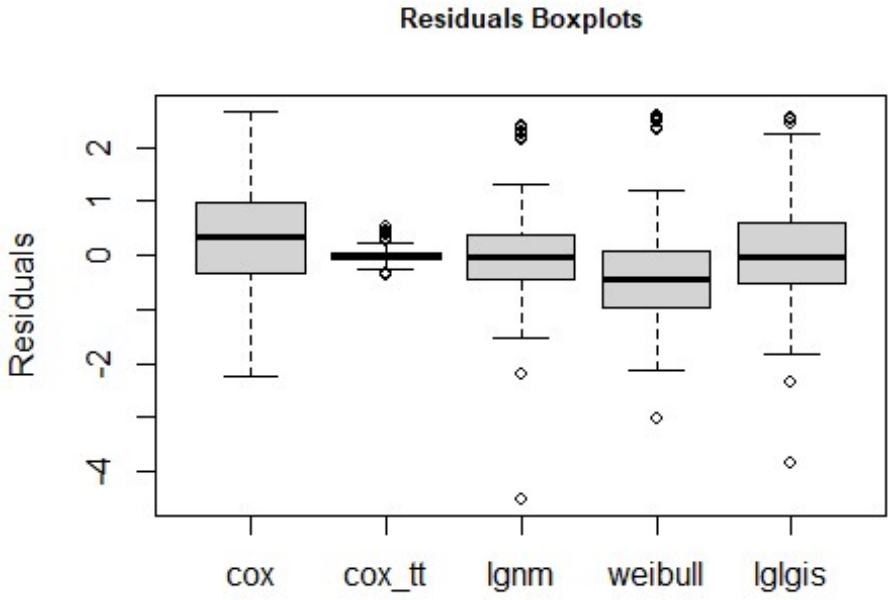
For all these models, I recommend the Cox model with time*treatment interaction covariate.

By the following residual plots, we can see that the cox model with a covariate of time interacting with treatment is the only one that doesn't violate the linear assumption for regression models. This model yields a much denser distribution of residuals, as shown in the boxplot, also suggesting a better goodness fit.

For the nonparametric model, there is no function form and can only include very few categorical variables, while cox model has potential to include more covariates into the model and can show how the treatment effect changes over time. Given that the hazard ratio between Chemo group and Chemo plus Radio group are proportional by adding the time interaction covariate, it is appropriate to fit the data by Cox.

Therefore, I recommend cox model with the time*treatment interaction covariate.


Cox
rr.cox
Index
Cox with time interaction covariate
rr.cox2
Index
weibull
rr.weibull
Index
weibull with time interaction covariate
resid.deviance3
Index
loglogistic
rr.loglogistic
Index
loglogistic with time interaction covariate
resid.deviance4
Index


Residuals Boxplots
Residuals
2
1
0
-2
-4
cox
cox_tt
lgnm
weibull
lglgis