

project2

Ze Li

2023-12-18

Libraries

```
library(tidyverse)
library(readr)
library(boot)
library(table1)
library(gridExtra)
library(MASS)
library(car)
library(dplyr)
library(leaps)
library(corrplot)
library(survival)
```

Data Clean

```
breastcancer_data =
  read_csv("Project_2_data.csv") |>
  janitor::clean_names()
```

```
## Rows: 4024 Columns: 16
## -- Column specification -----
## Delimiter: ","
## chr (11): Race, Marital Status, T Stage, N Stage, 6th Stage, differentiate, ...
## dbl (5): Age, Tumor Size, Regional Node Examined, Reginol Node Positive, Su...
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

```
summary(breastcancer_data)
```

```
##      age      race      marital_status      t_stage
## Min.   :30.00  Length:4024      Length:4024      Length:4024
## 1st Qu.:47.00  Class :character  Class :character  Class :character
## Median :54.00  Mode  :character  Mode  :character  Mode  :character
## Mean    :53.97
## 3rd Qu.:61.00
```

```

## Max.      :69.00
##   n_stage      x6th_stage      differentiate      grade
## Length:4024      Length:4024      Length:4024      Length:4024
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##   a_stage      tumor_size      estrogen_status      progesterone_status
## Length:4024      Min.      : 1.00      Length:4024      Length:4024
## Class :character  1st Qu.: 16.00      Class :character  Class :character
## Mode  :character  Median : 25.00      Mode  :character  Mode  :character
##                      Mean      : 30.47
##                      3rd Qu.: 38.00
##                      Max.      :140.00
## regional_node_examined reginol_node_positive survival_months
## Min.      : 1.00      Min.      : 1.000      Min.      : 1.0
## 1st Qu.: 9.00      1st Qu.: 1.000      1st Qu.: 56.0
## Median :14.00      Median : 2.000      Median : 73.0
## Mean   :14.36      Mean   : 4.158      Mean   : 71.3
## 3rd Qu.:19.00      3rd Qu.: 5.000      3rd Qu.: 90.0
## Max.   :61.00      Max.   :46.000      Max.   :107.0
##   status
## Length:4024
## Class :character
## Mode  :character
##
##
##

```

```

bc = breastcancer_data |>
  mutate(
    race = factor(race, levels = c("White", "Black", "Other")),
    marital_status = factor(marital_status, levels = c("Married", "Divorced",
                                                       "Single", "Widowed",
                                                       "Separated")),
    t_stage = factor(t_stage, levels = c("T1", "T2", "T3", "T4")),
    n_stage = factor(n_stage, levels = c("N1", "N2", "N3")),
    x6th_stage = factor(x6th_stage, levels = c("IIA", "IIIA", "IIIC", "IIB", "IIIB")),
    differentiate = factor(differentiate, levels = c("Poorly differentiated",
                                                     "Moderately differentiated",
                                                     "Well differentiated",
                                                     "Undifferentiated")),
    grade = factor(grade, levels = c("1", "2", "3", "anaplastic; Grade IV")),
    a_stage = factor(a_stage, levels = c("Distant", "Regional")),
    estrogen_status = factor(estrogen_status, levels = c("Negative", "Positive")),
    progesterone_status = factor(progesterone_status,
                                 levels = c("Negative", "Positive")),
    status = factor(status, levels = c("Dead", "Alive"))
  )

```

Descriptive statistics for all variables

```
summary(bc)
```

```
##      age      race      marital_status t_stage  n_stage  x6th_stage
## Min.   :30.00  White:3413  Married   :2643  T1:1603  N1:2732  IIA :1305
## 1st Qu.:47.00  Black: 291  Divorced : 486  T2:1786  N2: 820  IIIA:1050
## Median :54.00  Other: 320  Single   : 615  T3: 533  N3: 472  IIIC: 472
## Mean   :53.97                Widowed   : 235  T4: 102                IIB :1130
## 3rd Qu.:61.00                Separated:  45                IIIB: 67
## Max.   :69.00
##
##      differentiate      grade      a_stage
## Poorly differentiated :1111  1      : 543  Distant : 92
## Moderately differentiated:2351  2      :2351  Regional:3932
## Well differentiated      : 543  3      :1111
## Undifferentiated         :  19  anaplastic; Grade IV: 19
##
##
##      tumor_size      estrogen_status progesterone_status regional_node_examined
## Min.   : 1.00  Negative: 269  Negative: 698  Min.   : 1.00
## 1st Qu.: 16.00  Positive:3755  Positive:3326  1st Qu.: 9.00
## Median : 25.00
## Mean   : 30.47
## 3rd Qu.: 38.00
## Max.   :140.00
##
## reginol_node_positive survival_months  status
## Min.   : 1.000  Min.   : 1.0  Dead : 616
## 1st Qu.: 1.000  1st Qu.: 56.0  Alive:3408
## Median : 2.000  Median : 73.0
## Mean   : 4.158  Mean   : 71.3
## 3rd Qu.: 5.000  3rd Qu.: 90.0
## Max.   :46.000  Max.   :107.0
```

Fit a Cox Proportional Hazards Model

```
bc$survival_months <- as.numeric(bc$survival_months)
bc$status <- as.numeric(bc$status)
surv_obj <- Surv(time = bc$survival_months, event = bc$status)

# Fit the Cox model
cox_model <- coxph(surv_obj ~ age + race + marital_status + t_stage + n_stage + x6th_stage +
  differentiate + grade + a_stage + tumor_size + estrogen_status +
  progesterone_status + regional_node_examined + reginol_node_positive, data = bc)
summary(cox_model)

## Call:
## coxph(formula = surv_obj ~ age + race + marital_status + t_stage +
##      n_stage + x6th_stage + differentiate + grade + a_stage +
##      tumor_size + estrogen_status + progesterone_status + regional_node_examined +
##      reginol_node_positive, data = bc)
```

```
##
##   n= 4024, number of events= 3408
##
##               coef   exp(coef)   se(coef)      z
## age           -0.0004124  0.9995877  0.0020594 -0.200
## raceBlack      0.0725419  1.0752378  0.0713977  1.016
## raceOther     -0.0282799  0.9721162  0.0623962 -0.453
## marital_statusDivorced -0.0604365  0.9413536  0.0549382 -1.100
## marital_statusSingle -0.0482234  0.9529209  0.0496924 -0.970
## marital_statusWidowed -0.0303772  0.9700795  0.0782217 -0.388
## marital_statusSeparated 0.2309519  1.2597987  0.1848733  1.249
## t_stageT2      0.0125409  1.0126199  0.0857679  0.146
## t_stageT3     -0.0847136  0.9187754  0.1434176 -0.591
## t_stageT4     -0.3011978  0.7399314  0.2924931 -1.030
## n_stageN2     -0.1666794  0.8464709  0.1012458 -1.646
## n_stageN3     -0.0261531  0.9741859  0.1405331 -0.186
## x6th_stageIIIA 0.1632452  1.1773253  0.1182755  1.380
## x6th_stageIIIC NA          NA      0.0000000    NA
## x6th_stageIIB -0.0071028  0.9929224  0.0905328 -0.078
## x6th_stageIIIB 0.1569403  1.1699257  0.3163929  0.496
## differentiateModerately differentiated 0.0557485  1.0573318  0.0419336  1.329
## differentiateWell differentiated 0.1215000  1.1291894  0.0580404  2.093
## differentiateUndifferentiated -0.3997794  0.6704679  0.3199544 -1.249
## grade2         NA          NA      0.0000000    NA
## grade3         NA          NA      0.0000000    NA
## gradeanaplastic; Grade IV NA          NA      0.0000000    NA
## a_stageRegional -0.0051530  0.9948603  0.1508261 -0.034
## tumor_size     0.0009129  1.0009133  0.0018429  0.495
## estrogen_statusPositive 0.0915971  1.0959232  0.0909893  1.007
## progesterone_statusPositive 0.1628328  1.1768400  0.0545896  2.983
## regional_node_examined 0.0025479  1.0025511  0.0023305  1.093
## reginol_node_positive -0.0032452  0.9967601  0.0080101 -0.405
## Pr(>|z|)
## age           0.84127
## raceBlack      0.30962
## raceOther      0.65038
## marital_statusDivorced 0.27130
## marital_statusSingle 0.33183
## marital_statusWidowed 0.69776
## marital_statusSeparated 0.21158
## t_stageT2      0.88375
## t_stageT3      0.55474
## t_stageT4      0.30312
## n_stageN2      0.09970
## n_stageN3      0.85237
## x6th_stageIIIA 0.16752
## x6th_stageIIIC NA
## x6th_stageIIB 0.93747
## x6th_stageIIIB 0.61987
## differentiateModerately differentiated 0.18370
## differentiateWell differentiated 0.03632 *
## differentiateUndifferentiated 0.21149
## grade2         NA
## grade3         NA
```

```

## gradeanaplastic; Grade IV          NA
## a_stageRegional                    0.97275
## tumor_size                        0.62035
## estrogen_statusPositive            0.31409
## progesterone_statusPositive        0.00286 **
## regional_node_examined             0.27428
## reginol_node_positive              0.68538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                exp(coef) exp(-coef) lower .95 upper .95
## age                          0.9996      1.0004    0.9956    1.004
## raceBlack                    1.0752      0.9300    0.9348    1.237
## raceOther                    0.9721      1.0287    0.8602    1.099
## marital_statusDivorced       0.9414      1.0623    0.8453    1.048
## marital_statusSingle         0.9529      1.0494    0.8645    1.050
## marital_statusWidowed        0.9701      1.0308    0.8322    1.131
## marital_statusSeparated      1.2598      0.7938    0.8769    1.810
## t_stageT2                    1.0126      0.9875    0.8559    1.198
## t_stageT3                    0.9188      1.0884    0.6936    1.217
## t_stageT4                    0.7399      1.3515    0.4171    1.313
## n_stageN2                    0.8465      1.1814    0.6941    1.032
## n_stageN3                    0.9742      1.0265    0.7396    1.283
## x6th_stageIIIA              1.1773      0.8494    0.9337    1.484
## x6th_stageIIIC              NA          NA        NA        NA
## x6th_stageIIB               0.9929      1.0071    0.8315    1.186
## x6th_stageIIIB              1.1699      0.8548    0.6293    2.175
## differentiateModerately differentiated 1.0573    0.9458    0.9739    1.148
## differentiateWell differentiated    1.1292    0.8856    1.0078    1.265
## differentiateUndifferentiated      0.6705    1.4915    0.3581    1.255
## grade2                      NA          NA        NA        NA
## grade3                      NA          NA        NA        NA
## gradeanaplastic; Grade IV      NA          NA        NA        NA
## a_stageRegional              0.9949      1.0052    0.7403    1.337
## tumor_size                   1.0009      0.9991    0.9973    1.005
## estrogen_statusPositive       1.0959      0.9125    0.9169    1.310
## progesterone_statusPositive    1.1768      0.8497    1.0574    1.310
## regional_node_examined       1.0026      0.9975    0.9980    1.007
## reginol_node_positive        0.9968      1.0033    0.9812    1.013
##
## Concordance= 0.538 (se = 0.006 )
## Likelihood ratio test= 39.61 on 24 df,  p=0.02
## Wald test               = 38.29 on 24 df,  p=0.03
## Score (logrank) test = 38.42 on 24 df,  p=0.03

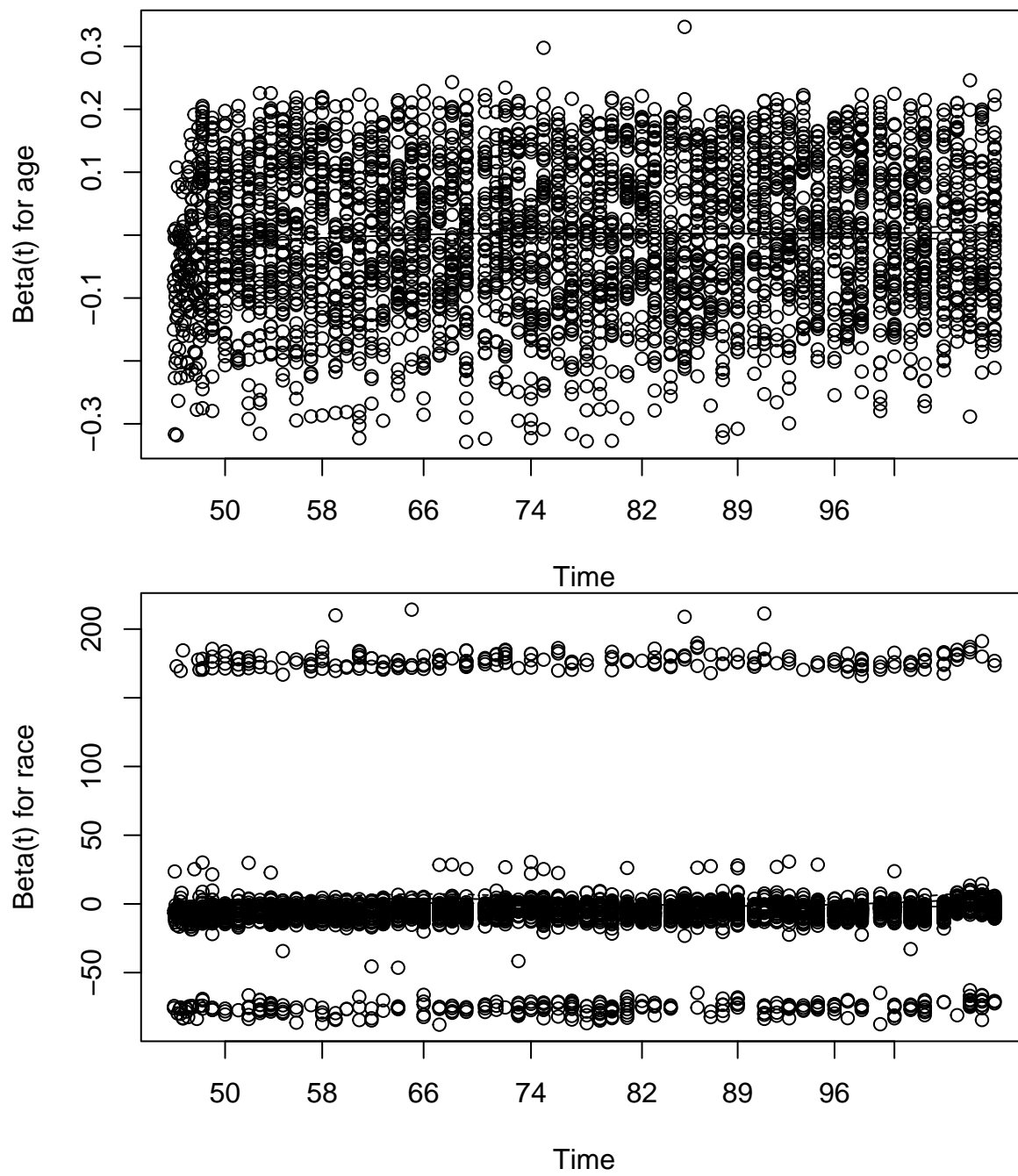
```

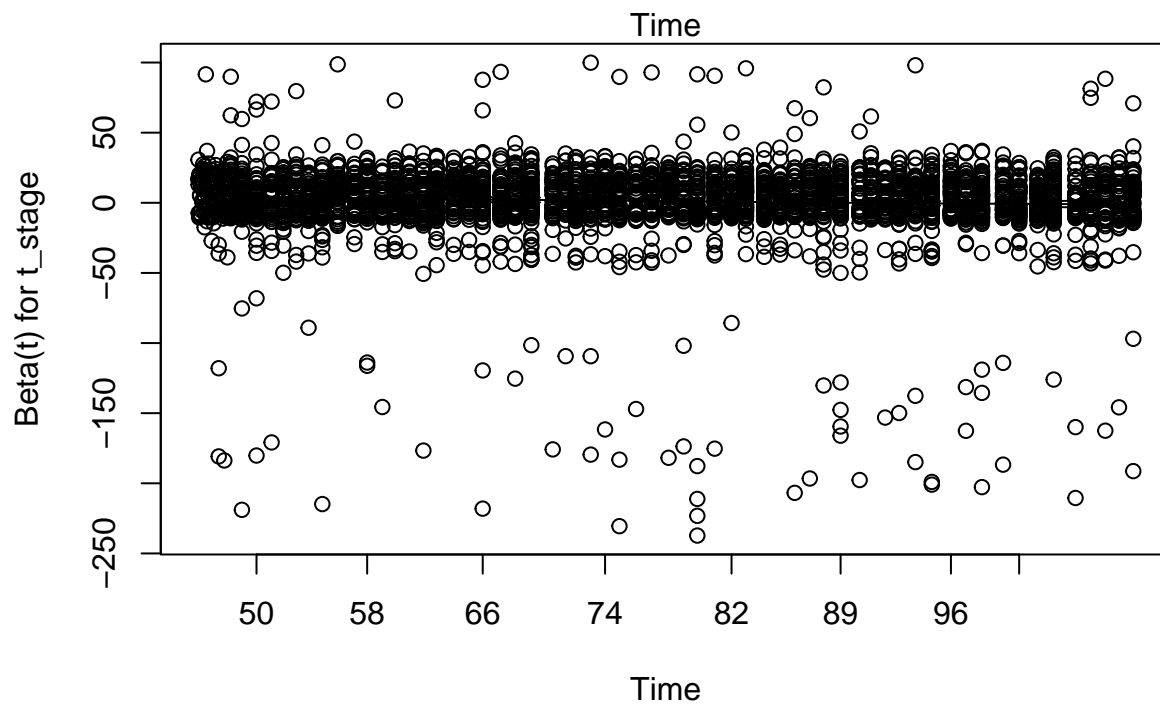
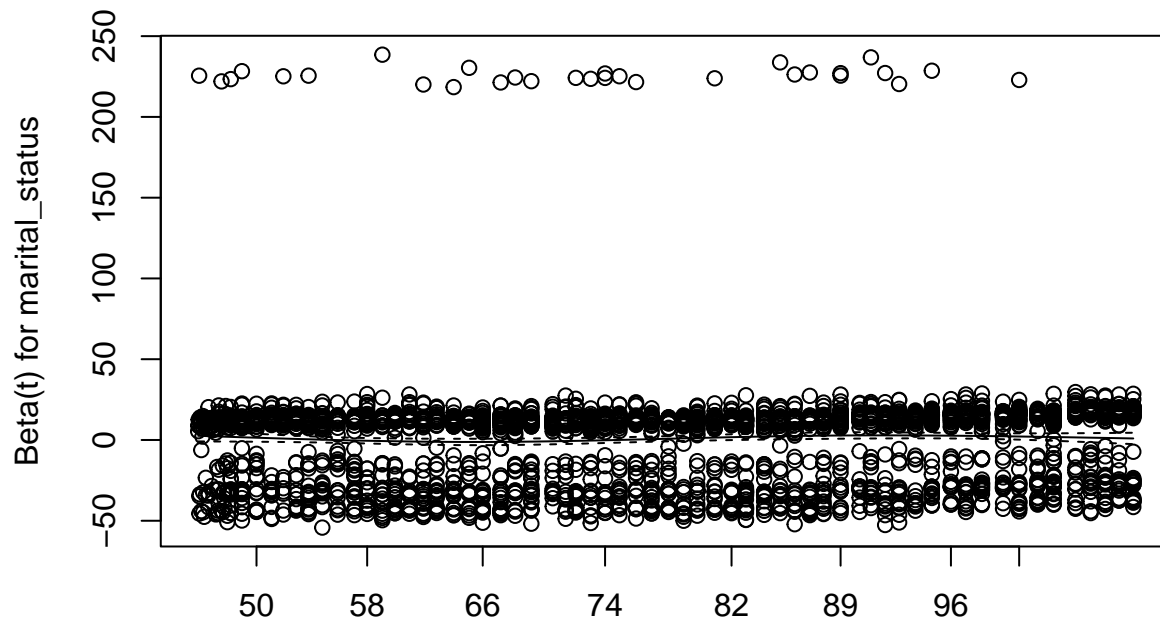
Check Proportional Hazards Assumption

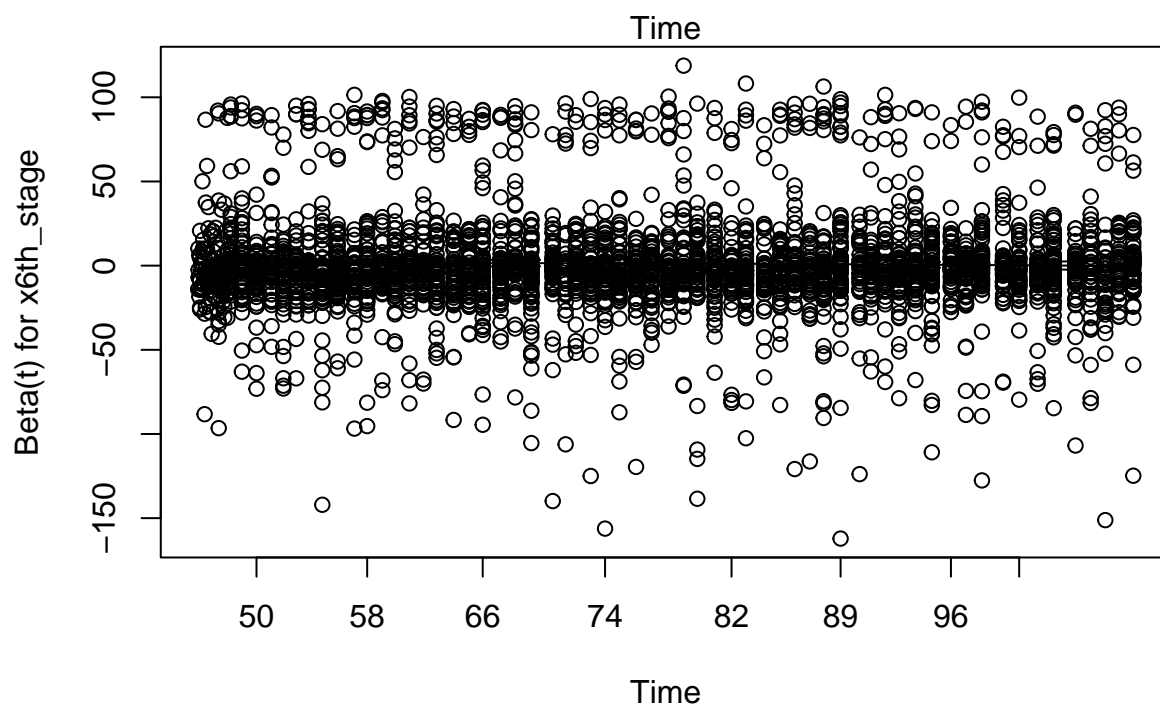
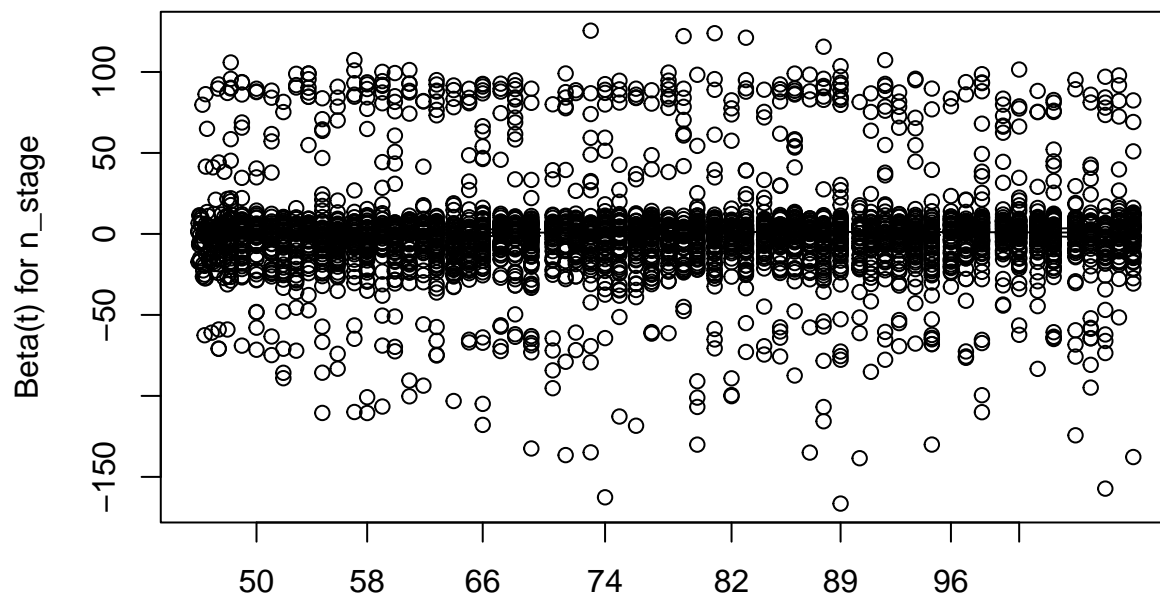
```

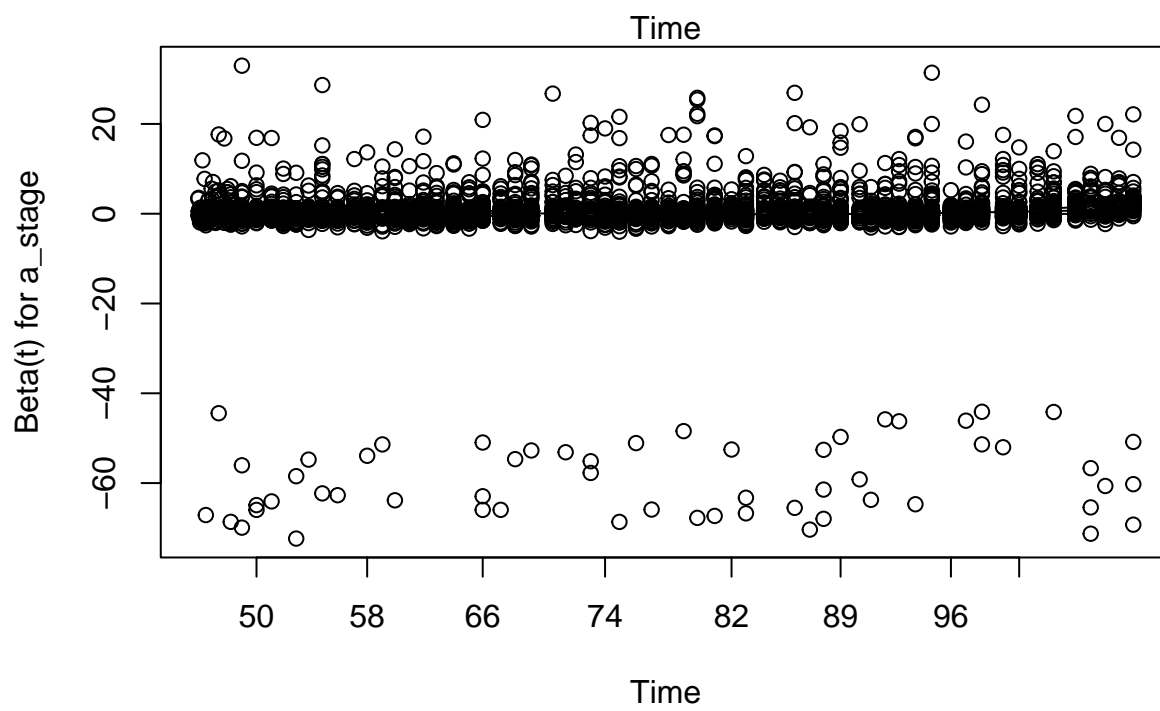
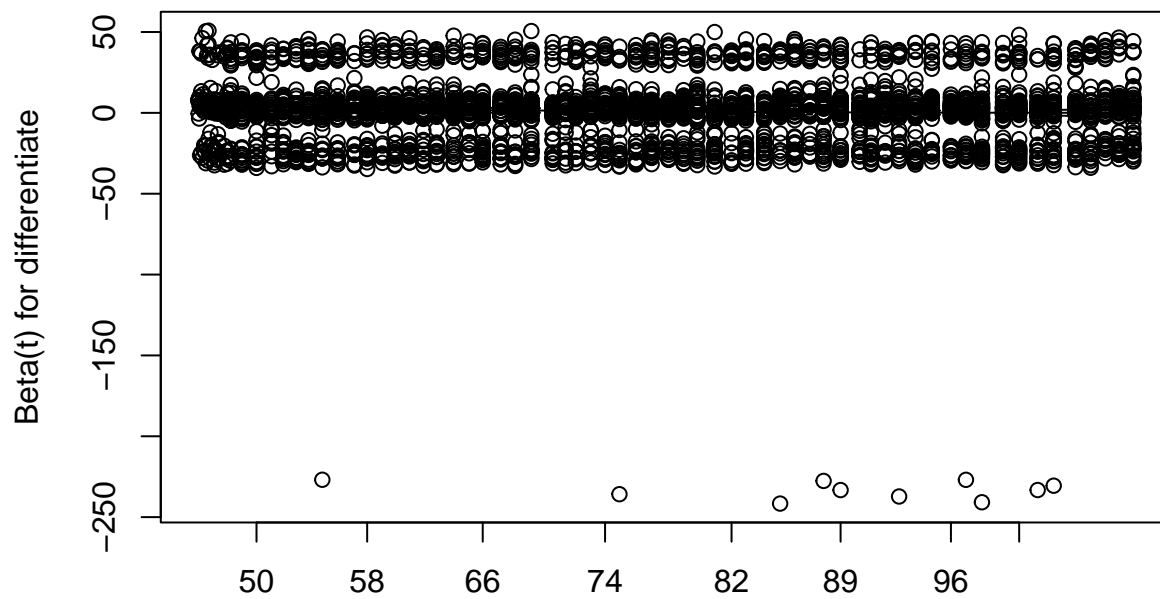
zph <- cox.zph(cox_model)
plot(zph)

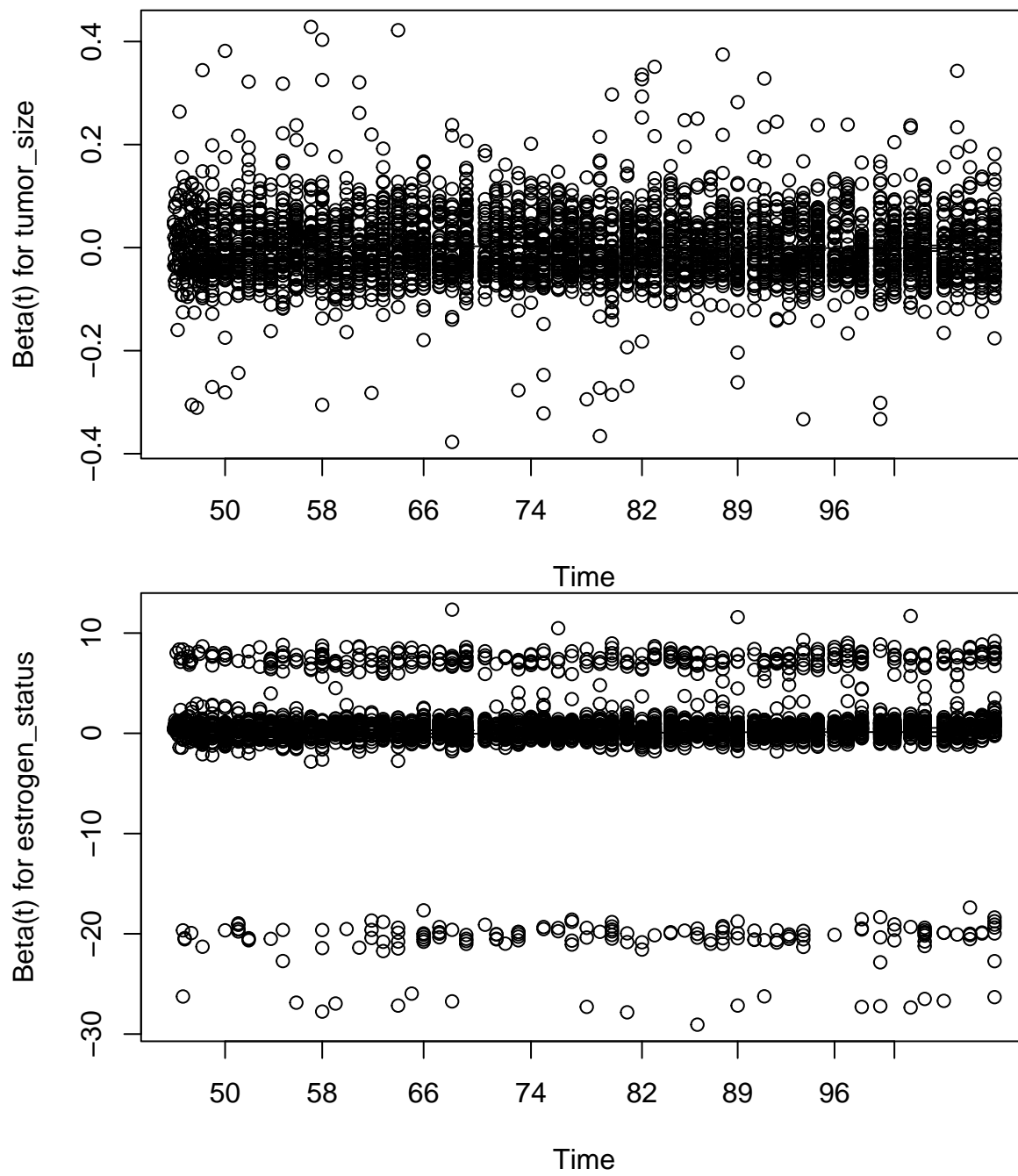
```

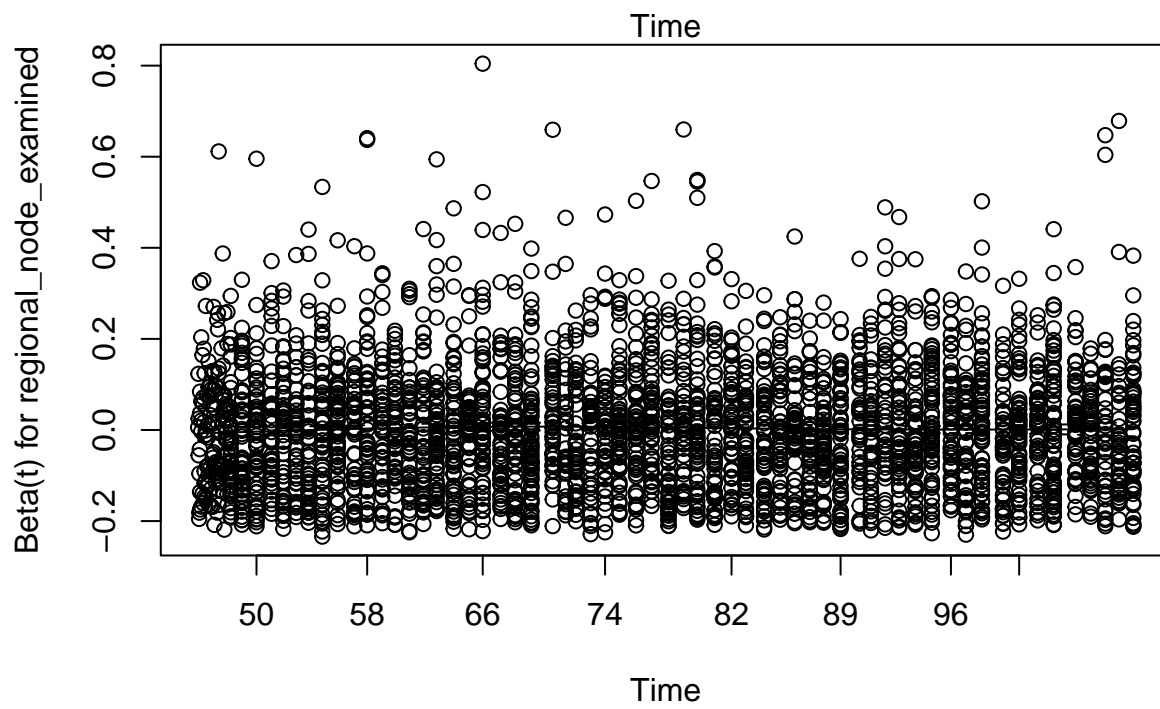
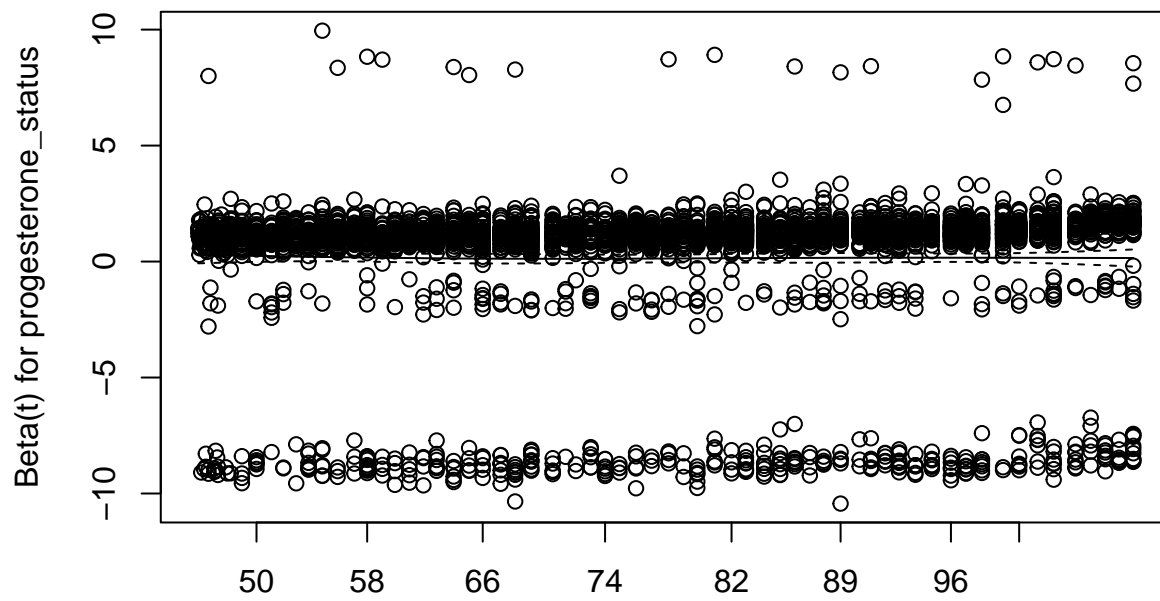


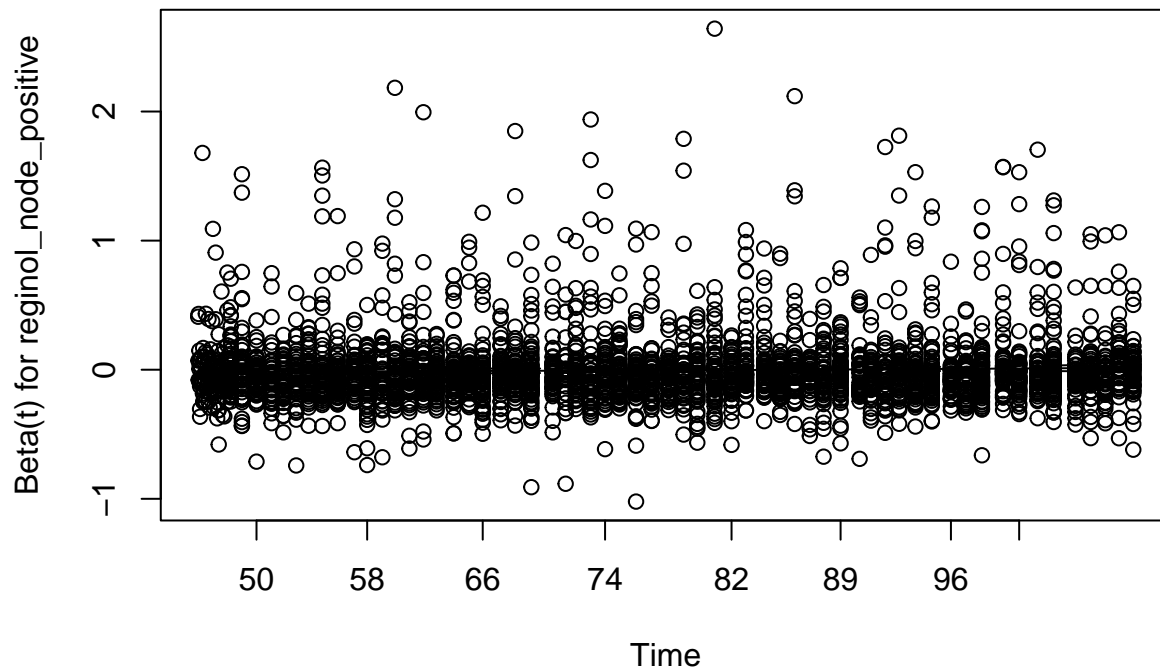








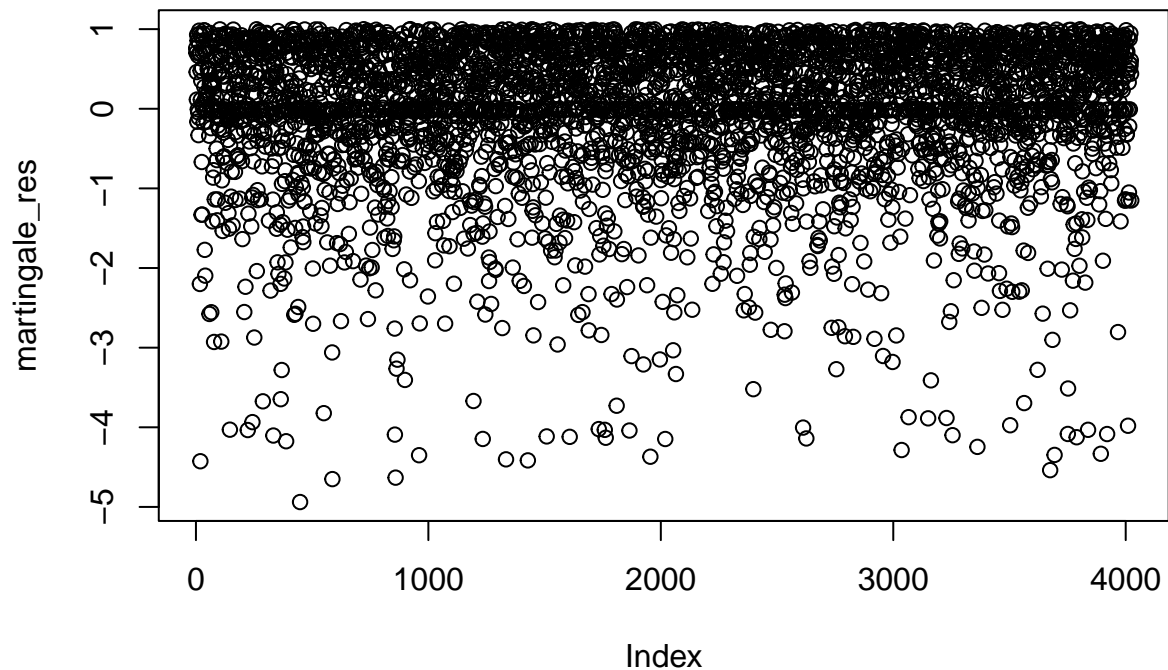




A non-random pattern or a significant global test (p-value) can indicate violations of the assumption.

Identify Influential Observations

```
martingale_res <- residuals(cox_model, type = "martingale")
plot(martingale_res)
```



It seems there are few observations with large negative Martingale residuals, which could be potential outliers or influential observations.