EDA 7장 과제

주의사항을 숙지하였고 모든 책임을 지겠습니다.

2019122041 송유진

---

1. 아래 자료를 분석하여라.

The data are part of a larger experiments to determine the effectiveness of blast furnace slags as agricultural liming materials on three types of soil, sandy loams(I), sandy clay loam(II), and loamy sand(III). The treatments were all applied at 4000 lbs per acre, and what was measured was the corn yield in bushels per acre.

Treatment	I	II	III
None	11.1	32.6	63.3
Coarse slag	15.3	40.8	65.0
Mediaum slag	22.7	52.1	58.8
Agricultural slag	23.8	52.8	61.4
Agricultural limestone	25.6	63.1	41.1
Agricultural slag + minor limestone	31.2	59.5	78.1
Agricultural limestone + minor elements	25.8	55.3	60.2

```
effect_data <- rbind(c(11.1,32.6,63.3),
                     c(15.3,40.8,65),
                     c(22.7,52.1,58.8),
                     c(23.8,52.8,61.4),
                     c(25.6,63.1,41.1),
                     c(31.2,59.5,78.1),
                     c(25.8,55.3,60.2))

dimnames(effect_data) <- list(c("None", "Coarse slag", "Mediaum slag", "Agricultural slag", "Ag
ricultural limestone","Agricultural slag + minor limestone", "Agricultural limestone + minor
elements"),c("sandy loams", "sandy clay loam","loamy sand"))

effect_data
```

문제에서 제시한 자료를 상단의 코드를 통해 생성해주었다. 그 결과는 아래와 같다.

```
                                      sandy loams sandy clay loam loamy sand
None                                         11.1            32.6       63.3
Coarse slag                                  15.3            40.8       65.0
Mediaum slag                                 22.7            52.1       58.8
Agricultural slag                            23.8            52.8       61.4
Agricultural limestone                       25.6            63.1       41.1
Agricultural slag + minor limestone          31.2            59.5       78.1
Agricultural limestone + minor elements      25.8            55.3       60.2
```

```
> knitr::kable(effect_data, align = "lccrr")
```

```
|                                      |sandy loams | sandy clay loam | loamy sand |
|:-------------------------------------|:-----------|:---------------:|:----------:|
|None                                  |11.1        |      32.6       |    63.3    |
|Coarse slag                           |15.3        |      40.8       |    65.0    |
|Mediaum slag                          |22.7        |      52.1       |    58.8    |
|Agricultural slag                     |23.8        |      52.8       |    61.4    |
|Agricultural limestone                |25.6        |      63.1       |    41.1    |
|Agricultural slag + minor limestone   |31.2        |      59.5       |    78.1    |
|Agricultural limestone + minor elements |25.8        |      55.3       |    60.2    |
>
```

해당 데이터는 농업용 liming material로서 3가지 종류의 토양에 blast furnace slags의 효과를 측정한 것이다. 효과는 각 토양의 acre 당 옥수수 생산량을 통해 살펴볼 수 있다.

slag : 원하는 금속이 원석에서 분리된 후 남은 유리 같은 부산물
liming : 칼슘, 마그네슘 등이 풍부한 물질들을 marl, chalk, limestone 등의 다양한 형태로 토양에 적용하는 것
적절한 liming은 식물의 성장이나 토양의 박테리아의 활동성을 증가

위의 표에서 None은 slag를 사용하지 않았을 때를 뜻한다.
또한, Coarse slag, Mediaum slag, Agricultural slag 등 다양한 slag들을 사용했을 때의 옥수수의 생산량을 통해 해당 slag이 plant growth에 어떤 영향을 미치고 있는지 확인할 수 있다.
토양의 종류 또한 sandy loams(모래비옥토), sandy clay loam(sandy loams보다 입자가 고운 토양), loamy sand(모래의 양이 가장 많은 토양)으로 나누어 옥수수 생산에 미치는 영향을 분석할 수 있다.

summary(effect_data)

```
> summary(effect_data)
  sandy loams     sandy clay loam   loamy sand
 Min.   :11.10   Min.   :32.60   Min.   :41.10
 1st Qu.:19.00   1st Qu.:46.45   1st Qu.:59.50
 Median :23.80   Median :52.80   Median :61.40
 Mean   :22.21   Mean   :50.89   Mean   :61.13
 3rd Qu.:25.70   3rd Qu.:57.40   3rd Qu.:64.15
 Max.   :31.20   Max.   :63.10   Max.   :78.10
```

해당 데이터를 summary() 함수를 통해 토양 관점에서 우선적으로 살펴보았다.

[sandy loams]
최솟값이 11.1, 최댓값이 31.2, median은 23.8
[sandy clay loam]
최솟값은 32.6, 최댓값은 63.1, median은 52.8
[loamy sand]
최솟값은 41.1, 최댓값은 78.1, median은 61.4

sandy -> sandy clay -> loamy로 갈수록 옥수수 생산량이 증가한다는 사실을 알 수 있다.

summary(t(effect_data))

```
> summary(t(effect_data))
      None          Coarse slag      Mediaum slag    Agricultural slag
 Min.   :11.10   Min.   :15.30   Min.   :22.70   Min.   :23.8
 1st Qu.:21.85   1st Qu.:28.05   1st Qu.:37.40   1st Qu.:38.3
 Median :32.60   Median :40.80   Median :52.10   Median :52.8
 Mean   :35.67   Mean   :40.37   Mean   :44.53   Mean   :46.0
 3rd Qu.:47.95   3rd Qu.:52.90   3rd Qu.:55.45   3rd Qu.:57.1
 Max.   :63.30   Max.   :65.00   Max.   :58.80   Max.   :61.4
 Agricultural limestone Agricultural slag + minor limestone
 Min.   :25.60          Min.   :31.20
 1st Qu.:33.35          1st Qu.:45.35
 Median :41.10          Median :59.50
 Mean   :43.27          Mean   :56.27
 3rd Qu.:52.10          3rd Qu.:68.80
 Max.   :63.10          Max.   :78.10
 Agricultural limestone + minor elements
 Min.   :25.80
 1st Qu.:40.55
 Median :55.30
 Mean   :47.10
 3rd Qu.:57.75
 Max.   :60.20
```

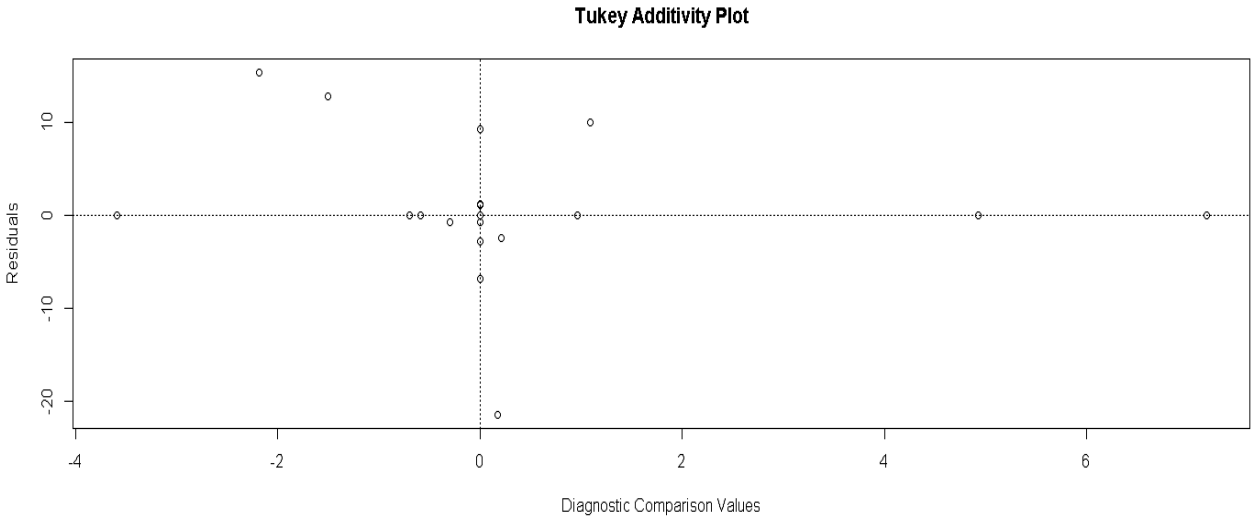
이번에는 t()를 적용하여 slag의 종류에 따라 옥수수 생산량을 살펴보았다.
None 그룹과 Coarse slag 그룹의 경우 다른 그룹에 비해 생산량이 현저하게 낮음을 확인할 수 있었다. 또한, Agricultural slag + minor limestone을 사용한 그룹의 경우 다른 그룹에 비해 생산량이 눈에 띄게 높다는 사실을 알 수 있었다.

```
> medpolish(effect_data)
1: 109.3
2: 84.6
Final: 84.6

Median Polish Results (Dataset: "effect_data")

Overall: 52.8

Row Effects:
                                  None                          Coarse slag                          Mediaum slag
                                 -13.4                                -9.2                                -1.8
                    Agricultural slag                Agricultural limestone     Agricultural slag + minor limestone
                                   0.0                                 1.1                                 6.7
Agricultural limestone + minor elements
                                   1.3

Column Effects:
    sandy loams sandy clay loam      loamy sand
          -28.3             0.0             8.6

Residuals:
                                     sandy loams sandy clay loam loamy sand
None                                         0.0            -6.8       15.3
Coarse slag                                  0.0            -2.8       12.8
Mediaum slag                                 0.0             1.1       -0.8
Agricultural slag                           -0.7             0.0        0.0
Agricultural limestone                       0.0             9.2      -21.4
Agricultural slag + minor limestone          0.0             0.0       10.0
Agricultural limestone + minor elements      0.0             1.2       -2.5
```

(row effects & col effects)

```
> med.e <-  medpolish(effect_data)
1: 109.3
2: 84.6
Final: 84.6
```

plot(med.e)



해당 데이터가 additive model에 적합한지 알아보기 위하여 비교값 대 잔차의 산점도를 그려보았다. 만약 scatter plot에 경향선이 나타난다면, 이는 additive model이 적합하지 않다는 것을 의미하고 경향선이 없으면 이는 additive model에 적합함을 의미한다.
이때 비교값은 row * col/all로 정의되는데, 이 비교값과 잔차가 의미있는 관계에 있다면 additive model은 부적절하다. 경향선의 slope 값 k는 re-expression값 p와 p=1-k 관계에 있다. 경향성의 기울기가 1이라면(승법적 모형), p=0으로, log expression이 additive model을 만든다. 따라서 x축에는 residual 값을 그리고, y축에는 비교값은 그려 경향선이 나타나는지, 나타나지 않는지를 살펴보면 된다.
해당 데이터에서는 점들의 경향을 찾을 수는 없다. 따라서, effect data는 additive model에 적합하다는 결론을 내릴 수 있다.

하단의 additive model을 통해 각 data를 설명할 수 있다.
data = all + row effect + column effect + residual = fit + residual
data effect는 additive 모델에 적합하기 때문에 위와 같이 표현할 수 있는데, row effect는 토양의 종류에 따른 effect에 관한 것이고, column effect는 slag의 유무, 종류에 따른 effect에 관한 것이다.

effect_data

	sandy loams	sandy clay loam	loamy sand
None	11.1	32.6	63.3
Coarse slag	15.3	40.8	65.0
Mediaum slag	22.7	52.1	58.8
Agricultural slag	23.8	52.8	61.4
Agricultural limestone	25.6	63.1	41.1
Agricultural slag + minor limestone	31.2	59.5	78.1
Agricultural limestone + minor elements	25.8	55.3	60.2

med.e$overall + outer(med.e$row, med.e$col, '+') + med.e$residuals

	sandy loams	sandy clay loam
None	11.1	32.6
Coarse slag	15.3	40.8
Mediaum slag	22.7	52.1
Agricultural slag	23.8	52.8
Agricultural limestone	25.6	63.1
Agricultural slag + minor limestone	31.2	59.5
Agricultural limestone + minor elements	25.8	55.3

	loamy sand
None	63.3
Coarse slag	65.0
Mediaum slag	58.8
Agricultural slag	61.4
Agricultural limestone	41.1
Agricultural slag + minor limestone	78.1
Agricultural limestone + minor elements	60.2

상단의 코드 및 출력결과를 통해 data = all + row effect + column effect + residual = fit + residual로 표현한 값과 실제 데이터의 값이 일치함을 확인할 수 있다.

```
> column <- matrix(c(-13.4, -9.2, -1.8, 0.0, 1.1, 6.7, 1.3), nrow=7, ncol=1)
> dimnames(column) <- list(c("None", "Coarse slag", "Mediaum slag", "Agricultural slag", "Agricultural limestone","Agricultural slag + minor limestone", "Agricultural limestone + minor elements"))
> knitr::kable(column, align = "lccrr",caption = "row effects")
```

Table: row effects

None	-13.4
Coarse slag	-9.2
Mediaum slag	-1.8
Agricultural slag	0.0
Agricultural limestone	1.1
Agricultural slag + minor limestone	6.7
Agricultural limestone + minor elements	1.3

[row effect / column effect]

3페이지 (row effects & col effects) 결과 참조

col effect는 -28.3, 0.0, 8.6으로, 토양의 차이에 따라 옥수수의 생산량에도 차이가 있다는 것을 알 수 있다. sandy loams 토양에는 비교적 옥수수 생산량이 제일 적고, loamy sand 토양에는 옥수수 생산량이 비교적 많다는 것을 확인할 수 있다. sandy loams의 col effect 값은 -28.3인데 이를 통해 옥수수 생산이 다른 두 토양에 비해 매우 적다는 사실을 알 수 있다.

sandy loams -> sandy clay loam -> loamy sand로 갈수록 옥수수 생산량이 많아지고 있다.

다음으로, 해당 데이터의 residual을 살펴보았다.

```
residual_matrix <- matrix(c(0.0,0.0,0.0,-0.7,
                            0.0,0.0,0.0,-6.8,
                            -2.8, 1.1,0.0,9.2,0.0,
                            1.2,15.3,12.8,-0.8,0.0,
                            -21.4, 10.0, -2.5), ncol=3, nrow=7)

dimnames(residual_matrix) <- list(c("None", "Coarse slag", "Mediaum slag",
"Agricultural slag","Agricultural limestone","Agricultural slag + minor
limestone", "Agricultural limestone + minor elements"),c("sandy loams",
"sandy clay loam", "loamy sand"))
```

```
knitr::kable(residual_matrix, align = "lccrr", caption= "residual")
```

Table: residual

	sandy loams	sandy clay loam	loamy sand
None	0.0	-6.8	15.3
Coarse slag	0.0	-2.8	12.8
Mediaum slag	0.0	1.1	-0.8
Agricultural slag	-0.7	0.0	0.0
Agricultural limestone	0.0	9.2	-21.4
Agricultural slag + minor limestone	0.0	0.0	10.0
Agricultural limestone + minor elements	0.0	1.2	-2.5

[None]
다른 slag들에 비해 sandy clay loam에서의 생산량이 비교적 적다. raw data의 값도 32.6으로, 다른 그룹들과 비교해 보았을 때 적은 수치임을 알 수 있다. sandy loams -> sandy clay loam 증가 정도 또한 20 정도인데 이는 다른 그룹들과 비교했을 때 현저하게 낮은 수치임을 알 수 있다. 따라서, -6.8이라는 작은 residual가 도출되었음을 알 수 있다.
그러나 loamy sand의 경우 residual 값이 15.3으로 sandy loams와 sandy clay loam과 비교했을 때 생산량이 급격히 많아졌으며 다른 group들과 비교했을 때도 loamy sand에서의 생산량은 차이가 많지 않음을 알 수 있다.

[Coarse slag]
None의 경우와 비슷한 양상을 보이고 있다. sandy loams, sandy clay loam에 비해 loamy sand의 경우 12.8이라는 큰 residual 값이 나왔음을 알 수 있다.

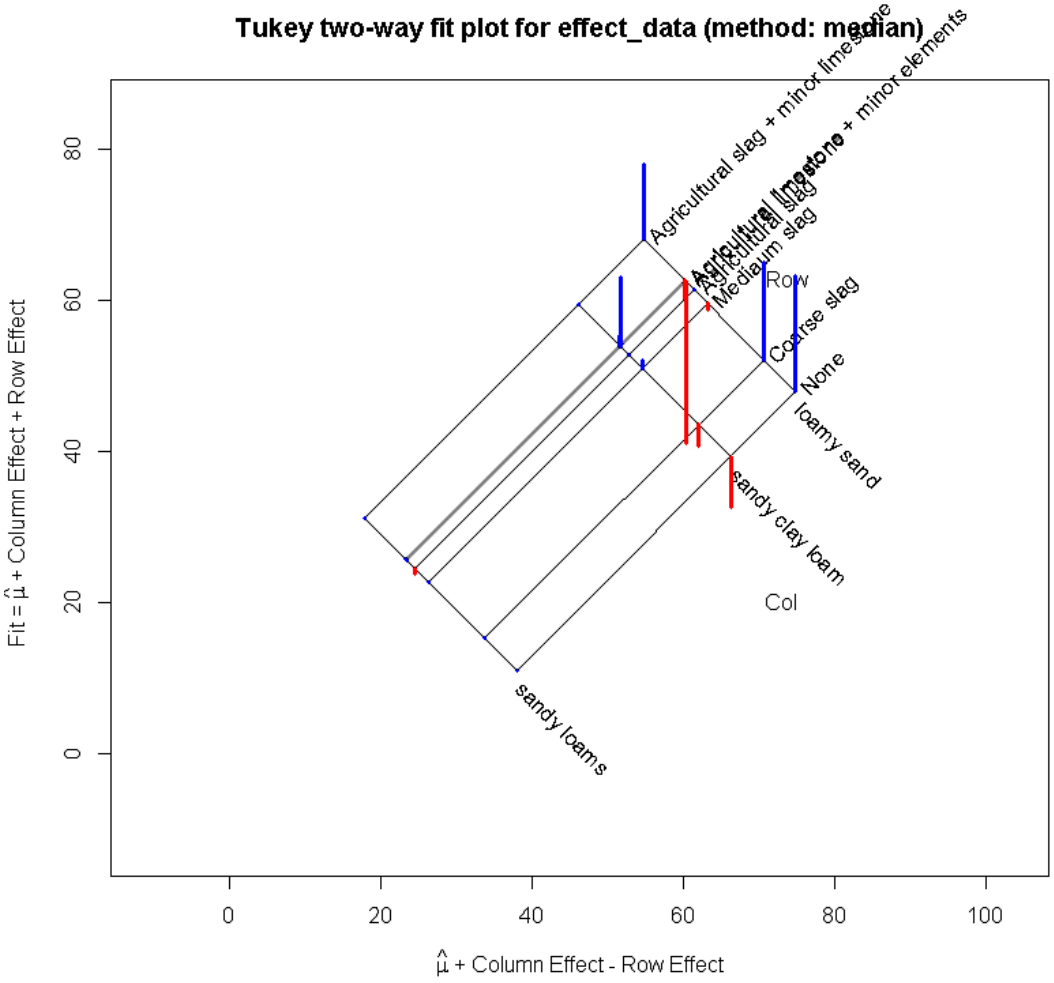
[Mediaum slag]
fitting이 비교적 잘 되었기 때문에 눈에 띄는 residual은 보이지 않았다.

[Agricultural limestone]
다른 group에서의 옥수수 생산량은 sandy loams -> sandy clay loam -> loamy sand 토양으로 갈수록 점차 증가하는 양상을 보이지만, 이 경우는 sandy clay loam 토양에서 가장 많은 생산량을 보였다. 이로 인해 andy clay loam 토양에서의 residual 값이 9.2로 큰 값이 나왔고 loamy sand 토양에서의 residual이 -21.4로 다른 그룹들과 비교했을 때 매우 작은 값이 도출되었다.

[Agricultural slag + minor limestone]
loamy sand 토양에서 생산량은 78.1이다. Agricultural slag + minor limestone를 loamy sand 토양에서 사용했을 때, 가장 많은 옥수수 생산량을 기록했다. raw data에서도 loamy sand 토양에서의 생산량이 비교적 많았는데 이로 인해 10.0이라는 residual 값이 도출되었음을 유추해볼 수 있다.

[Agricultural limestone + minor elements]
다른 그룹과 비교했을 때 눈에 띄는 차이는 발견되지 않았다.

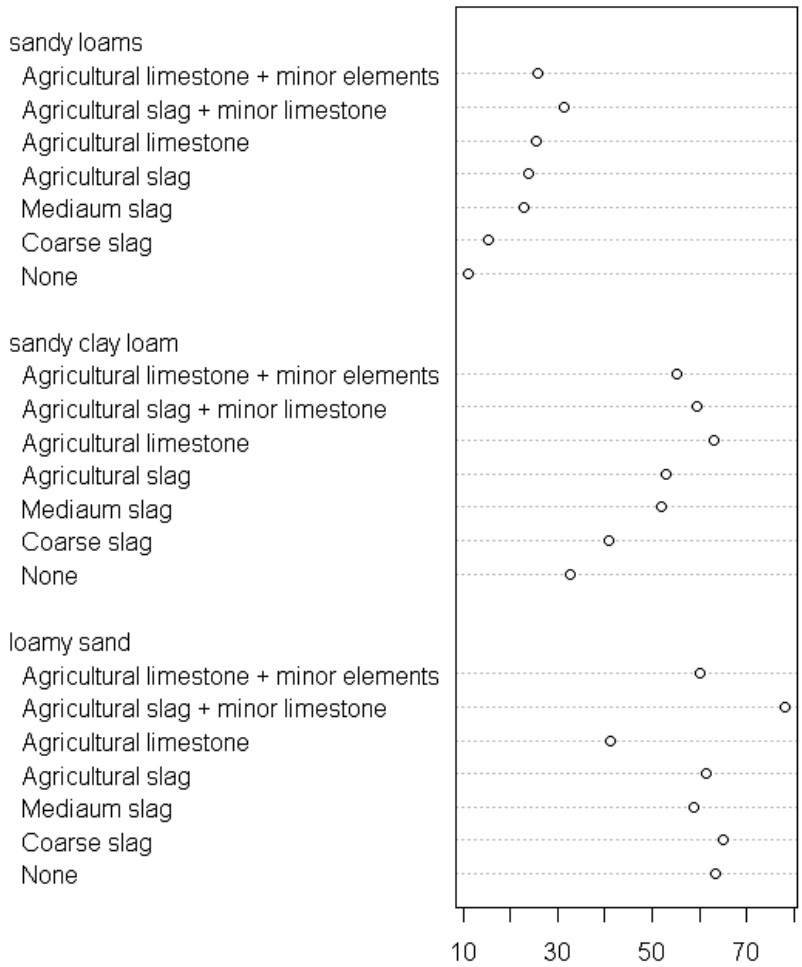
```
install.packages("twoway")
library("twoway")
plot(twoway(effect_data,method = "median"))
```



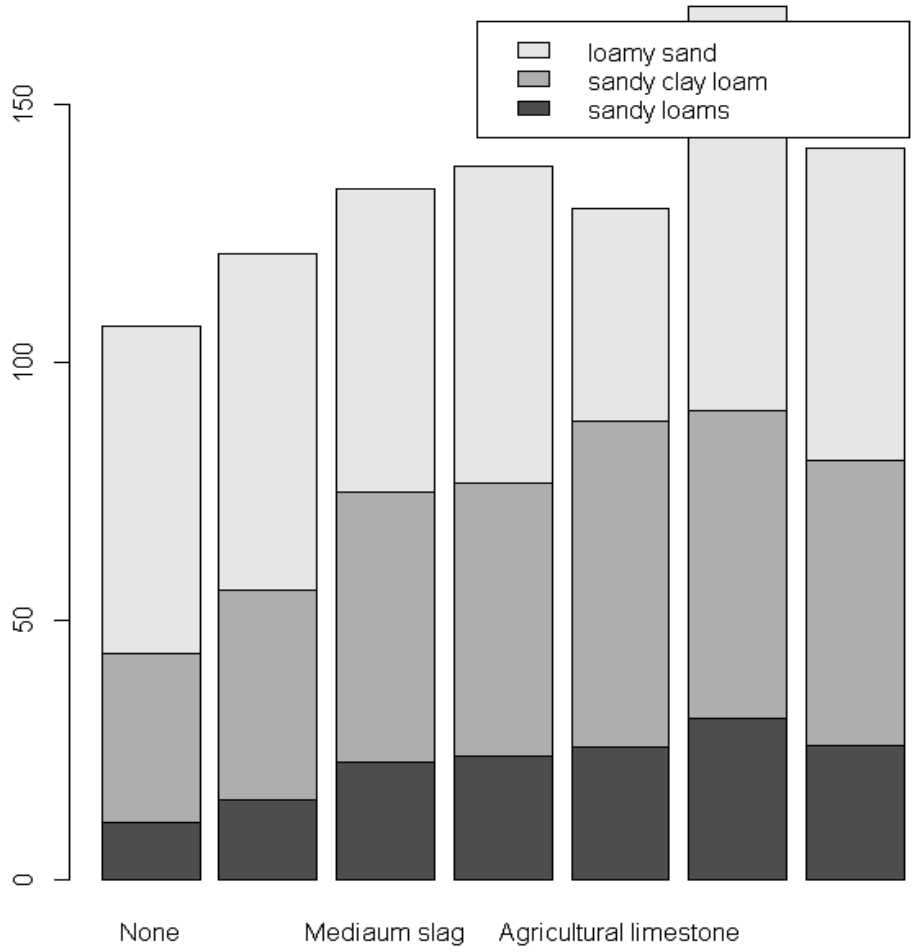
다음으로, two way plot을 그려보았다.
two way plot의 y축은 all+row effect+col effect로서 residual을 제외한 fitting된 부분의 값을 나타내고, residual은 양이면 파란색 선, 음이면 빨간색 선으로 표시해준다.
two way plot을 통해 토양의 종류 그리고 slag의 종류에 따라 생산량의 차이가 크다는 사실을 알 수 있었다. 가장 많은 생산량을 기록한 것은 Agricultural slag + minor limestone였고, 가장 적은 생산량을 기록한 것은 None이었다.

그러나 None의 경우 loamy sand 토양에서 파란색 양의 residual을 가졌고 비교적 높은 생산량을 보였다. 또한, Agricultural limestone의 경우 loamy sand에서 빨간색으로 표시된 큰 음의 residual을 보인다. 즉, 해당 slag가 해당 토양에서 덜 effective하다는 것을 알 수 있다.
결론적으로, sandy loams -> sandy clay loam -> loamy sand으로 갈수록 옥수수 생산량이 증가한다는 사실을 알 수 있다.

dotchart(effect_data)

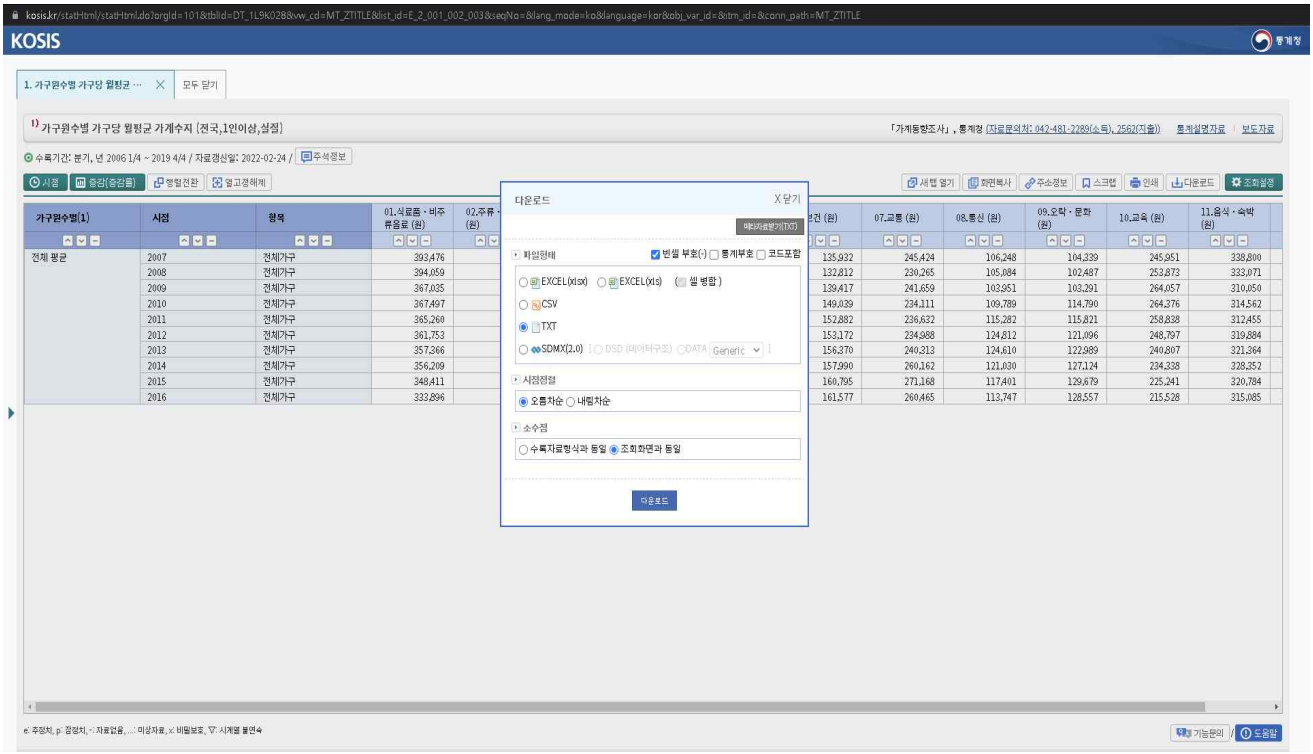


```
barplot(t(effect_data), legend=colnames(effect_data))
```



dot chart와 bar chart를 추가적으로 그려봄으로써 데이터를 살펴보았다. 이전 분석들에서 얻을 수 있었던 경향성 및 인사이트를 해당 chart를 통해서도 확인할 수 있었으며 더욱 직관적인 시각화 결과를 얻을 수 있었다.

2. 교과서 7장 가구 소비지출에 대한 통계청에서 얻을 수 있는 최근 10년간 자료로 2원 분석을 하여라.




```
> ##KOSIS
> setwd("D://2022-1(3-2)//2022-01_탐자분//Data//")
> consume=read.table("consume.txt", header=TRUE)
> consume
     식료품비주류음료 주류담배 의류신발 주거수도광열 가정용품가사서비스   보건   교통   통신 오락문화   교육 음식숙박
2007           393476    43534   155344       242357             87757 135932 245424 106248   104339 245951   338800
2008           394059    42705   152121       246695             81363 132812 230265 105084   102487 253873   333071
2009           367035    40176   146420       247254             80332 139417 241659 103951   103291 264057   310050
2010           367497    41404   157448       260929             91380 149039 234111 109789   114790 264376   314562
2011           365260    41205   163141       263807             91498 152882 236632 115282   115821 258838   312455
2012           361753    40629   165306       265292             93035 153172 234988 124812   121096 248797   319884
2013           357366    39583   161715       265619             96491 156370 240313 124610   122989 240807   321364
2014           356209    38966   153840       257982             99168 157990 260162 121030   127124 234338   328352
2015           348411    30256   143780       267780             95360 160795 271168 117401   129679 225241   320784
2016           333896    31419   137742       266894             97539 161577 260465 113747   128557 215528   315085
     기타상품서비스
2007         225220
2008         206207
2009         198058
2010         209434
2011         219584
2012         224248
2013         207601
2014         212505
2015         205959
2016         199537
```

KOSIS 국가 통계 포털을 통해 가구원수별 가구당 월평균 가계수지 (전국, 1인이상) 데이터를 추출하여 분석에 활용했다. 2017년부터 소비지출 부문 기준이 변경되어 2007년부터 2016년까지의 데이터를 이용하였다.

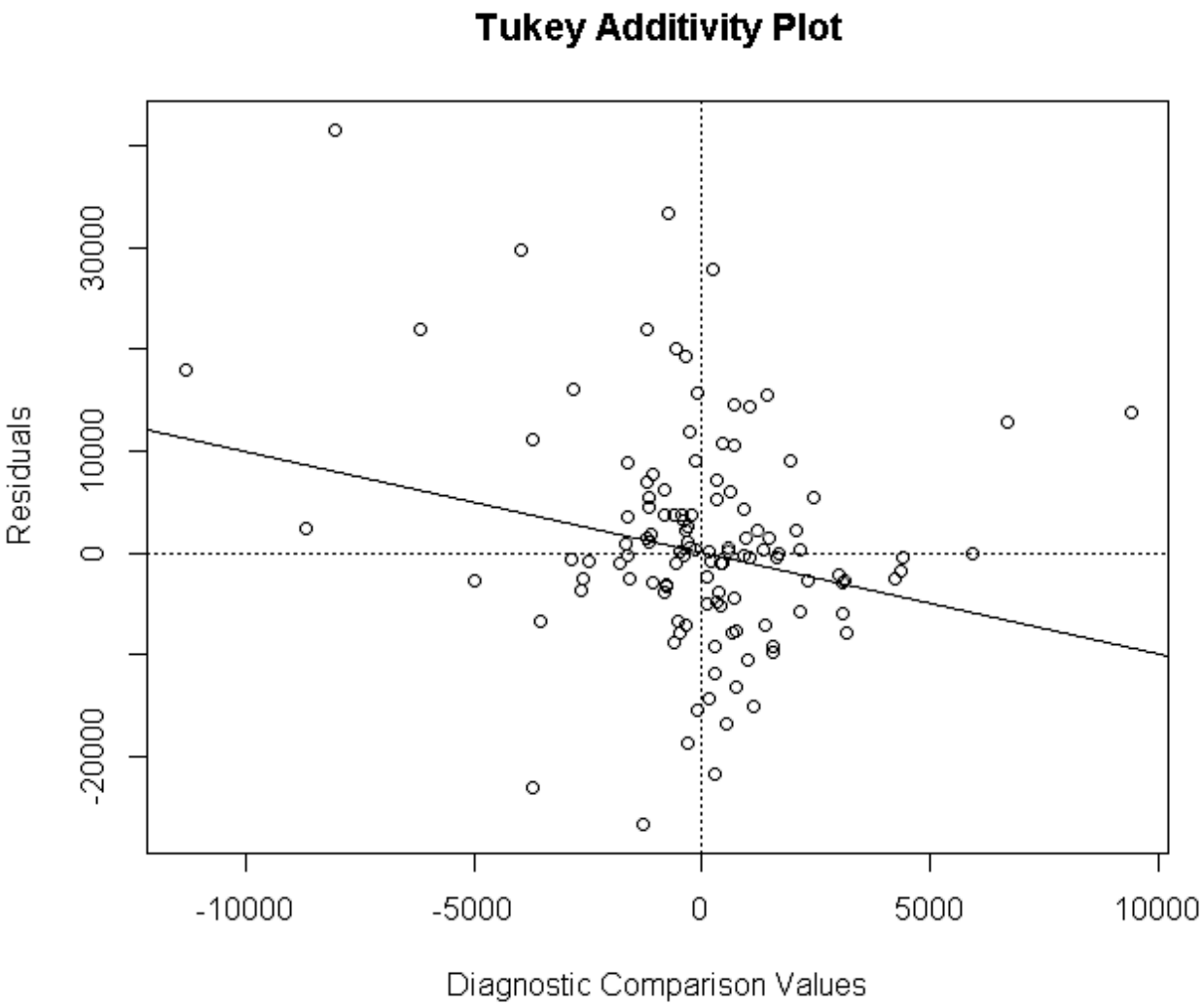
```
> summary(consume)
 식료품비주류음료     주류담배          의류신발         주거수도광열     가정용품가사서비스       보건             교통
 Min.   :333896   Min.   :30256   Min.   :137742   Min.   :242357   Min.   :80332   Min.   :132812   Min.   :230265
 1st Qu.:356498   1st Qu.:39120   1st Qu.:147845   1st Qu.:249936   1st Qu.:88663   1st Qu.:141823   1st Qu.:235399
 Median :363507   Median :40403   Median :154592   Median :262368   Median :92267   Median :153027   Median :240986
 Mean   :364496   Mean   :38988   Mean   :153686   Mean   :258461   Mean   :91392   Mean   :149999   Mean   :245519
 3rd Qu.:367382   3rd Qu.:41354   3rd Qu.:160648   3rd Qu.:265537   3rd Qu.:96208   3rd Qu.:157585   3rd Qu.:256478
 Max.   :394059   Max.   :43534   Max.   :165306   Max.   :267780   Max.   :99168   Max.   :161577   Max.   :271168
      통신           오락문화          교육           음식숙박       기타상품서비스
 Min.   :103951   Min.   :102487   Min.   :215528   Min.   :310050   Min.   :198058
 1st Qu.:107133   1st Qu.:106952   1st Qu.:235955   1st Qu.:314693   1st Qu.:206021
 Median :114515   Median :118459   Median :247374   Median :320334   Median :208518
 Mean   :114195   Mean   :117017   Mean   :245181   Mean   :321441   Mean   :210835
 3rd Qu.:120123   3rd Qu.:126090   3rd Qu.:257597   3rd Qu.:326605   3rd Qu.:217814
 Max.   :124812   Max.   :129679   Max.   :264376   Max.   :338800   Max.   :225220
```

summary() 함수를 통해 데이터를 요약해보았다. 식료품비주류음료, 주류담배, 의류신발, 주도수도광열, 가종용품기사서비스, 보건, 교통, 통신, 오락문화, 교육, 음식숙박, 기타상품서비스와 같이 12개 부문으로 나누어져 있음을 확인했다.

다음으로, medpolish() 함수를 통해 모형의 적합 결과를 확인해보았다. 비교값 대 잔차의 산점도 Plot에서 어떠한 의미 있는 패턴이 나타난다면 모형이 적절하지 않기 때문에 변환을 수행해야만 한다.

```
> consume.out=medpolish(consume)
1: 955090
2: 888883
Final: 882134.7
```

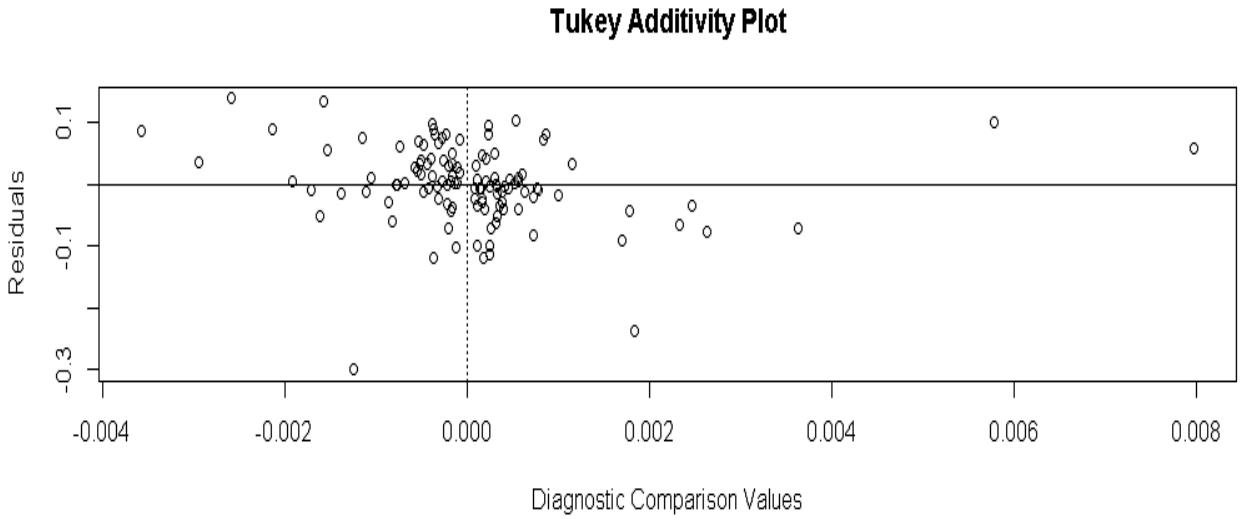
```
attach(consume.out)
plot(consume.out)
abline(0,-1)
```



기울기 1인 직선(abline(0,1))을 그려 확인해본 결과 어느 정도의 선형성을 볼 수 있었다. 따라서, 데이터에 log변환을 수행한 이후 다시 잔차의 산점도를 확인해보았다.

```
> detach(consume.out)
> consume.log.out=medpolish(log(consume))
1: 5.664742
2: 5.45445
Final: 5.418962
```

```
attach(consume.log.out)
plot(consume.log.out)
abline(0,0)
```



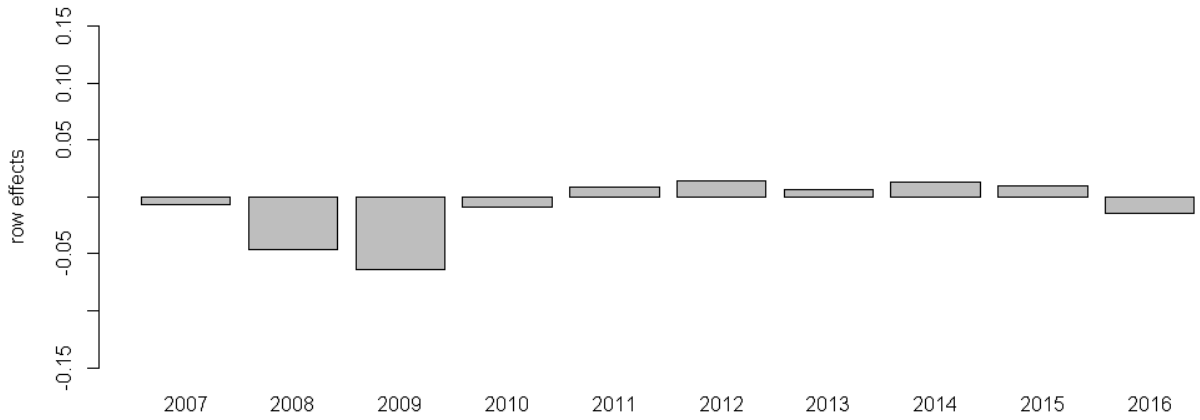
log변환 후 비교 값과 잔차의 산점도을 그려보았다. 위 Plot은 경향선을 찾기 더욱 어려웠기 때문에 변환 전 Plot에 비해 더욱 적합한 모형이라 판단할 수 있었다.

log변환 후 아래의 코드를 통해 결과를 살펴보고자 했으나 추세를 확인하기에는 한계가 있었다.

```
> consume.log.out

Median Polish Results (Dataset: "log(consume)")

Overall: 12.117

Row Effects:
       2007        2008        2009        2010        2011        2012        2013        2014        2015
-0.006926934 -0.046465856 -0.064061775 -0.009221311  0.008418391  0.013691902  0.006926934  0.013005909  0.010018130
       2016
-0.014754018

Column Effects:
   식료품비주류음료              주류담배              의류신발         주거수도광열 가정용품가사서비스                  보건
          0.6754906          -1.5087741          -0.1487970           0.3610646          -0.6874739          -0.1896810
              교통                통신             오락문화                교육            음식숙박      기타상품서비스
          0.2885657          -0.4653930          -0.4416311           0.2982625           0.5573822           0.1438860

Residuals:
     식료품비주류음료    주류담배    의류신발 주거수도광열 가정용품가사서비스          보건        교통        통신   오락문화
2007        0.0972166  0.0800035 -0.0078738    -0.0729656           -0.040267 -0.1004771  0.012109 -0.0711439 -0.113037
2008        0.1382361  0.1003162  0.0106994    -0.0156858           -0.076379 -0.0841583 -0.012109 -0.0426209 -0.091407
2009        0.0847886  0.0568659 -0.0099018     0.0041735           -0.071536 -0.0180276  0.053784 -0.0358654 -0.065997
2010        0.0312061  0.0321331  0.0078738     0.0031652            0.002482 -0.0061295 -0.032789 -0.0360652 -0.015283
2011        0.0074606  0.0096755  0.0257537    -0.0035050           -0.013867  0.0016892 -0.039718 -0.0048839 -0.023981
2012       -0.0074606 -0.0096755  0.0336636    -0.0031652           -0.002482 -0.0016892 -0.051963  0.0692699  0.015283
2013       -0.0128969 -0.0289929  0.0184658     0.0048316            0.040757  0.0257393 -0.022790  0.0744151  0.037559
2014       -0.0222187 -0.0507821 -0.0375356    -0.0304205            0.062044  0.0299670  0.050493  0.0391857  0.064548
2015       -0.0413657 -0.3007893 -0.1021765     0.0098432            0.025875  0.0505534  0.094915  0.0117305  0.087435
2016       -0.0591468 -0.2382989 -0.1203064     0.0313012            0.073240  0.0801770  0.079417  0.0048839  0.103518
            교육     음식숙박 기타상품서비스
2007  0.0045569  0.06571491      0.07087881
2008  0.0757977  0.08819955      0.02222055
2009  0.1327244  0.03417337     -0.00050414
2010  0.0790913 -0.00621948      0.00050414
2011  0.0402817 -0.03057991      0.03019063
2012 -0.0045569 -0.01235545      0.04593485
2013 -0.0304335 -0.00097448     -0.02443479
2014 -0.0637437  0.01445831     -0.00716622
2015 -0.1003495 -0.00587211     -0.03546684
2016 -0.1196575  0.00097448     -0.04237213
```

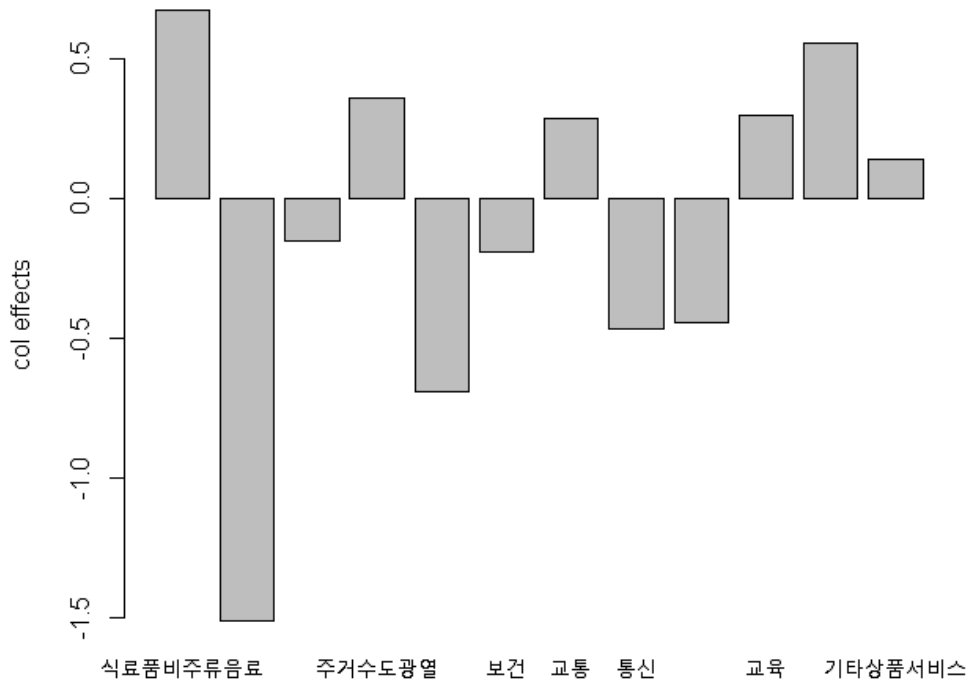
따라서, 아래의 코드를 통해 plot을 그려보았다.

```
barplot(consume.log.out$row,ylim=c(-0.15,0.15),ylab="row effects")
```



2007년 이후 2008년, 2009년에 행효과가 눈에 띄게 감소했지만 그 이후로는 행효과가 어느정도 증가했음을 알 수 있었다.

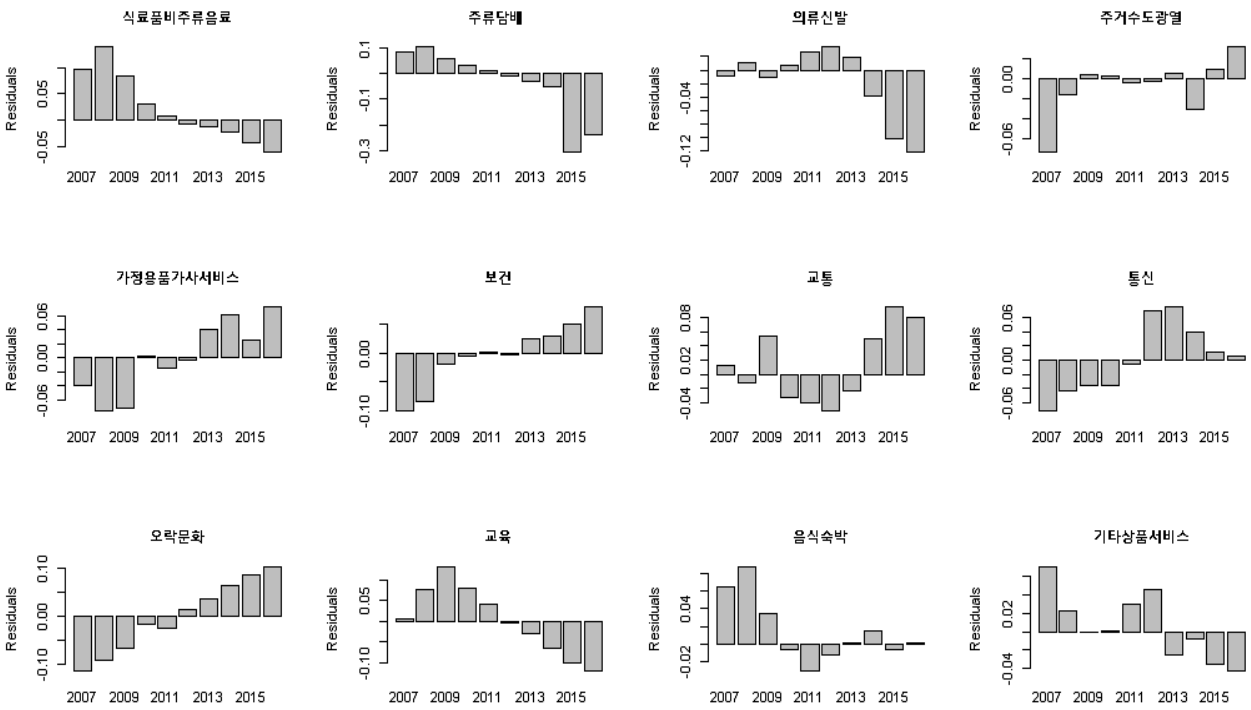
```
barplot(consume.log.out$col,ylab="col effects")
```



다음으로 열효과를 살펴보기 위해 상단의 코드를 실행하여 plot을 그려보았다.
식료품비주류음료가 가장 큰 값을 가지고 있었다.

더욱 자세하게 살펴보기 위해 아래의 코드를 실행하여 plot을 그려보았다.

```
par(mfrow=c(3,4))
barplot(consume.log.out$residuals[,1],
main="식료품비주류음료",ylab="Residuals");barplot(consume.log.out$residuals[,2],main="주류담배 ",ylab="Residuals")

barplot(consume.log.out$residuals[,3],main="의류신발“,
ylab="Residuals");barplot(consume.log.out$residuals[,4],main="주거수도광열",ylab="Residuals")

barplot(consume.log.out$residuals[,5],main="가정용품가사서비스”,
ylab="Residuals");barplot(consume.log.out$residuals[,6],main="보건",ylab="Residuals")
barplot(consume.log.out$residuals[,7],
main="교통",ylab="Residuals");barplot(consume.log.out$residuals[,8],main="통신",ylab="Residuals")

barplot(consume.log.out$residuals[,9],
main="오락문화",ylab="Residuals");barplot(consume.log.out$residuals[,10],main="교육",ylab="Residuals")

barplot(consume.log.out$residuals[,11],main="음식숙박“,
ylab="Residuals");barplot(consume.log.out$residuals[,12],main="기타상품서비스",ylab="Residuals")
```



잔차를 통해 시간의 흐름에 따른 각 부문의 증감소 추세를 확인해볼 수 있었다.
보건, 오락문화의 경우 시간의 흐름에 따라 지출이 지속적으로 늘어나는 추세를 확인할 수 있었다. 그에 비해 식료품비주류음료는 시간의 흐름에 따라 감소하는 양상을 살펴볼 수 있었다.

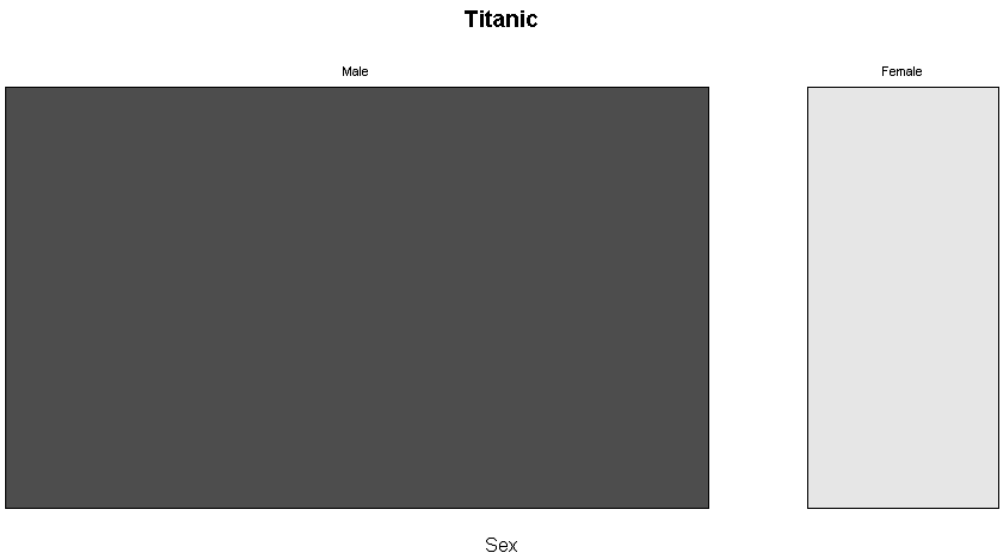
y축의 경우 각 부문 별로 범위가 다르다는 것을 알 수 있었는데 주류담배가 가장 큰 변동폭을 보임을 알 수 있었고 오락문화 또한 0.1 정도로 변동폭이 큰 편이었다.

3. 타이타닉 자료를 모자이크플랏으로 그리고 각 그룹별 또는 그룹 조합별의 생존율을 비교 분석하여라.

```
> head(Titanic)
, , Age = Child, Survived = No

      Sex
Class  Male Female
  1st     0      0
  2nd     0      0
  3rd    35     17
  Crew    0      0

, , Age = Adult, Survived = No

      Sex
Class  Male Female
  1st   118      4
  2nd   154     13
  3rd   387     89
  Crew  670      3

, , Age = Child, Survived = Yes

      Sex
Class  Male Female
  1st     5      1
  2nd    11     13
  3rd    13     14
  Crew    0      0

, , Age = Adult, Survived = Yes

      Sex
Class  Male Female
  1st    57    140
  2nd    14     80
  3rd    75     76
  Crew  192     20
```

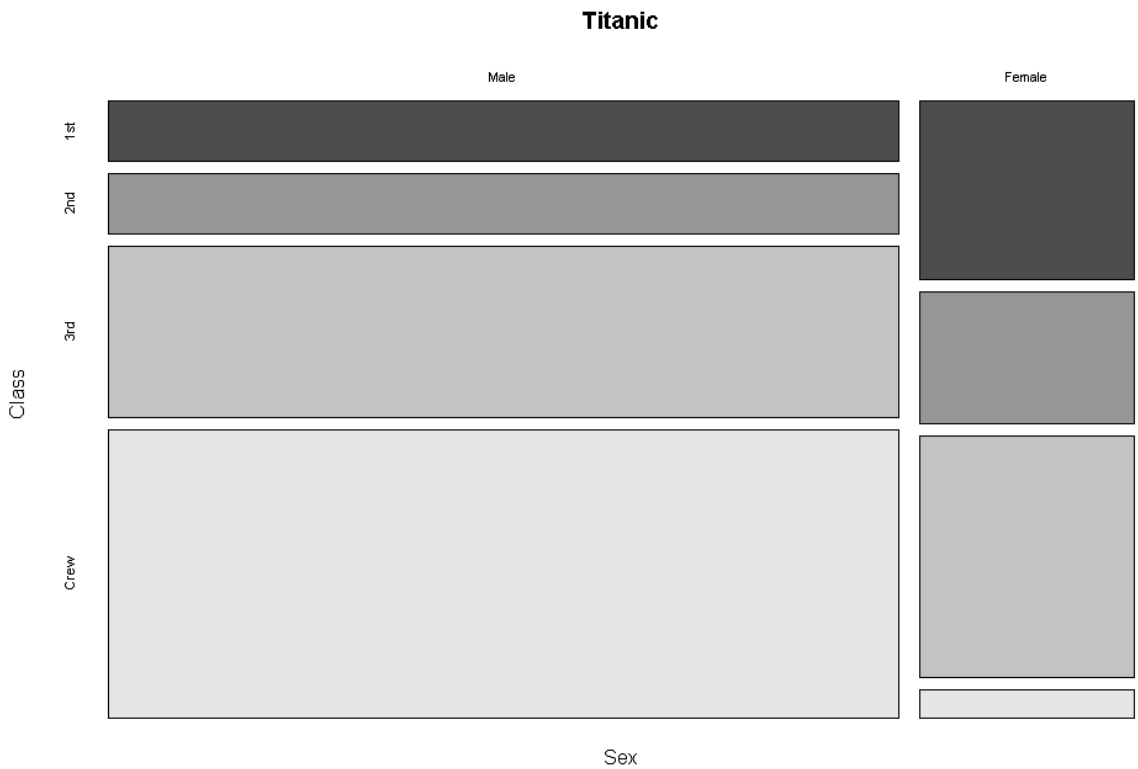
Titanic 데이터를 불러온 뒤, head()를 통해 대략적인 데이터를 확인해보았다.
Titanic 데이터는 Class, Sex, Age, Survived 총 4개의 변수로 구성된 4-dimensional array이다. 4개의 변수에 대한 2201개의 observation들이 있다. Class는 1st, 2nd, 3rd, Crew 총 4개의 level로 구성되어 있고, Sex는 Male, Female로 구성되어 있으며, Age는 Child, Adult, Survuved는 No, Yes로 구성되어 있다.

```
mosaicplot(~ Sex, data = Titanic, color = TRUE)
```



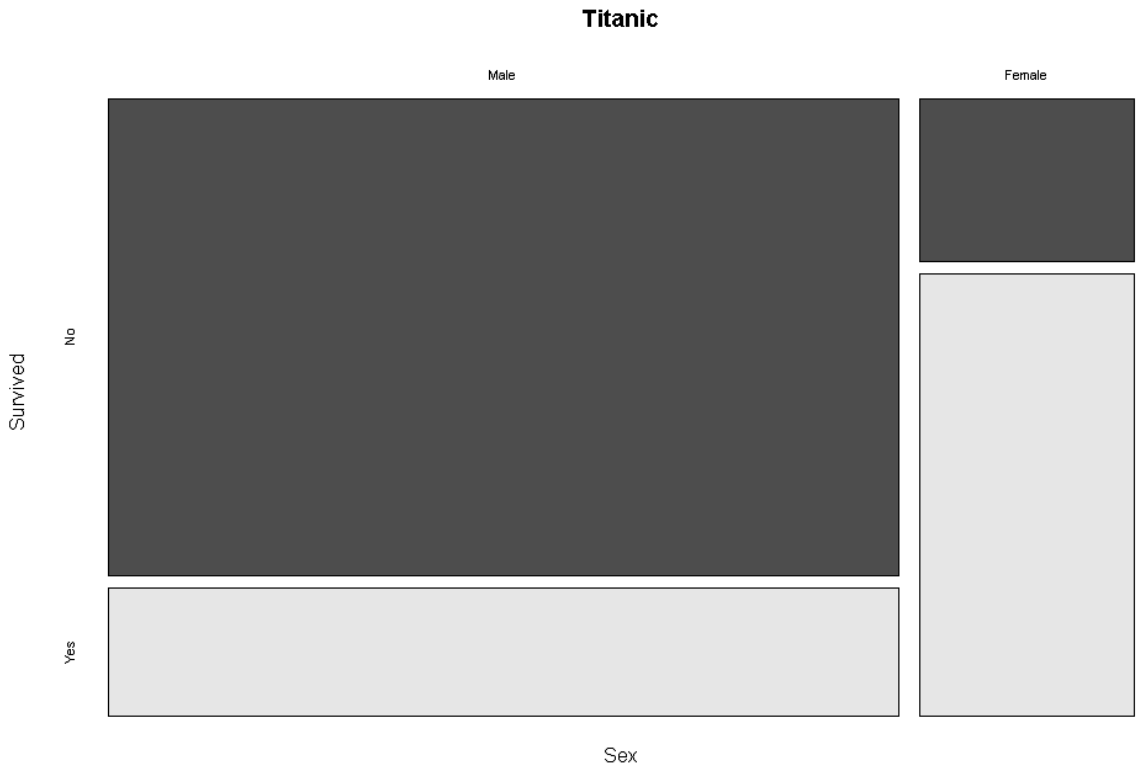
상단의 코드를 실행한 후 출력된 모자이크 plot은 Titanic에 탑승한 여성과 남성의 수를 보여준다. 남성의 탑승객 수가 여성 탑승객 수 보다 3배 이상 많은 것을 확인할 수 있었다.

```
mosaicplot(~ Sex+Class, data = Titanic, color = TRUE)
```



또한, 상단의 코드를 실행하여 성별과 Class 간의 관계를 살펴보고자 했다. 3등석, 선원 class와 다르게 남성의 1등석, 2등석 탑승 비율은 비슷했고 여성 또한 비슷한 비율을 보였다.

mosaicplot(~Sex+Survived,data=Titanic,color=T)

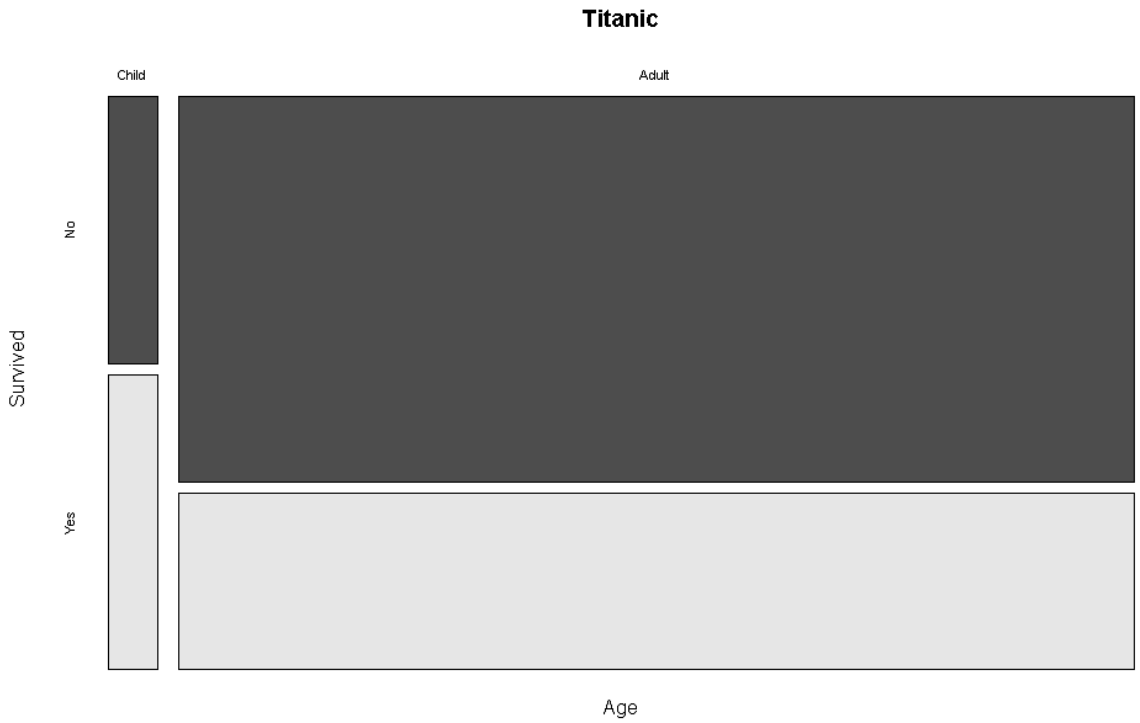


다음으로, 상단의 코드를 실행한 후 출력된 모자이크 plot을 살펴보았다. 남성 생존자 비율보다 여성 생존자 비율이 훨씬 높음을 확인할 수 있었다.

```
> library("vcd")
필요한 패키지를 로딩중입니다: grid
경고메시지(들):
패키지 'vcd'는 R 버전 4.0.5에서 작성되었습니다
> chisq.test(structable(~Sex+Survived,data=Titanic))

        Pearson's Chi-squared test with Yates' continuity correction

data:  structable(~Sex + Survived, data = Titanic)
X-squared = 454.5, df = 1, p-value < 2.2e-16
```

다음으로, 카이제곱 테스트를 통해 수치적으로 집단 간의 유의미한 차이가 있는지 확인해보았다. 귀무가설은 '성별에 따른 생존비율의 차이가 없다'이고, 대립가설은 차이가 있다는 것이다. p-value가 매우 작기 때문에 귀무가설을 기각할 수 있다. 즉, 성별에 따라 생존비율의 차이는 존재가 존재한다는 결론에 이를 수 있다.

mosaicplot(~Age+Survived,data=Titanic,color=T)

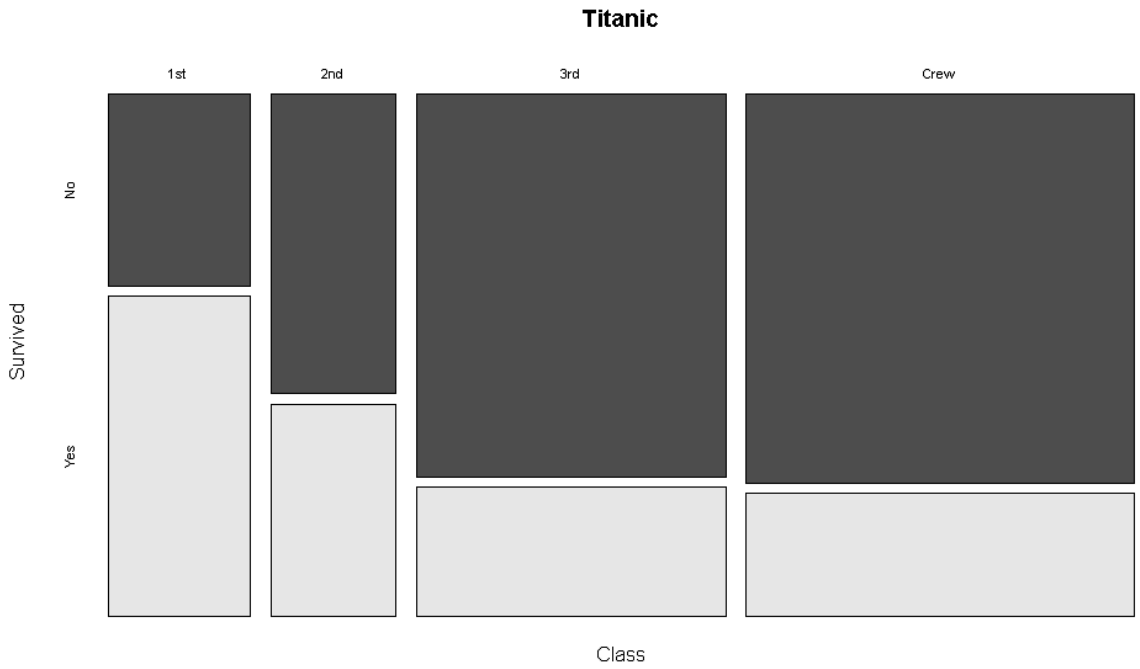


다음으로, 상단의 코드를 실행하여 나이에 따라 생존 비율의 차이를 보고자 mosaic plot을 확인해보았다. 해당 plot을 통해 어른 집단의 생존율이 낮음을 알 수 있었다.

```
> chisq.test(structable(~Age+Survived,data=Titanic))

        Pearson's Chi-squared test with Yates' continuity correction

data:  structable(~Age + Survived, data = Titanic)
X-squared = 20.005, df = 1, p-value = 7.725e-06
```

해당 분석 결과를 카이제곱 테스트를 통해 수치적으로 확인해보았다. p-value가 매우 작기 때문에 귀무가설을 기각할 수 있다. 즉, 연령에 따라 생존비율의 차이는 존재가 존재한다는 결론에 이를 수 있다. 결론적으로, 따라서 아이와 성인의 생존율에는 차이가 있고, 아이의 생존율이 성인의 생존율보다 높다는 결론을 내릴 수 있었다.

mosaicplot(~Class+Survived,data=Titanic, color=T)

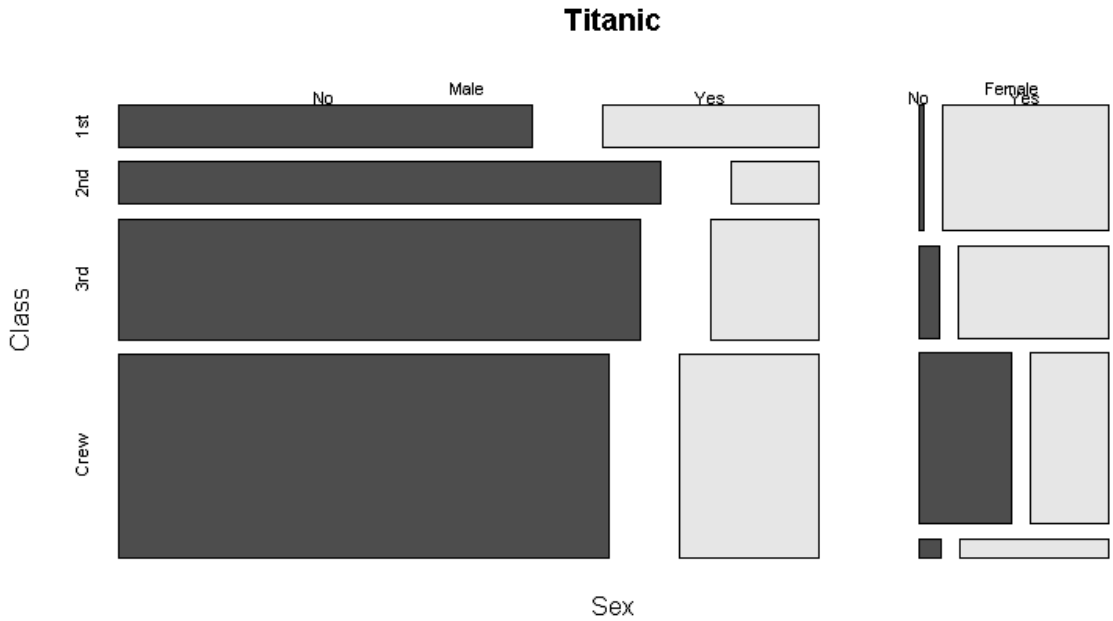


다음으로 class와 생존 여부의 관계를 나타낸 mosaic plot을 그려보았다. 1등석과 2등석에 탄 승객들은 3등석과 선원들에 비해 더 높은 생존율을 보였다.

```
> chisq.test(structable(~Class+Survived,data=Titanic))

        Pearson's Chi-squared test

data:  structable(~Class + Survived, data = Titanic)
X-squared = 190.4, df = 3, p-value < 2.2e-16
```

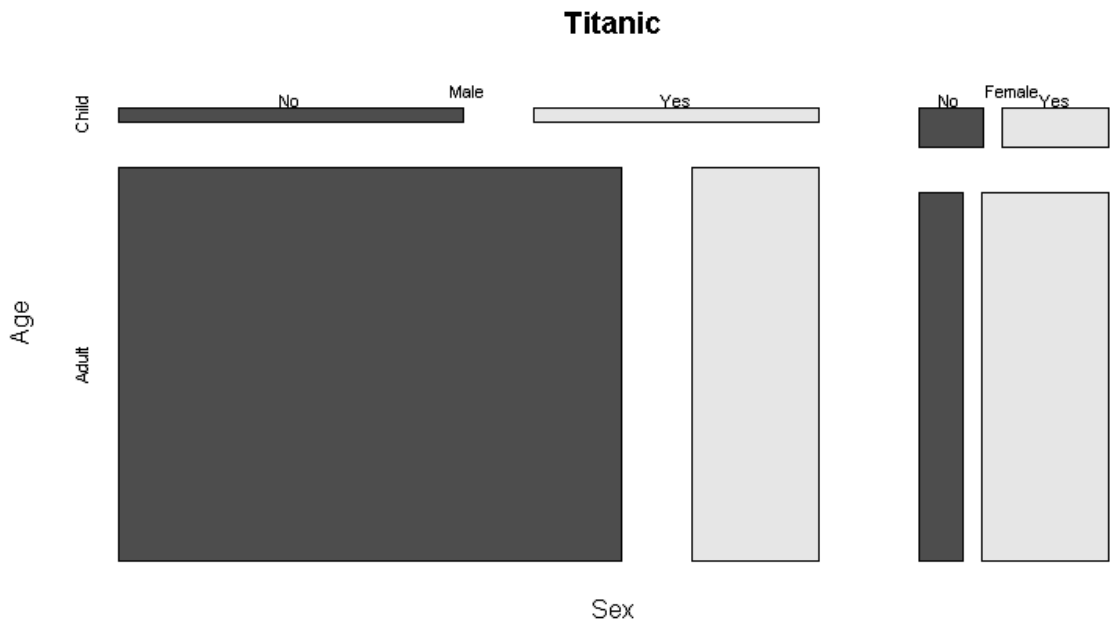
카이제곱 테스트 결과, p-value가 매우 작은 값으로, Class간 생존율의 차이가 있다는 결론을 내릴 수 있었다.

```
mosaicplot(~ Sex+Class+Survived, data = Titanic, color = TRUE)
```



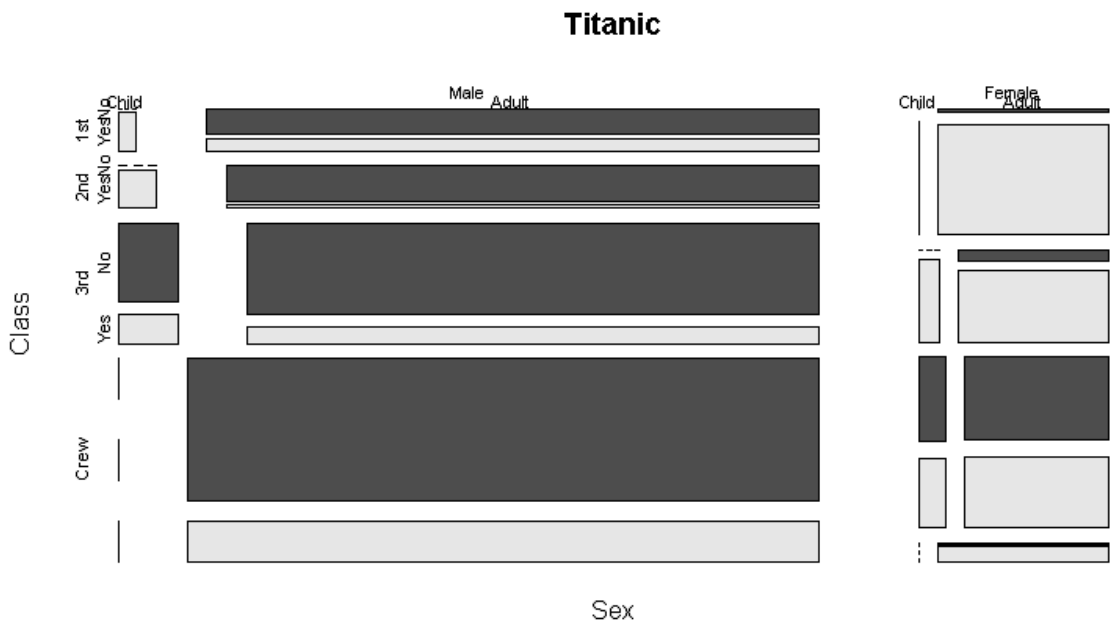
다음으로, Sex, Class 그리고 Survived 세 변수의 관계를 살펴보고자 mosaic plot을 그려보았다. 남성의 경우, 2등석에 탑승한 승객들의 사망률이 제일 높지만 다른 class에서는 특별한 차이를 볼 수 없었다. 그러나 여성의 경우, 3등석->2등석->1등석으로 갈수록 생존율이 높아진다는 것을 알 수 있었다. 또한, 남성 선원에 비해 여성 선원의 생존율이 훨씬 더 높다는 것을 알 수 있다.

```
mosaicplot(~ Sex+Age+Survived, data = Titanic, color = TRUE)
```



상단의 코드를 실행하여 출력된 mosaic plot을 보면 Child 그룹 사이의 성별에 따른 생존율의 차이는 거의 없었다. 그러나 어른 그룹 사이의 성별에 따른 생존율의 차이는 비교적 컸다. 여성 어른들의 생존율이 남성 어른에 비해 더 크다는 사실을 확인할 수 있었다.

```
mosaicplot(~ Sex + Class + Age + Survived, data = Titanic, color = TRUE)
```



마지막으로 4가지 변수를 모두 고려하여, Sex, Class, Age에 따른 생존율의 차이를 살펴보고자

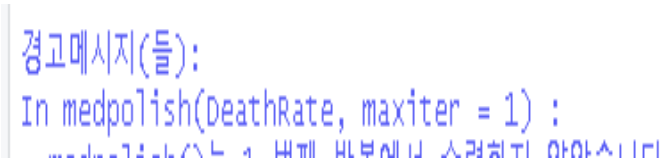
상단의 코드를 실행하여 결과를 출력해보았다. 우선 남성의 경우, 남자 어린아이의 경우, 1등석/2등석 아이들의 생존율은 어른 그리고 3등석, crew의 아이들과 비교했을 때 더 높은 생존율을 보였다. 여성의 경우는 아이와 어른 모두 1, 2등석에 탄 승객들과 crew의 생존율이 높았지만 3등석에 탄 여자 어른, 여자 아이 승객의 생존율은 이들에 비해 낮았다.
전반적으로 남성보다 여성의 생존율이 훨씬 높았고 남성에 비해 여성은 class에 따른 생존율의 차이가 컸다. 또한, 어른에 비해 아이들의 생존율이 더 높았다.

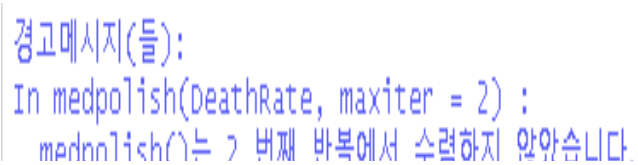
4. 유인물에 있는 자료 중 암발생과 흡연에 대한 자료만 사용하여 R의 중간값 다듬기로 분석하여라. 비교값 그래프로 변환이 필요한지 점검하여라. 투키가 그렸던 도시별 온도에 대한 격자모양 그래프와 같은 것을 그려라. (흰 종이 또는 모눈 종이 위에 자를 사용하여 직접 그려라. R로 그려도 좋으나 프로그램 시간이 오래 걸릴 수 있다.)

```
DeathRate <- rbind(c(0.07, 0.47, 0.86, 1.66),
                   c(0.00, 0.13, 0.09, 0.21),
                   c(0.41, 0.36, 0.10, 0.31),
                   c(0.44, 0.54, 0.37, 0.74),
                   c(0.55, 0.26, 0.22, 0.34),
                   c(0.64, 0.72, 0.76, 1.02))
> colnames(DeathRate)=c("None","1-14","15-24","25+")
> rownames(DeathRate)=c("Lung","Upper respiratory","Stomach","Colon and rectum",
+                       "Prostate","Other")
> DeathRate
                 None 1-14 15-24  25+
Lung             0.07 0.47  0.86 1.66
Upper respiratory 0.00 0.13  0.09 0.21
Stomach          0.41 0.36  0.10 0.31
Colon and rectum  0.44 0.54  0.37 0.74
Prostate         0.55 0.26  0.22 0.34
Other            0.64 0.72  0.76 1.02
> medpolish(DeathRate, maxiter=1)
1: 3.47

Median Polish Results (Dataset: "DeathRate")

Overall: 0.4125

Row Effects:
            Lung Upper respiratory          Stomach  Colon and rectum          Prostate
          0.2525           -0.3025          -0.0775           0.0775          -0.1125
           Other
          0.3275

Column Effects:
         None          1-14         15-24           25+
-7.500000e-02 -6.938894e-18 -5.000000e-02  1.750000e-01

Residuals:
                   None   1-14  15-24    25+
Lung              -0.520 -0.195  0.245  0.820
Upper respiratory -0.035  0.020  0.030 -0.075
Stomach            0.150  0.025 -0.185 -0.200
Colon and rectum   0.025  0.050 -0.070  0.075
Prostate           0.325 -0.040 -0.030 -0.135
Other             -0.025 -0.020  0.070  0.105
```

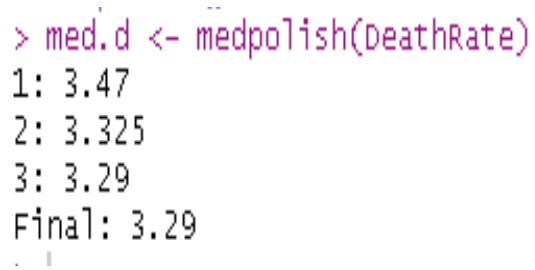
```
> medpolish(DeathRate, maxiter=2)
1: 3.47
2: 3.325

Median Polish Results (Dataset: "DeathRate")

Overall: 0.37125

Row Effects:
             Lung Upper respiratory           Stomach  Colon and rectum          Prostate
          0.29375           -0.29375          -0.14125           0.13125          -0.13125
            Other
          0.36875

Column Effects:
     None      1-14     15-24       25+
-0.07000  0.02875 -0.00375  0.18500

Residuals:
                     None     1-14    15-24     25+
Lung              -0.5250 -0.22375  0.19875  0.8100
Upper respiratory -0.0075  0.02375  0.01625 -0.0525
Stomach            0.2500  0.10125 -0.12625 -0.1050
Colon and rectum   0.0075  0.00875 -0.12875  0.0525
Prostate           0.3800 -0.00875 -0.01625 -0.0850
Other             -0.0300 -0.04875  0.02375  0.0950
```

강의 Script를 토대로 암발생과 흡연에 대한 사망률 데이터를 생성했다. 그 후 medpolish()를 이용하여 중간값 다듬기를 진행해보았다.



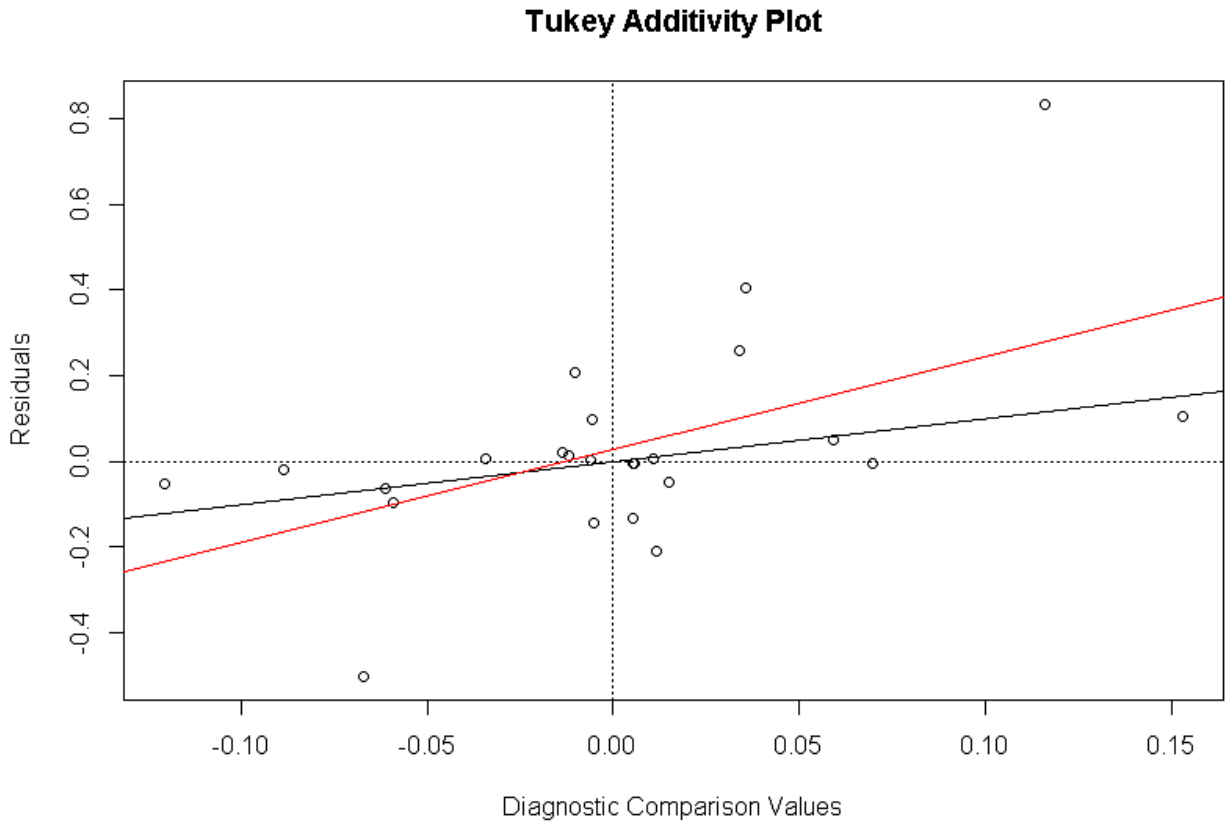


maxiter=1 또는 2일 때는 반복에서 수렴하지 않았다는 경고 메시지를 확인할 수 있었다.



maxiter를 설정하지 않고 분석한 결과 maxiter=3에서 결과값이 나오는 것을 확인하였다.

```
plot(med.d)
abline(0,1)
abline(lm(as.vector(med.d$residuals) ~
            as.vector(outer(med.d$row,med.d$col, "*")/med.d$overall)),col="red")
```



(검은선: 기울기1의 직선, 빨간선: lm()추정 선)
lm()함수를 이용한 기울기 값은 2.16778이었다. 상단의 중간값 다듬기 결과값을 살펴보면 lung의 25+ 잔차값이 다른 잔차에 비해 8.33으로 크다는 것을 알 수 있었다.
그래서 이를 0으로 바꾼 후 기울기 값을 산출했고 그 결과 1.031이 도출되어 log 변환이 적절하다는 것을 알 수 있었다.

로그 변환 후 중간값 다듬기에서 Error 메시지를 확인했기 때문에 이를 해결하고자 수업 Script에 주어진 대로 0.03를 넣어 중간값 다듬기를 진행하였다.

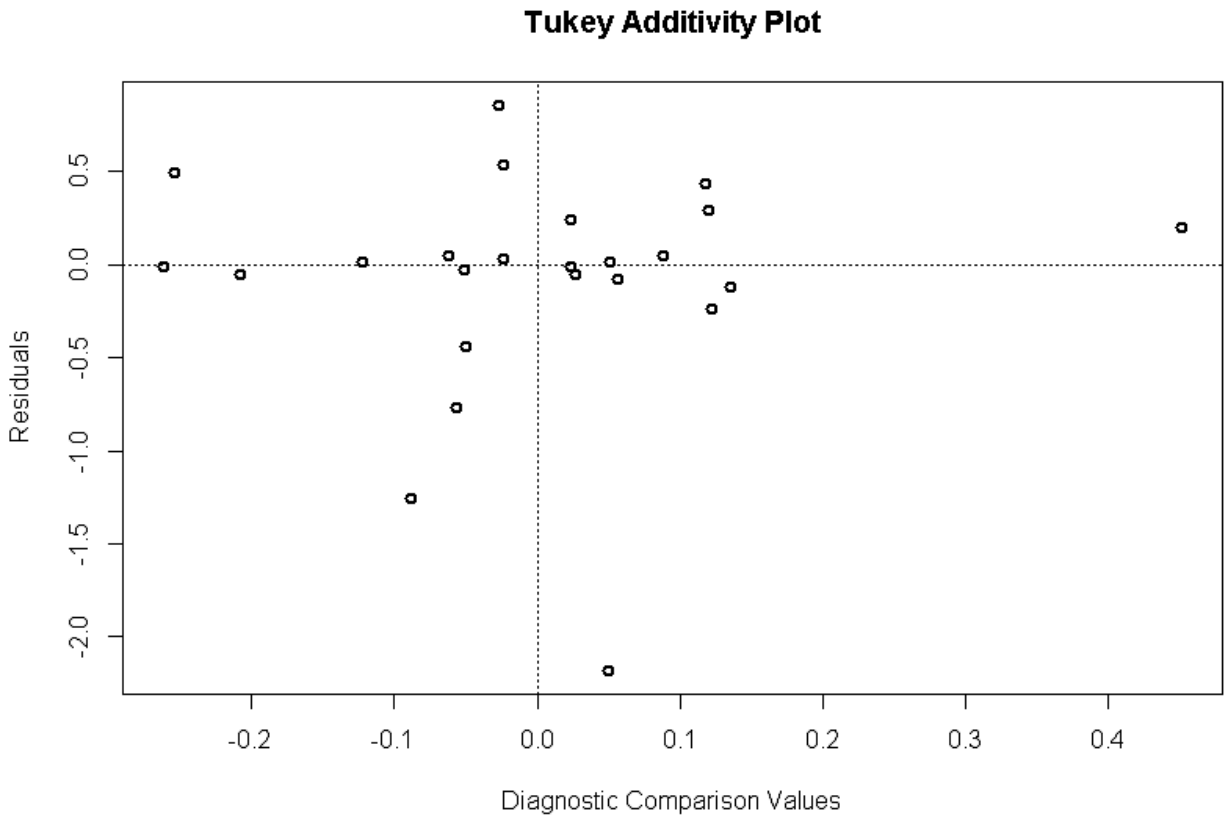
그 결과는 아래와 같다.

```
> DeathRate[2,1] <- 0.03
> (medpolish.log.DeathRate=medpolish(log(DeathRate)))
1: 8.763755
2: 8.444269
Final: 8.444269

Median Polish Results (Dataset: "log(DeathRate)")

Overall: -1.034899

Row Effects:
             Lung Upper respiratory           Stomach  Colon and rectum          Prostate
        0.6384951        -1.1326654        -0.3080093         0.3080093        -0.3391828
            Other
        0.6558099

Column Effects:
       None        1-14       15-24         25+
-0.08064436  0.08064436 -0.19021349  0.41233823

Residuals:
                      None      1-14     15-24       25+
Lung              -2.182212 -0.439263  0.435795  0.490883
Upper respiratory -1.258349  0.046699 -0.050168  0.194578
Stomach            0.531955  0.240613 -0.769463 -0.240613
Colon and rectum  -0.013446  0.030059 -0.077149  0.013446
Prostate           0.856889 -0.053636  0.050168 -0.117066
Other              0.013446 -0.030059  0.294866 -0.013446
```

위의 결과를 통해 Other과 Lung에서 사망률이 높게 나타나고 있으며 담배를 피지 않은 사람의 사망률이 낮다는 것을 알 수 있었다.

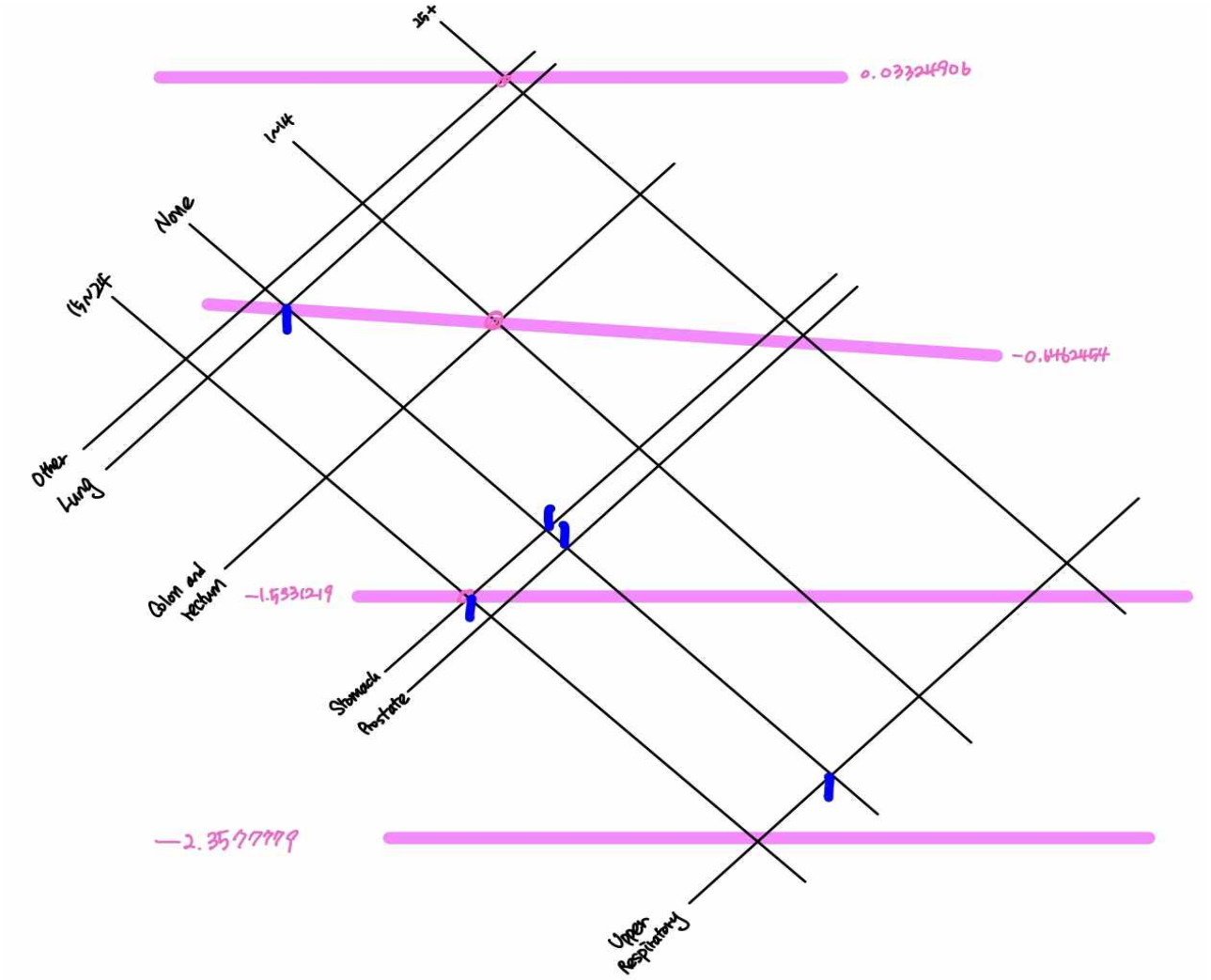
다음으로 비교값과 잔차의 산점도를 그려보았다.

```
plot(medpolish.log.DeathRate,lwd=2)
```



몇몇 잔차의 값이 변환 전과 같이 큰 값을 가지고 있다. 하지만 lm()함수를 이용하여 기울기 값이 0.1임을 확인할 수 있었고 변환 전에 비해 뚜렷한 패턴이 나타나지 않았기 때문에 log변환이 가장 적합한 변환이라 생각했다.

위 log변환 자료를 토대로 격자모양 그래프를 직접 그려보았다.
(격자모양 그래프를 그릴 수 있는 library를 찾을 수 없어 아이패드를 통해 수기로 그렸습니다)



19개의 잔차의 절대값은 0.5보다 작으며 표시한 잔차의 절대값은 모두 0.5보다 크다.
25gram 이상 흡연하는 경우 Other 암의 사망률이 가장 높았다. 또한, 15-24gram 흡연자의 Upper respiratory 암의 사망률의 가장 낮은 것을 확인할 수 있었다. 모든 종류의 암에서 25gram 이상 흡연했을 때의 사망률이 가장 높게 나타났다.