Extra Credit

Pomelo Wu

2022-12-14

Introduction

Breast cancer is still of primary concern for women. The Cancer Organization in the United States suggests that the average risk of women encountering breast cancer is about 13%. Therefore, understanding what treatments can help patients with breast cancer is beneficial. In this report, I am interested in examining whether patients having node-positive breast cancer and receiving the hormone trreatment experience increased time to death/recurrence compared to patients who received standard treatments. The dataset is obtained from a 1984-1989 trial conducted by the German Breast Cancer Study Group (GBSG).

Methods

This dataset contains 686 observations and 12 variables in total. The primary response variable of interest is the recurrence free survival time(RFS), and the primary predictor variable of interest is whether patients receive standard treatment or hormone therapy treatment.

1. Exploratory Data Analysis

In this dataset, 440 patients received standard treatment and 246 patients received hormone treatment. Patients who had records of their cancer recurrence or death in the dataset are 299, and patients who had no records of their cancer recurrence or death in the dataset are 387. Among those who received standard treatment, 235 individuals had no records of recurrence/death, and 205 had records. Among those who received hormone treatment, 152 individuals had no records and 94 had records (See Figure 1). Based on Figure 1, the proportion of individuals having recurrence records versus not among individuals receiving different treatments seems quite even. As for our primary variable of interest, `RFS time`, we can observe a very *distinct distribution pattern* for patients receiving standard treatment versus receiving hormone treatment. Because of this characteristic, we need to be careful in choosing the model.

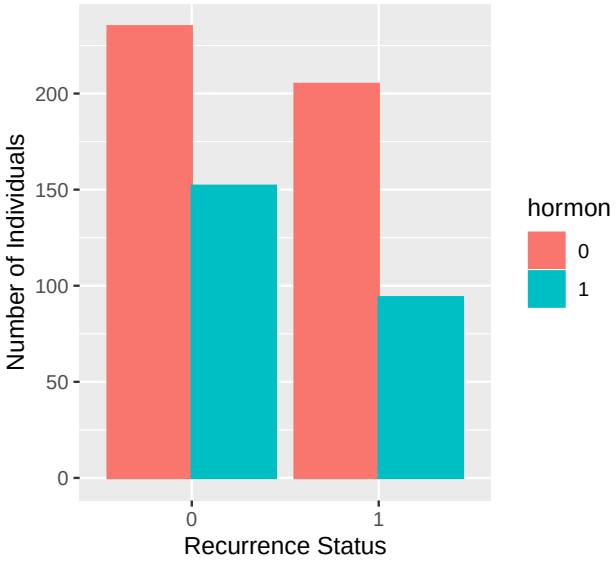


Figure 1: Recurrence Status for Patients Receiving Hormone Treatment vs. Standard Treatment

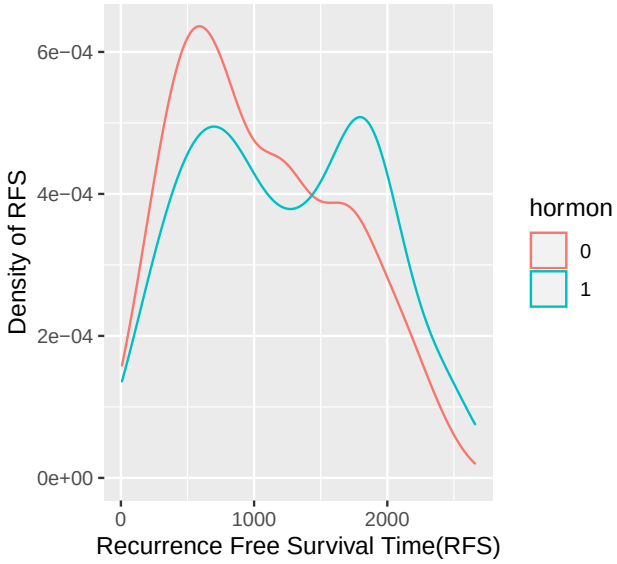


Figure 2: Density Distribution of RFS Time for Patients with/without Recurrence/Death

Some important descriptive statistics as well as demonstrative figures of other predictor/control variables are in Appendix I. All other variables except `age` show similar patterns and distributions among individuals who received hormone treatment and individuals who received standard treatment, meaning that the mean and standard deviations do not differ much. There are two primary possibilities: a) age might be a contributory factor to the difference in RFS time, which would be reflected when I utilized the model to analyze, and b) the data is biased and needs more control. For the purpose of this project, I kept `age` in the final analysis and discussed this potential limitation.

2. Model Selection

In this analysis, our primary variable of interest is RFS. This variable needs careful consideration. For individuals who had not yet experienced death or recurrence, the total survival time is unknown. In some other circumstances, researchers failed to follow up. Therefore, the total RFS for individuals who had not experienced death/recurrence is not accurate. To address this issue, I did some research and found that there are a few model options, including `Kaplan Meier Analysis` and `Cox Regression`. `Kaplan Meier Analysis` does not have rigorous assumptions and can help analyze the risk of a patient over time. However, it cannot estimate the effect of other covariates. In this analysis, nevertheless, I am also interested in whether other factors can contribute to the RFS. As a result, I decided to adopt `Cox Regression` in this analysis.

`Cox regression` helps investigate the effect of the predictor variable on the time a specified event takes to happen. Since our primary variable of interest is the survival time, the Cox regression serves this purpose perfectly. Cox Regression has these `assumptions`: (1) One of the assumptions is the proportional hazards (PH), which suggests that the predictor variable should have a constant impact on the risk over time, and (2) the other one is the relationship between the log hazard and each covariate is linear. The first one is the most important assumption, and these assumptions can be checked in the residual plots.

After determining to use the Cox regression model, I also utilized `bidirectional stepwise methods` to select variables. The result confirms the inclusion of all the variables I chose: `age`,`size`,`grade`,`nodes`,`pgr`,`er`,and `meno`.

Results

1. Model Assessment

I utilized the `scaled Schoenfeld residuals against the transformed time` to test the proportional hazards (PH) assumption. The global Schoenfeld test p has a value of 0.0047, and based on the graph, there are no obvious patterns for each predictor variable over time. Therefore, the PH assumption is held. Additionally, we often assumes that the covariates have a linear relationship over time, therefore, I also plotted martingale residuals and partial residuals against a continuous variable to test this assumption (See Appendix for a sample code and graph). After testing these assumptions, I failed to find any alarming results, meaning that the assumptions are not violated. Therefore, we could utilize the model to conduct our analysis. Finally, I also checked whether `influential points` exist in the analysis or not using deviance test specific to the Cox regression model. The results does not suggest any significant influential points that are detrimental to our statistical analysis (Figure 4).

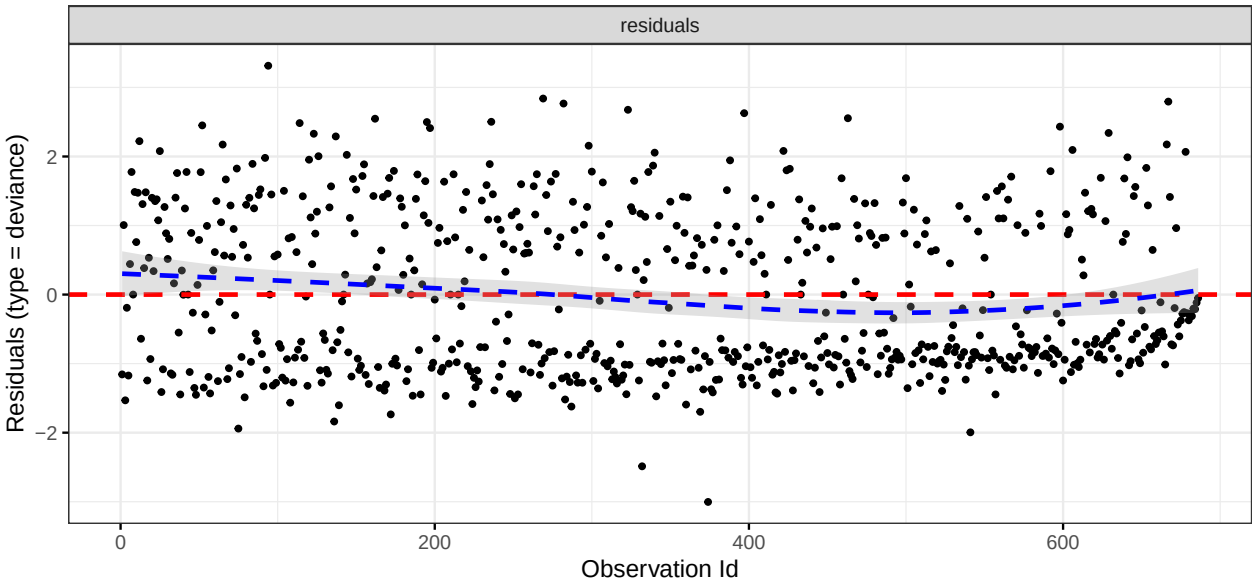


Figure 4: Influential Points Check

2. Results Interpretation

The results suggest that hormone treatment, size, grade, number of nodes, and progesterone receptors are statistically significant and might contribute to the RFS time. In cox regression Model, if the sign of β is negative, it suggests a reduced risk. If the sign of β is positive, then the risk is increased. Based on the model results, compared to the standard treatment, patients receiving the hormone treatment have 0.71 time less in the expected hazard with 95% confidence intervals between 0.55 and 0.92 ($p<0.01$). In other words, patients with hormone treatment are likely to have 29% reduced risk compared to patients with standard treatment. One unit increase in size is associated with 1.008 time more expected risk with 95% confidence intervals between 0.004 and 1.95 ($p<0.05$). One unit increase in the level of tumor grade is associated with the 32% increase (1.32 times more likely) in the expected hazard with 95% confidence intervals between 1.08 and 1.63 ($p<0.01$). One unit increase in the number of nodes is associated with 5% increase (or 1.05 times more likely) with 95% confidence intervals between 1.036 and 1.065 in the expected risk ($p<0.001$), and one unit increase in the progesterone receptors is related to 0.997 times decrease with 95% confidence intervals between 0.997 and 0.999 in the expected hazard ($p<0.001$).

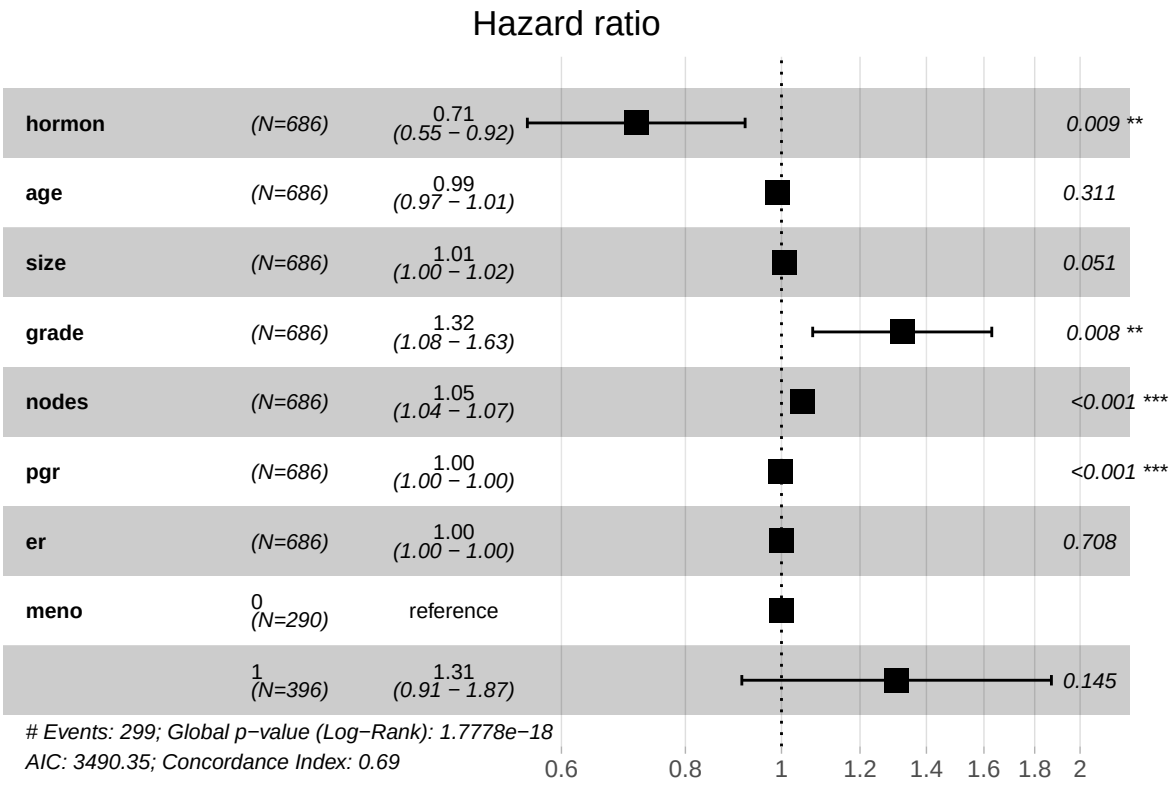


Figure 3: Hazard Ratio for all Predictor Variables

Importantly, to examine the robustness of the model, I also utilized a package to conduct the **model validation**. The internal prediction accuracy is 67.77%, which is a relative high value and suggests that this model is quite accurate.

Conclusion

1. Key Findings
   Using Cox Regression Model, I found that patients receiving hormone treatment have a potentially lower risk of encountering recurrence/death of breast cancer compared to patients receiving standard treatment. Other variables, `size`, `grade`, `the number of nodes`, and `progesterone receptors` also contribute to the expected hazard of breast cancer. While one unit increase in `size`, `grade`, and `the number of nodes` increase the risk, one unit increase in `progesterone receptors` reducess risk.
2. Limitations

As I stated above, the variable `age` might need more exmination and transformation in statistical analysis. However, because I did not find any violation of assumptions for analysis, I believe including age variable did not make any detriments in my analysis. Additionally, arguably, in randomized experiments, the characteristics of patients are random enough. On the other hand, it is also likely that age is associated with increased expect hazard in breast cancer though the model results suggest otherwise. As a result, this analysis might consider transforming some variables using `stratification` and/or `adding covariate*time interaction` to increase the robustness of the analysis. Finally, this research does not include external validations because of technical issues. Adding `external validations` such as k-fold validation might also be beneficial.

3. Future Direction
   Future researchers could fine-tune the analysis, meaning that they can focus on speicfic subgroups of the patients to better understand the effect of hormone treatment. For instance, `race` and `income` are often important variables in survival analysis. Moreover, though I applied KM analysis during the model selection stage, I did not compare the result. Additionally, there might be other helpful methods in building model. Therefore, future studies can also consider comparing different analyses and compare the accuracy, validity, and robustness of these models.

Appendix I: Tables and Figures

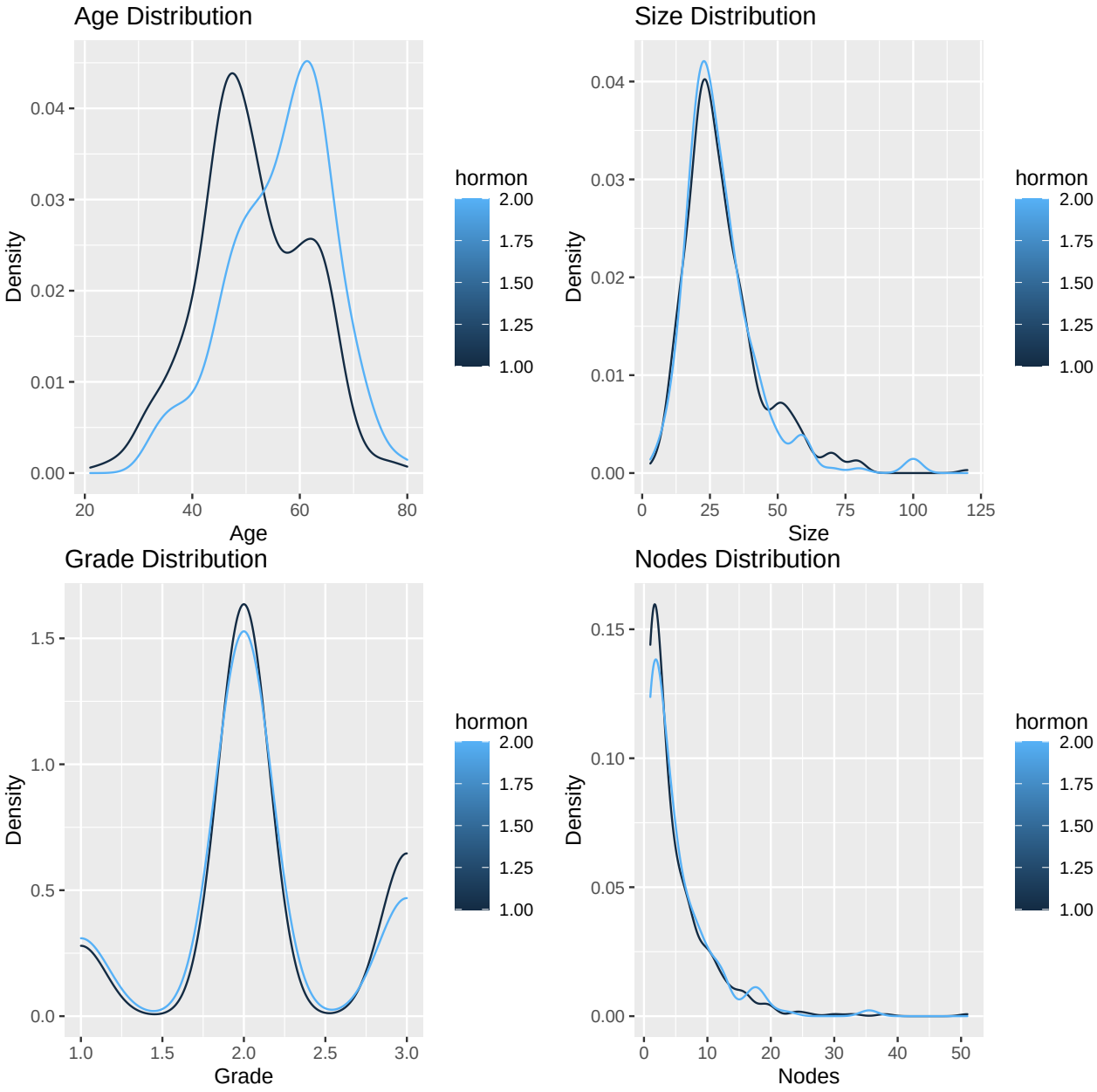
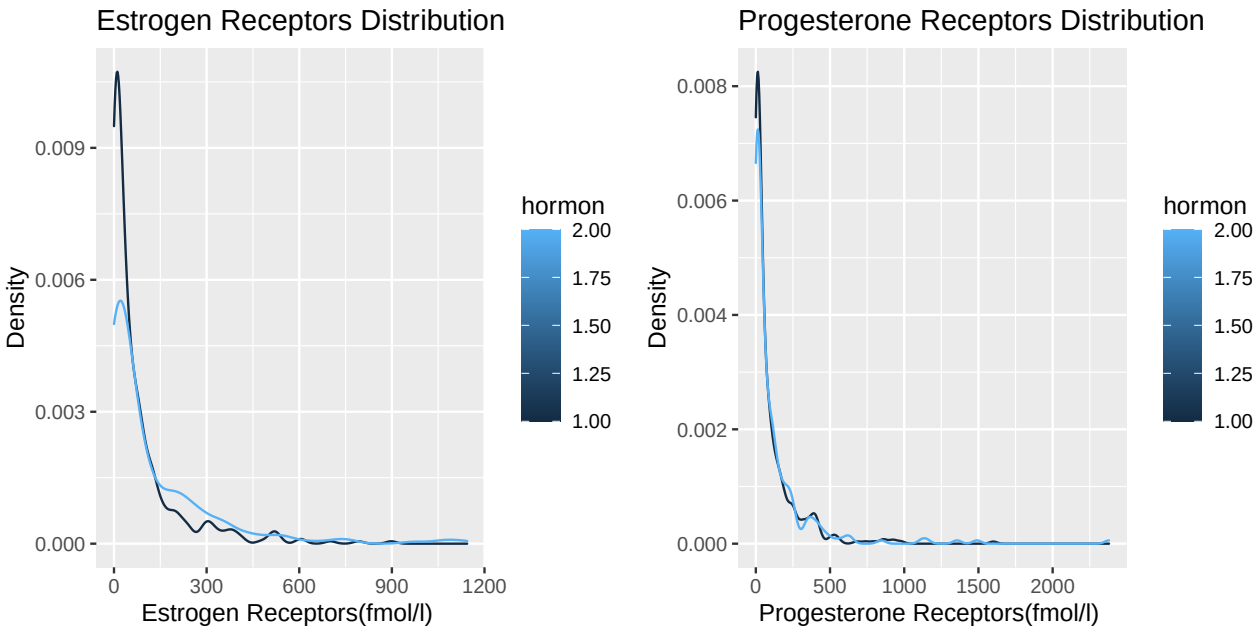


Table 1: Table 1: Summary Statistics for Patients with Hormone Treatment

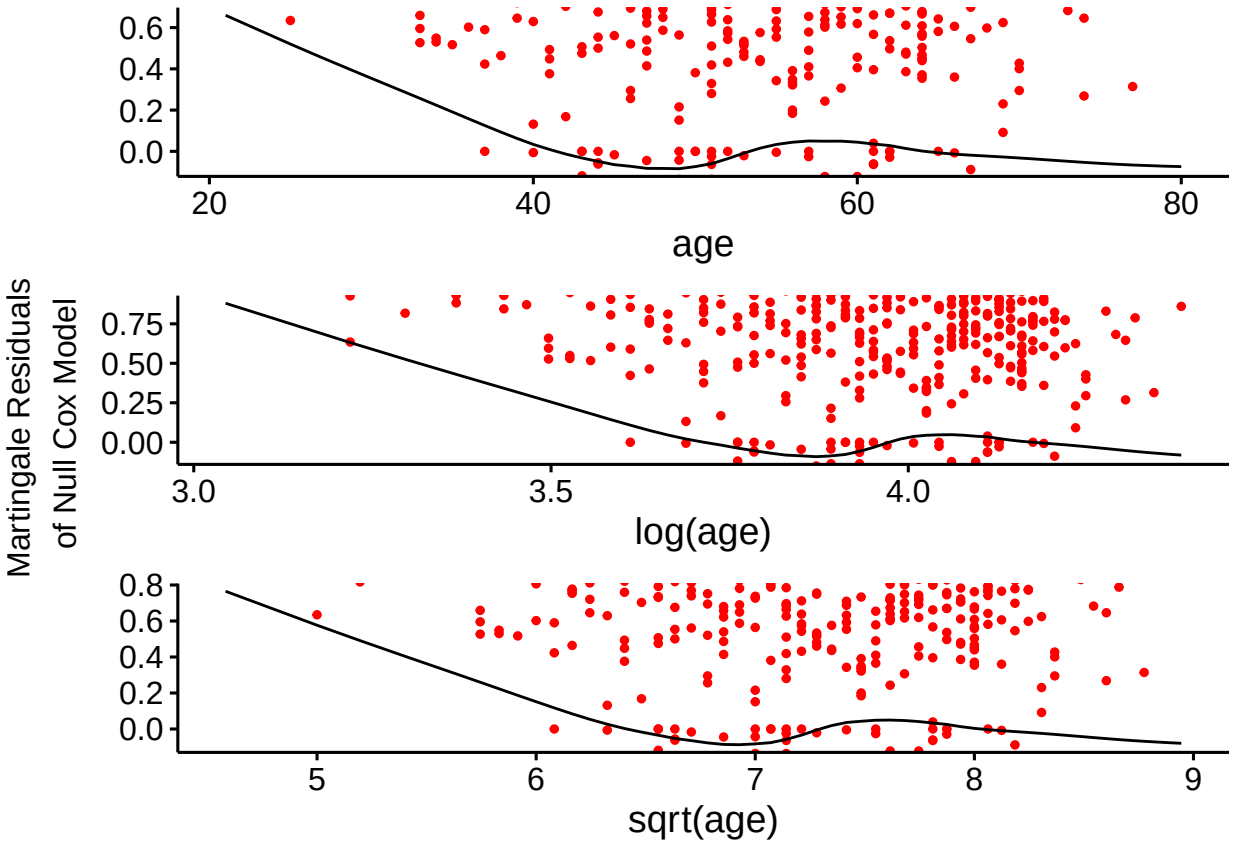
	age	size	grade	nodes	er	pgr	rfstime
	Min. :32.00	Min. : 4.0	Min. :1.000	Min. : 1.00	Min. : 0.0	Min. : 0.00	Min. : 15.0
	1st Qu.:50.00	1st Qu.: 20.0	1st Qu.:2.000	1st Qu.: 1.00	1st Qu.: 9.0	1st Qu.: 7.25	1st Qu.: 695.8
	Median :58.00	Median : 25.0	Median :2.000	Median : 3.00	Median : 46.0	Median : 35.00	Median :1220.5
	Mean :56.62	Mean : 28.8	Mean :2.069	Mean : 5.13	Mean : 125.8	Mean : 124.29	Mean :1240.3
	3rd Qu.:63.00	3rd Qu.: 35.0	3rd Qu.:2.000	3rd Qu.: 7.00	3rd Qu.: 182.5	3rd Qu.: 133.00	3rd Qu.:1818.0
	Max. :80.00	Max. :100.0	Max. :3.000	Max. :36.00	Max. :1144.0	Max. :2380.00	Max. :2659.0

Table 2: Table 2: Summary Statistics for Patients with Standard Treatment

	age	size	grade	nodes	er	pgr	rfstime
	Min. :21.00	Min. : 3.00	Min. :1.000	Min. : 1.000	Min. : 0.00	Min. : 0	Min. : 8.0
	1st Qu.:45.00	1st Qu.: 20.00	1st Qu.:2.000	1st Qu.: 1.000	1st Qu.: 8.00	1st Qu.: 7	1st Qu.: 547.8
	Median :50.00	Median : 25.00	Median :2.000	Median : 3.000	Median : 32.00	Median : 32	Median : 967.0
	Mean :51.06	Mean : 29.62	Mean :2.143	Mean : 4.943	Mean : 79.72	Mean : 102	Mean :1059.7
	3rd Qu.:59.00	3rd Qu.: 35.00	3rd Qu.:3.000	3rd Qu.: 7.000	3rd Qu.: 92.25	3rd Qu.: 130	3rd Qu.:1573.0
	Max. :80.00	Max. :120.00	Max. :3.000	Max. :51.000	Max. :898.00	Max. :1600	Max. :2563.0




```
## Warning: arguments formula is deprecated; will be removed in the next version;
## please use fit instead.
```


Martingale Residuals of Null Cox Model
0.6
0.4
0.2
0.0
20
40
60
80
age
0.75
0.50
0.25
0.00
3.0
3.5
4.0
log(age)
0.8
0.6
0.4
0.2
0.0
5
6
7
8
9
sqrt(age)

#Appendix II: KM analysis

```
km <- with(breast,Surv(rfstime,status))
km_fit <- survfit(Surv(rfstime, status) ~ 1, data=breast)
summary(km_fit, times = c(1,30,60,90*(1:10)))
```

```
## Call: survfit(formula = Surv(rfstime, status) ~ 1, data = breast)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    686       0    1.000 0.00000        1.000        1.000
##    30    679       0    1.000 0.00000        1.000        1.000
##    60    676       0    1.000 0.00000        1.000        1.000
##    90    671       1    0.999 0.00149        0.996        1.000
##   180    657      11    0.982 0.00512        0.972        0.992
##   270    638      14    0.961 0.00749        0.946        0.976
##   360    603      30    0.916 0.01080        0.895        0.937
##   450    573      25    0.877 0.01276        0.853        0.903
##   540    538      31    0.830 0.01467        0.801        0.859
##   630    490      35    0.775 0.01635        0.744        0.808
##   720    465      14    0.753 0.01694        0.720        0.787
##   810    417      24    0.713 0.01790        0.678        0.749
##   900    383      20    0.678 0.01865        0.642        0.715
```

Appendix III: Code for Extra Credit Assignment

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(ggplot2)
library(tidyverse)
library(gridExtra)
library(knitr)
library(jtools)
library(ISLR)
library(arm)
library(caret)
library(e1071)
library(pROC)
library(stargazer) #Table
library(kableExtra)
library(MASS) #Model selection
library(rms) #VIF
library(cobalt)
library(survival)
library(survminer)
library(StepReg)
rm(list=ls())
breast <- read.csv('/Users/wuyuyou/Desktop/breastcancer.csv', header=T, sep = ",", dec=",")
str(breast)
head(breast)
breast$meno <- as.factor(breast$meno)
breast$status <- as.factor(breast$status)
breast$hormon <- as.factor(breast$hormon)
num <- breast[,c(3,5:9,11)]
cat <- breast[,c(4,10,12)]
summary(num)
summary(cat)
hist(breast$pgr)
hist(breast$er)
hist(breast$size)
hist(breast$rfstime)
control <- subset(breast, breast$hormon==0)
treated <- subset(breast, breast$hormon==1)
summary(control)
summary(treated)
p1 <- ggplot(breast,aes(x=status,group=hormon))+
  geom_bar(position="dodge",aes(color=hormon,fill=hormon))+
   labs(x="Recurrence Status",y="Number of Individuals",
        caption='Figure 1: Recurrence Status for Patients Receiving
        Hormone Treatment vs. Standard Treatment')+
   theme(plot.caption = element_text(size=11,hjust=0.5))
p2<- ggplot(breast,aes(x=rfstime,y=..density..,group=hormon))+
  geom_density(aes(color=hormon))+
  labs(x="Recurrence Free Survival Time(RFS)",y="Density of RFS",
       caption='Figure 2: Density Distribution of RFS Time for
       Patients with/without Recurrence/Death')+
  theme(plot.caption = element_text(size=11,hjust=0.5))
```

```
grid.arrange(p1,p2,ncol=2)
breast$hormon <- as.integer(breast$hormon)
breast$status <- as.integer(breast$status)
formula = Surv(rfstime, status) ~ hormon+age+size+grade+nodes+pgr+er+meno
stepwiseCox(formula,
breast,
selection=c("bidirection"),
select="HQ",
method=c("efron"),
sle=0.15,
sls=0.15,
weights=NULL,
best=NULL)
breast$status <- as.integer(breast$status)
surv_object <- Surv(time=breast$rfstime,event=breast$status)
mod <- coxph(surv_object~hormon+age+size+grade+nodes+pgr+er+meno, data=breast)
summary(mod)
test <- cox.zph(mod)
ggcoxzph(test)
ggcoxfunctional(Surv(rfstime, status) ~ age+log(age)+sqrt(age),breast)
ggcoxfunctional(Surv(rfstime, status) ~ size+log(size)+sqrt(size),breast)
ggcoxfunctional(Surv(rfstime, status) ~ grade+log(grade)+sqrt(grade),breast)
ggcoxfunctional(Surv(rfstime, status) ~ nodes+log(nodes)+sqrt(nodes),breast)
ggcoxdiagnostics(mod, type = "dfbeta", linear.predictions = FALSE,ggtheme = theme_bw())
ggcoxdiagnostics(mod, type = "deviance", linear.predictions = FALSE, ggtheme = theme_bw())+
  labs(caption='Figure 4: Influential Points Check')+
  theme(plot.caption = element_text(size=11,hjust=0.5))
ggforest(mod,data=breast)+labs(
       caption='Figure 3: Hazard Ratio for all Predictor Variables')+
  theme(plot.caption = element_text(size=11,hjust=0.5))
mod2 <- cph(Surv(rfstime, status)~hormon+age+size+grade+nodes+pgr+er+meno,
            data=breast,
            x=TRUE, y=TRUE)
v1 <- validate(mod2, method="boot", B=40, bw=FALSE, rule="aic",
type="residual", sls=.05, aics=0, force=NULL, estimates=TRUE,
pr=FALSE, dxy=TRUE, u, tol=1e-9)
c_index <- abs(v1[1,5])/2 + 0.5
c_index
p1 <- ggplot(breast,aes(x=age,y=..density..,group=hormon))+
  geom_density(aes(color=hormon))+
  labs(title="Age Distribution",
       x="Age",y="Density")
p2 <- ggplot(breast,aes(x=size,y=..density..,group=hormon))+
  geom_density(aes(color=hormon))+
  labs(title="Size Distribution",
       x="Size",y="Density")
p3 <- ggplot(breast,aes(x=grade,y=..density..,group=hormon))+
  geom_density(aes(color=hormon))+
  labs(title="Grade Distribution",
       x="Grade",y="Density")
p4 <- ggplot(breast,aes(x=nodes,y=..density..,group=hormon))+
  geom_density(aes(color=hormon))+
  labs(title="Nodes Distribution",
```

```
       x="Nodes",y="Density")
p5 <- ggplot(breast,aes(x=er,y=..density..,group=hormon))+
  geom_density(aes(color=hormon))+
  labs(title="Estrogen Receptors Distribution",
       x="Estrogen Receptors(fmol/l) ",y="Density")
p6 <- ggplot(breast,aes(x=pgr,y=..density..,group=hormon))+
  geom_density(aes(color=hormon))+
  labs(title="Progesterone Receptors Distribution",
       x="Progesterone Receptors(fmol/l) ", y="Density")
p7 <- ggplot(breast,aes(x=meno,group=hormon))+
  geom_bar(position="dodge",aes(color=hormon,fill=hormon))+
   labs(title="Meno Status",
        x="Meno Status",y="Number of Individuals")
grid.arrange(p1,p2,ncol=2)
grid.arrange(p3,p4,ncol=2)
grid.arrange(p5,p6,ncol=2)
knitr::kable(summary(treated[c('age','size','grade','nodes','er','pgr','rfstime')]),
             caption = "Table 1: Summary Statistics for Patients with Hormone Treatment")
knitr::kable(summary(control[c('age','size','grade','nodes','er','pgr','rfstime')]),
             caption = "Table 2: Summary Statistics for Patients with Standard Treatment")
ggcoxfunctional(Surv(rfstime, status) ~ age+log(age)+sqrt(age),breast)
km <- with(breast,Surv(rfstime,status))
km_fit <- survfit(Surv(rfstime, status) ~ 1, data=breast)
summary(km_fit, times = c(1,30,60,90*(1:10)))
```