Quiz

Haewon Lee

# 1. Quiz 데이터 주무르기 + 생존분석

## 1.1 Dataframe Index 하는 방법

### 1.1.1 일단 dataframe load 하여 주십시오.

library(magrittr)  
library(tidyverse)  
library(stringr)  
library(reshape2)  
library(gt)  
load("S.rdata")  
df<-S

### 1.1.2 df에서 생존분석과 상관없는 개인정보를 제거하려고 합니다. 컬럼변수명들을 다음과 같이 출력하려면 어떻게 하면 될까요? (colnames, t 함수 참고)

[,1]   
 [1,] "hospital"   
 [2,] "ChartNo"   
 [3,] "Ca\_Dx\_date"   
 [4,] "Sex"   
 [5,] "OpDate"   
 [6,] "OpAge"   
 [7,] "Birthdate"   
 [8,] "NeoAdj"   
 [9,] "optype"   
[10,] "VATS"   
[11,] "Ctsolid\_size"   
[12,] "CT\_size"   
[13,] "path\_size"   
[14,] "Cell\_type"   
[15,] "Patho"   
[16,] "survival"   
[17,] "rec\_free"   
[18,] "Recur\_date"   
[19,] "recurrence"   
[20,] "Op\_mortality"   
[21,] "death"   
[22,] "Death\_date"   
[23,] "last\_fu"   
[24,] "AJCC7"   
[25,] "Tstage"   
[26,] "Nstage"   
[27,] "Mstage"   
[28,] "Differentiation"  
[29,] "Adj\_RT"   
[30,] "adj\_CT"   
[31,] "survival.60"   
[32,] "type"   
[33,] "stage"

### 1.1.3 df에서 다음의 컬럼들을 제거하려고 합니다.

개인정보관련 컬럼들

|  |
| --- |
| hospital, ChartNo, survival.60, last\_fu, AJCC7 |

각자 편한 방법으로 제거하여 주십시오.

제거한 다음에 다음과 같은 결과를 내주십시오.

df %>% colnames

[1] "Ca\_Dx\_date" "Sex" "OpDate" "OpAge"   
 [5] "Birthdate" "NeoAdj" "optype" "VATS"   
 [9] "Ctsolid\_size" "CT\_size" "path\_size" "Cell\_type"   
[13] "Patho" "survival" "rec\_free" "Recur\_date"   
[17] "recurrence" "Op\_mortality" "death" "Death\_date"   
[21] "Tstage" "Nstage" "Mstage" "Differentiation"  
[25] "Adj\_RT" "adj\_CT" "type" "stage"

이번에는 다음의 컬럼들을 제거하려고 합니다.

|  |
| --- |
| Ca\_Dx\_date, OpDate, Birthdate, Death\_date |

colnames(df) 에서 date 또는 Date라는 이름을 가진 변수를 골라서 제거하는 방법으로 컬럼들을 제거하여 주십시오. (regxepr 함수 참고)

[1] "Ca\_Dx\_date" "OpDate" "Birthdate" "Recur\_date" "Death\_date"

제거한 다음에 다음과 같은 결과를 내주십시오.

df %>% colnames

[1] "Sex" "OpAge" "NeoAdj" "optype"   
 [5] "VATS" "Ctsolid\_size" "CT\_size" "path\_size"   
 [9] "Cell\_type" "Patho" "survival" "rec\_free"   
[13] "recurrence" "Op\_mortality" "death" "Tstage"   
[17] "Nstage" "Mstage" "Differentiation" "Adj\_RT"   
[21] "adj\_CT" "type" "stage"

## 1.2 생존분석

### 1.2.1 데이터 준비

library(survival)  
library(moonBook)  
data(cancer) ## survival package내에 있는 각종 암관련 built in 데이터  
rotterdam %>% str

'data.frame': 2982 obs. of 15 variables:  
 $ pid : int 1 2 3 4 5 6 7 8 9 10 ...  
 $ year : int 1992 1984 1983 1985 1983 1983 1993 1988 1988 1988 ...  
 $ age : int 74 79 44 70 75 52 40 53 60 52 ...  
 $ meno : int 1 1 0 1 1 0 0 1 1 0 ...  
 $ size : Factor w/ 3 levels "<=20","20-50",..: 1 2 1 2 1 1 1 1 1 2 ...  
 $ grade : int 3 3 2 3 3 3 2 2 2 3 ...  
 $ nodes : int 0 0 0 0 0 0 0 0 0 5 ...  
 $ pgr : int 35 36 138 0 260 139 13 1 627 316 ...  
 $ er : int 291 611 0 12 409 303 4 4 151 976 ...  
 $ hormon: int 0 0 0 0 0 0 0 0 0 0 ...  
 $ chemo : int 0 0 0 0 0 0 0 0 0 0 ...  
 $ rtime : num 1799 2828 6012 2624 4915 ...  
 $ recur : int 0 0 0 0 0 0 0 0 0 0 ...  
 $ dtime : num 1799 2828 6012 2624 4915 ...  
 $ death : int 0 0 0 0 0 0 0 0 0 0 ...

### 1.2.2 psm 모델 만들 준비

rotterdam 데이터를 살펴보면 age, meno, size, grade, nodes, pgr, er, hormon, chemo 등의 변수가 생존에 영향을 미칠 수 있는 변수로 추정됩니다. 이들 변수들이 생존기간 (dtime, death)에 미치는 모델을 만들어 주십시오.

psm 모델을 만들기 위해서는 비모수 변수를 factor로 바꿔야 합니다. rotterdam 데이터를 rtd.df 라는 데이터 프레임으로 옮기고 자료구조를 psm 모델에 맞게 변형해 주세요. factor로 변형할 필요가 있는 변수들이 있는지 다음과 같은 방법으로 확인하면 됩니다.

rtd.df <- rotterdam  
hline <- function(x){  
 paste0("\n", paste0(rep("-",x),collapse = ""),"\n",collapse = "")  
}  
for (i in 1:length(rtd.df)) {  
 colnames(rtd.df)[i] %>% cat  
 "\n" %>% cat  
 temp <- rtd.df[,i]  
 temp %>% unique %>% length %>%  
 paste(.," cases \n") %>% cat ### unique한 값이 몇개인지  
 temp %>% summary %>% print ### 연속변수인지 요약  
 if(length(unique(temp))<4 ){ ### unique value가 4개 미만이면 factor  
 temp %>% table %>% print  
 if(sum(unique(temp)%in% 0:1)==2){ ### unique value가 0,1 이면  
 rtd.df[,i] = factor(temp, labels = c("No","Yes")) ### No, Yes 로 자동변환  
 } else{  
 rtd.df[,i] = factor(temp) #### 0,1 이 아닌 경우에는 그냥 factor로  
 }  
 }  
 hline(52) %>% cat  
}

pid  
2982 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 1.0 753.2 1504.5 1505.0 2254.8 3007.0   
  
----------------------------------------------------  
year  
16 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 1978 1986 1988 1988 1990 1993   
  
----------------------------------------------------  
age  
67 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 24.00 45.00 54.00 55.06 65.00 90.00   
  
----------------------------------------------------  
meno  
2 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.00 0.00 1.00 0.56 1.00 1.00   
.  
 0 1   
1312 1670   
  
----------------------------------------------------  
size  
3 cases   
 <=20 20-50 >50   
 1387 1291 304   
.  
 <=20 20-50 >50   
 1387 1291 304   
  
----------------------------------------------------  
grade  
2 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 2.000 2.000 3.000 2.734 3.000 3.000   
.  
 2 3   
 794 2188   
  
----------------------------------------------------  
nodes  
28 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.000 0.000 1.000 2.712 4.000 34.000   
  
----------------------------------------------------  
pgr  
611 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0 4.0 41.0 161.8 198.0 5004.0   
  
----------------------------------------------------  
er  
642 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0 11.0 61.0 166.6 202.8 3275.0   
  
----------------------------------------------------  
hormon  
2 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0000 0.0000 0.0000 0.1137 0.0000 1.0000   
.  
 0 1   
2643 339   
  
----------------------------------------------------  
chemo  
2 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0000 0.0000 0.0000 0.1945 0.0000 1.0000   
.  
 0 1   
2402 580   
  
----------------------------------------------------  
rtime  
2182 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 36.0 823.5 1940.0 2097.9 3198.8 7043.0   
  
----------------------------------------------------  
recur  
2 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0000 0.0000 1.0000 0.5091 1.0000 1.0000   
.  
 0 1   
1464 1518   
  
----------------------------------------------------  
dtime  
2215 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 36 1607 2638 2605 3555 7043   
  
----------------------------------------------------  
death  
2 cases   
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0000 0.0000 0.0000 0.4266 1.0000 1.0000   
.  
 0 1   
1710 1272   
  
----------------------------------------------------

rtd.df %>% str

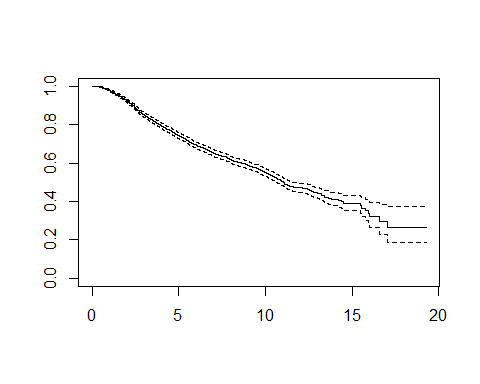
'data.frame': 2982 obs. of 15 variables:  
 $ pid : int 1 2 3 4 5 6 7 8 9 10 ...  
 $ year : int 1992 1984 1983 1985 1983 1983 1993 1988 1988 1988 ...  
 $ age : int 74 79 44 70 75 52 40 53 60 52 ...  
 $ meno : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 1 1 2 2 1 ...  
 $ size : Factor w/ 3 levels "<=20","20-50",..: 1 2 1 2 1 1 1 1 1 2 ...  
 $ grade : Factor w/ 2 levels "2","3": 2 2 1 2 2 2 1 1 1 2 ...  
 $ nodes : int 0 0 0 0 0 0 0 0 0 5 ...  
 $ pgr : int 35 36 138 0 260 139 13 1 627 316 ...  
 $ er : int 291 611 0 12 409 303 4 4 151 976 ...  
 $ hormon: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
 $ chemo : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
 $ rtime : num 1799 2828 6012 2624 4915 ...  
 $ recur : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
 $ dtime : num 1799 2828 6012 2624 4915 ...  
 $ death : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...

#### Surv 객체 만들기

Surv 함수로 survival 객체를 만들어 rtd.S에 주세요. rtd.S의 생존곡선을 그려주세요.

years <- dtime/365.2422  
units(years) <- "Year"

위와 같이 생존기간 단위를 year로 변경해 주세요



Surv function

Surv(time, time2, event, …)

time : right censored follow up time

event : death =1, alive=0 -> 주의 ! No=1, Yes=2 이렇게 factor로 되어 있는 경우에는 event = death 같은 방법으로 assign 하면 1=death, 2=left censored로 오인됨.

이걸 해결하려면 death==“Yes” 값으로 지정하면 됩니다.

**rms package 사용시에는 미리 attach를 해줘야 합니다.**

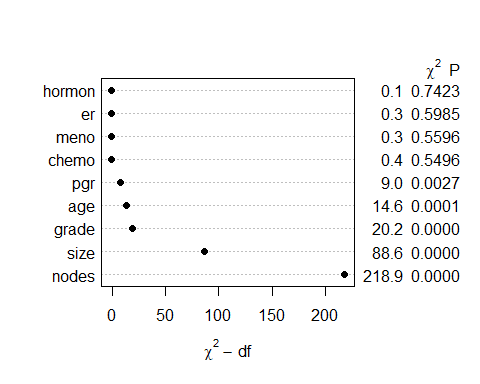
#### psm fit 모델 만들기

모델을 만들기 전에 datadist 함수와 datadist option을 지정해 주세요

age meno size grade nodes pgr er hormon chemo  
Low:effect 45.00000 <NA> <NA> <NA> 0 4.000 11.000 <NA> <NA>  
Adjust to 54.00000 Yes <=20 3 1 41.000 61.000 No No  
High:effect 65.00000 <NA> <NA> <NA> 4 198.000 202.750 <NA> <NA>  
Low:prediction 28.00000 No <=20 2 0 0.000 0.000 No No  
High:prediction 84.00335 Yes >50 3 24 1757.064 1700.131 Yes Yes  
Low 24.00000 No <=20 2 0 0.000 0.000 No No  
High 90.00000 Yes >50 3 34 5004.000 3275.000 Yes Yes  
  
Values:  
  
meno : No Yes   
size : <=20 20-50 >50   
grade : 2 3   
hormon : No Yes   
chemo : No Yes

age,meno,size,grade,nodes,pgr,er,hormon,chemo 변수들을 이용한 psm model을 만들어 주세요

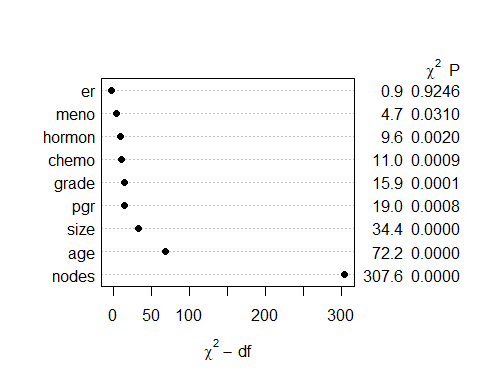
rtdfit.psm 이라는 변수에 모델을 넣고 anova plot을 그려 주세요



#### 연속변수들 변형

연속변수들에는 rcs(x,5) 와 같은 transform 함수를 적용하여 psm model 만들어서

rtdfit.psm2 라는 변수에 모델을 넣고 anova plot을 그려 주세요

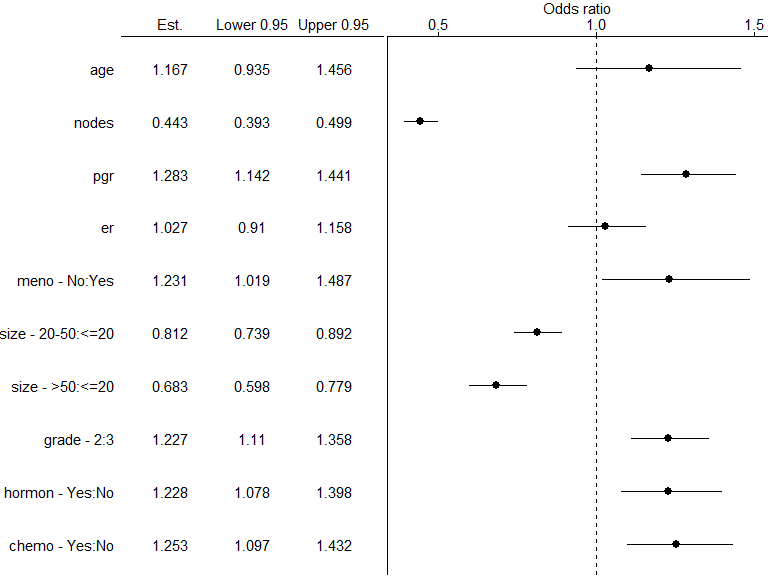


#### forest plot

위의 모델에 대한 forest plot을 그려 주세요

rms package의 psm class는 ormPlot의 forestplot 함수로만 그릴 수 있습니다.

library(ormPlot)



#### validate를 해주세요

rtdfit.psm2 모델에 대해서 bootstrap method로 200번 반복하여 검증해 주세요

validate 함수 결과를 rtdfpsm.val에 넣고 결과값을 보여주세요

rtdfpsm.val <- validate(rtdfit.psm2, method = "boot", B=200)  
rtdfpsm.val

index.orig training test optimism index.corrected n  
Dxy 0.4448 0.4472 0.4390 0.0082 0.4366 200  
R2 0.2318 0.2362 0.2256 0.0107 0.2211 200  
Intercept 0.0000 0.0000 0.0703 -0.0703 0.0703 200  
Slope 1.0000 1.0000 0.9728 0.0272 0.9728 200  
D 0.0778 0.0797 0.0755 0.0042 0.0736 200  
U -0.0002 -0.0002 -0.0004 0.0001 -0.0004 200  
Q 0.0780 0.0799 0.0758 0.0041 0.0740 200  
g 0.6233 0.6335 0.6155 0.0180 0.6053 200

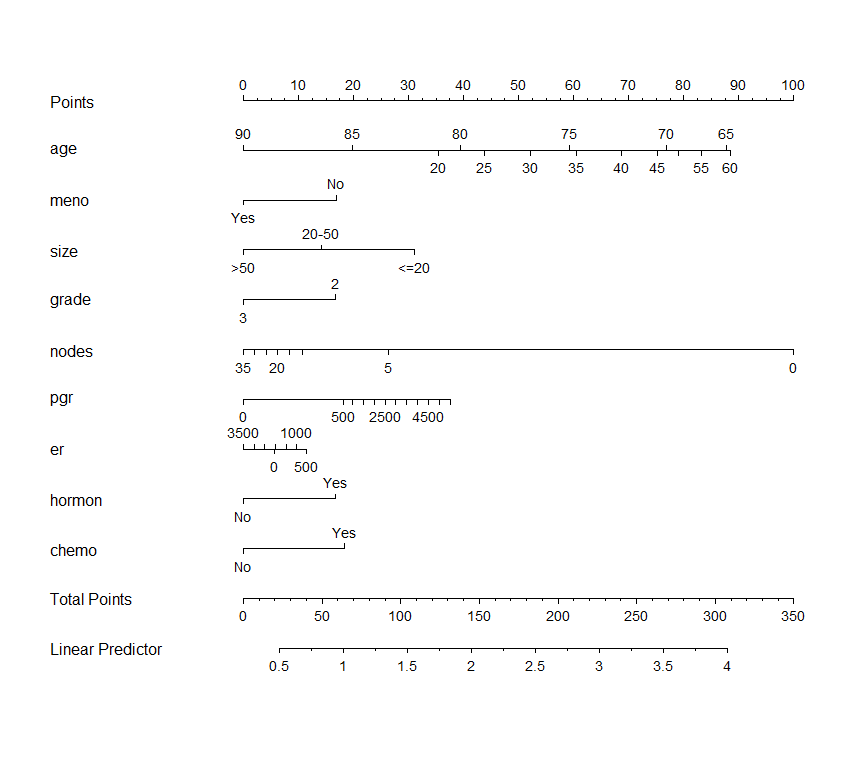
#### C-index를 구해주세요

corrected 를 이용하여 계산해 주세요

[1] 0.7182917

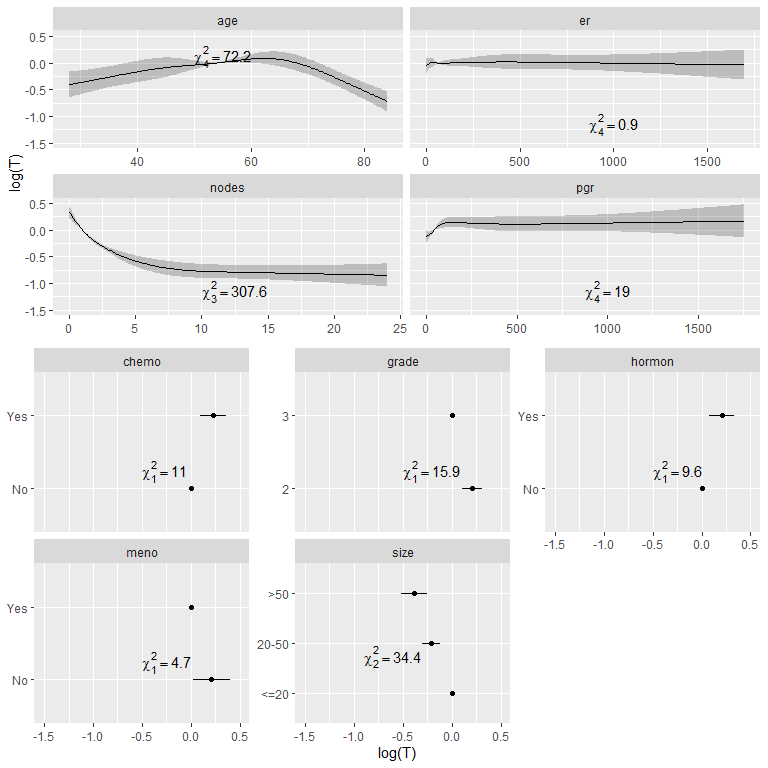
#### Nomogram 그리기

rtdfit.psm2 모델에 대해서 nomogram을 구해주세요



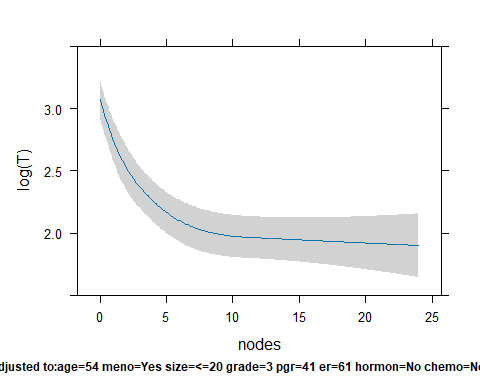
#### Predict 그래프

ggplot 함수를 이용해서 rtdfit.psm2 각 변수별 Odds ratio 그래프를 그려주세요



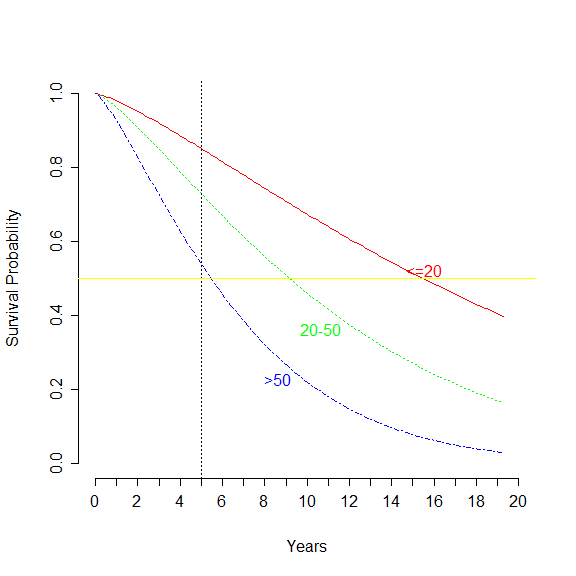
rtdfit.psm2 에서 nodes에 대한 predict 그래프를 그려주세요

plot(Predict(rtdfit.psm2,nodes))



#### survival curve

단일 변수 size에 대한 psm 모델을 만들어서 생존곡선을 그려 주세요



## 1.3 두개의 변수들의 상관관계 살펴보기

다음과 같은 crosstab table을 만들어 보세요 dcast(Nsatge~Tstage)

| Nstage | 1a | 1b | 2a | 2b | 3 | 4 | is |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 126 | 110 | 326 | 87 | 134 | 36 | 3 |
| 1 | 16 | 26 | 147 | 45 | 62 | 38 | 0 |
| 2 | 21 | 39 | 169 | 51 | 106 | 41 | 0 |
| 3 | 0 | 3 | 11 | 0 | 8 | 1 | 0 |

## 1.4 Factor 다루기

### 1.4.1 Factor 만들기

factor 함수를 이용하여 Tstage, Nstage, Mstage를 factor 변수로 만들어 보시오.

다음과 같은 결과를 만들어주세요

df$Tstage %>% table

.  
 1a 1b 2a 2b 3 4 is   
163 178 653 183 310 116 3

df$Nstage %>% table

.  
 0 1 2 3   
822 334 427 23

df$Mstage %>% table

.  
 0 1a 1b   
1553 34 19

위의 결과를 보면 Tstage의 level은 1a 1b 2a 2b 3 4 is의 순서입니다. 이것을 is 1a 1b 2a 2b 3 4의 순서로 바꾸시오.

다음과 같은 결과를 만들어주세요

df$Tstage %>% table

.  
 is 1a 1b 2a 2b 3 4   
 3 163 178 653 183 310 116

| Nstage | is | 1a | 1b | 2a | 2b | 3 | 4 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 3 | 126 | 110 | 326 | 87 | 134 | 36 |
| 1 | 0 | 16 | 26 | 147 | 45 | 62 | 38 |
| 2 | 0 | 21 | 39 | 169 | 51 | 106 | 41 |
| 3 | 0 | 0 | 3 | 11 | 0 | 8 | 1 |

is 1a 1b 2a 2b 3 4 형태의 label을 cis 1a 1b 2a 2b 3 4 로 바꿔주세요

다음과 같은 결과를 만들어주세요

df$Tstage %>% table

.  
cis 1a 1b 2a 2b 3 4   
 3 163 178 653 183 310 116

두변수의 Cramer’s V 분석을 해주세요

| Nstage | cis | 1a | 1b | 2a | 2b | 3 | 4 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 3 | 126 | 110 | 326 | 87 | 134 | 36 |
| 1 | 0 | 16 | 26 | 147 | 45 | 62 | 38 |
| 2 | 0 | 21 | 39 | 169 | 51 | 106 | 41 |
| 3 | 0 | 0 | 3 | 11 | 0 | 8 | 1 |

Cramer V   
 0.1419

CT\_size와 Tstage, Nstage의 ANOVA test

aov(CT\_size ~ Nstage, data=df) %>% summary

Df Sum Sq Mean Sq F value Pr(>F)   
Nstage 3 144 48.10 9.26 4.51e-06 \*\*\*  
Residuals 1602 8321 5.19   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

boxplot(CT\_size ~ Tstage, data=df, col=rainbow(9))

