EDA 7장 과제

2017122007 정헌영

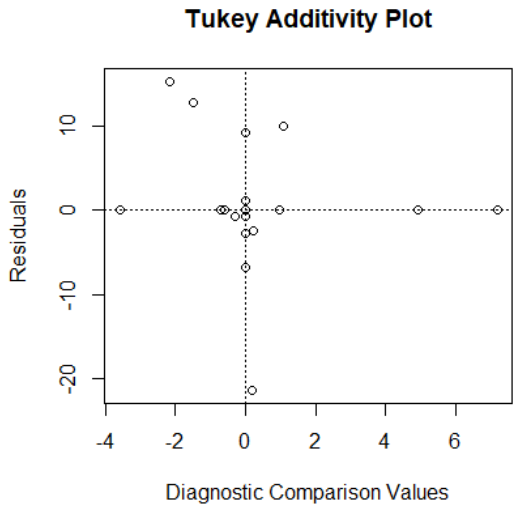
1. 아래 자료를 분석하여라.

soil에 저장된 데이터 :

	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

```
med.s <- medpolish(soil)
```

```
outer(med.s$row,med.s$col, "*")/med.s$overall
```



비교값과 잔차의 산점도를 그려보았다. 산점도에 뚜렷한 직선 경향이 없으므로 모형이 적절하다고 가정한다.

```
med.s
#Residuals:
```

	I	II	III
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

설명:

자료를 다듬기 한 결과이다.

sandy loams(I)의 경우 생산량 차이를 거의 찾아볼 수 없다. Agricultural slag 에서 다소 감소를 보여준다.

sandy clay loam(II)의 경우 None, Coarse slag에서 생산량을 저하시켰으나 Agricultural limestone 에서 생산량 증가를 보여준다.

loamy sand(III)의 경우 None, Coarse slag, Agricultural slag + minor limestone 에서 뚜렷히 높은 생산량을 보여주었으나 Agricultural limestone 에서는 생산량이 급격히 낮아지는것이 확인된다.

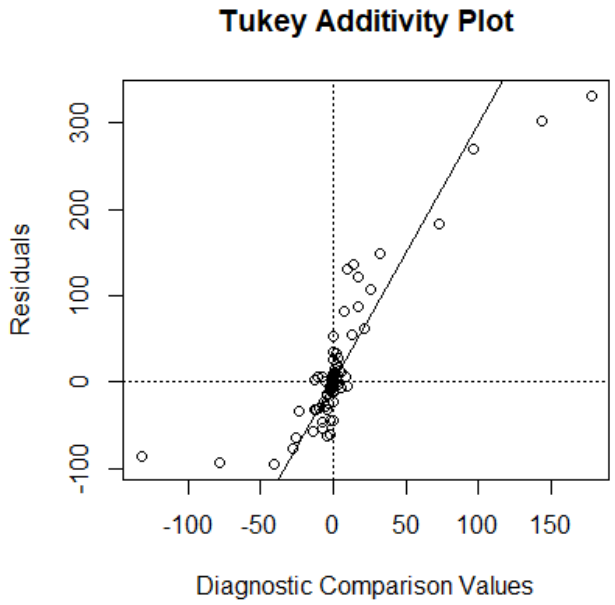
II와 III은 전체적으로 slag와 agricultural liming materials에 각각 반대의 영향을 주는 것으로 추측된다.

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

s :

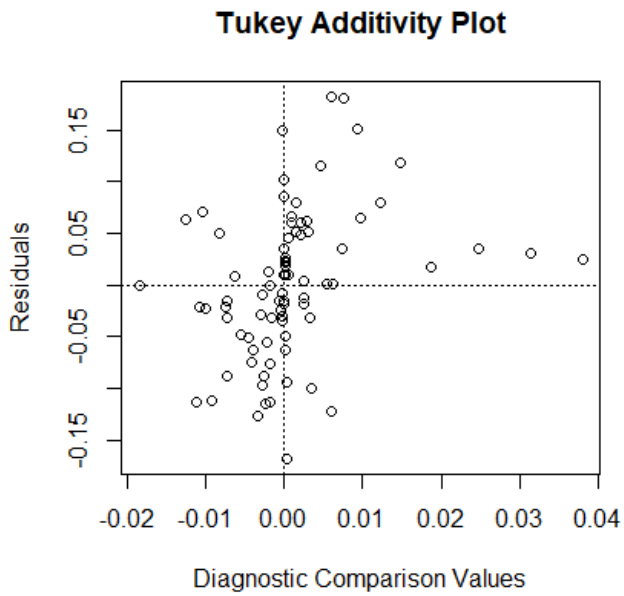
	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
소비지출	2302	2303	2306	2327	2340	2473	2601	2692	2732	2766
식료품	638	624	633	660	673	709	757	793	820	868
주거비	292	302	303	304	306	304	318	315	322	330
교육비	349	340	335	321	313	317	317	318	314	297
의료비	132	138	148	151	152	162	167	177	184	186
교통비	267	269	268	266	257	260	265	263	257	254
통신비	161	173	175	176	174	172	170	166	165	168
기타지출	463	456	445	449	466	550	607	660	671	663

```
med.s <- medpolish(s)
plot(med.s)
abline(0,3)
```



산점도를 확인하면 기울기가 3를 가진 경향성을 가진 것을 확인 할 수 있다. 따라서 자료가 승법모형에 더 적절할 것으로 생각되므로 양변에 로그를 취해 가법적 형태로 변환한다.

```
logmed.s <- medpolish(log(s))
plot(logmed.s)
```



로그 변환 후 더이상 경향성을 찾아볼 수 없으므로 로그변환 모형이 적절하다.

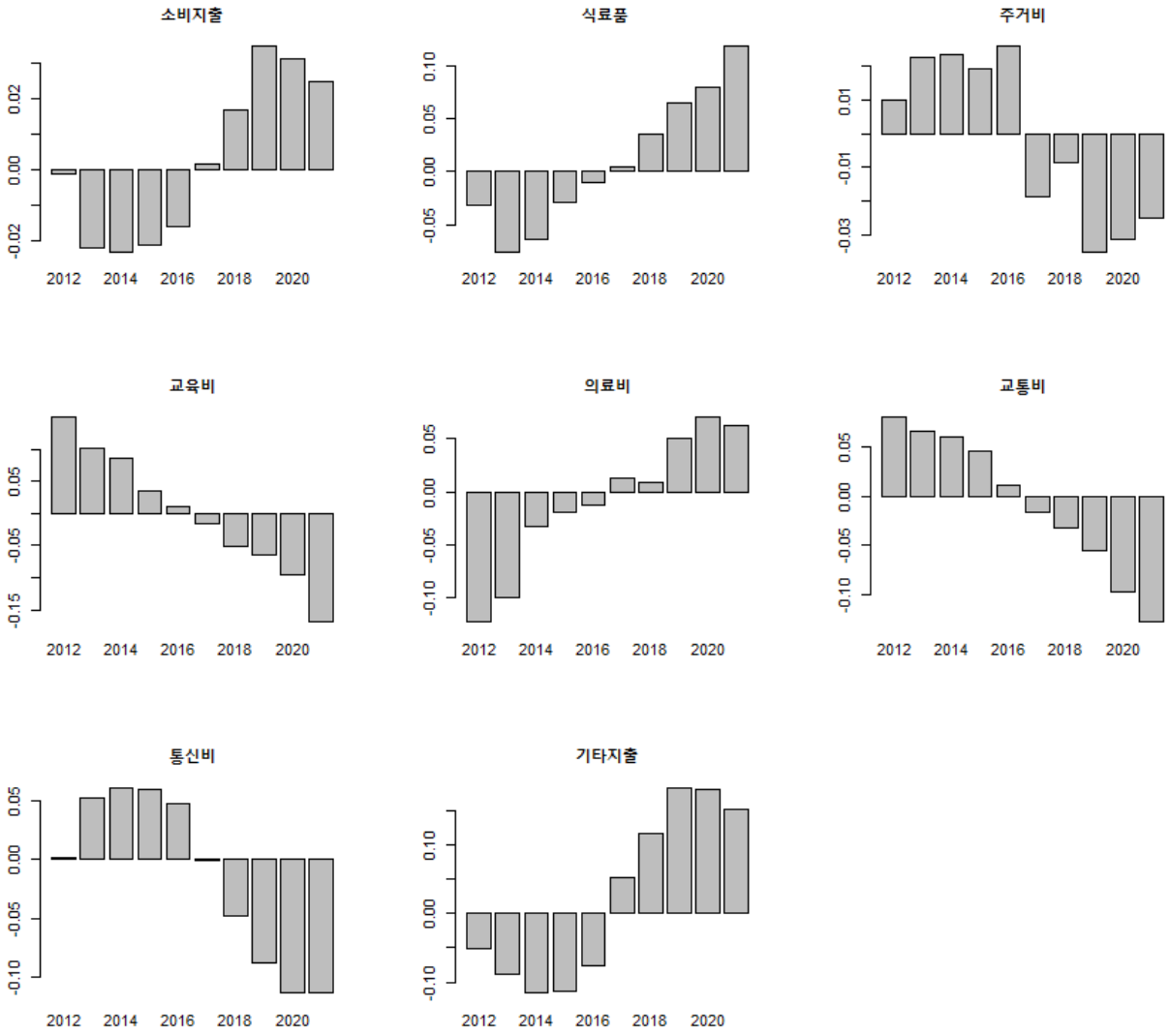
```
logmed.s
```

Residuals:

	2012	2013	2014	2015	2016	2017	2018
소비지출	-0.0013909	-0.022221	-0.023223	-0.021450	-0.0160856	0.0013909	0.0169970
식료품	-0.0321926	-0.075645	-0.063628	-0.029152	-0.0098529	0.0044527	0.0351026
주거비	0.0098115	0.022221	0.023223	0.019224	0.0255754	-0.0187865	-0.0086207
교육비	0.1497890	0.102399	0.085280	0.035297	0.0098529	-0.0152531	-0.0501110
의료비	-0.1228414	-0.099654	-0.031999	-0.019224	-0.0128301	0.0130811	0.0086207
교통비	0.0796393	0.065838	0.059810	0.045026	0.0103994	-0.0157996	-0.0316093
통신비	0.0013909	0.052014	0.061205	0.059610	0.0479745	-0.0013909	-0.0479448
기타지출	-0.0516597	-0.088158	-0.114880	-0.113225	-0.0762684	0.0516597	0.1154123

	2019	2020	2021
소비지출	0.034742	0.031128	0.025041
식료품	0.064920	0.080036	0.118469
주거비	-0.034742	-0.031128	-0.025041
교육비	-0.063604	-0.094627	-0.168742
의료비	0.050134	0.070555	0.062912
교통비	-0.055828	-0.097270	-0.127467
통신비	-0.088398	-0.112805	-0.113241
기타지출	0.182480	0.180645	0.150197

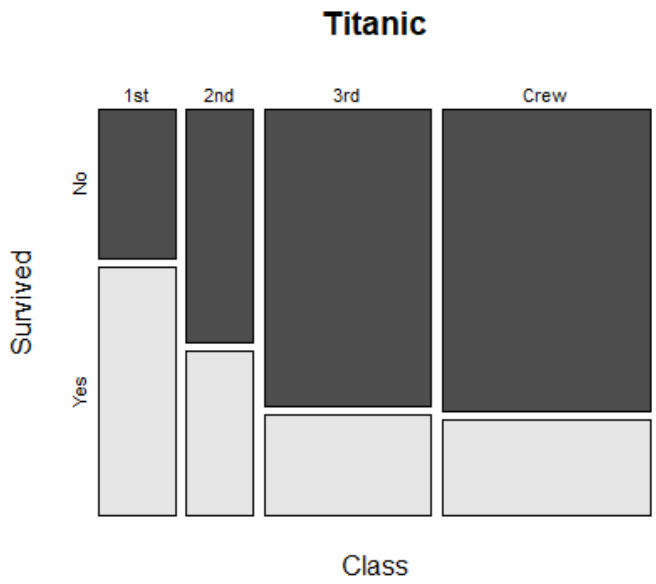
위에서 구한 resudual을 barplot() 함수로 각각 그려본다. 전체적인 지출은 증가했으며, 식료품, 의료비,



기타지출은 증가세인 반면 주거비, 교육비, 교통비, 통신비는 감소세이다. 특히 주거비가 2017년에 급격한 비중 감소를 보였다.
소비지출이 증가랬다는 점에 미루어 생각해볼때 감소세가 나타난 항목인 주거비, 교육비, 교통비, 통신비는 고정비용의 성격이 있어 소득이 증가함에 따라 지출의 증가가 크지 않았기 때문에 상대적으로 지출비중의 하락을 보였을 것으로 생각된다.

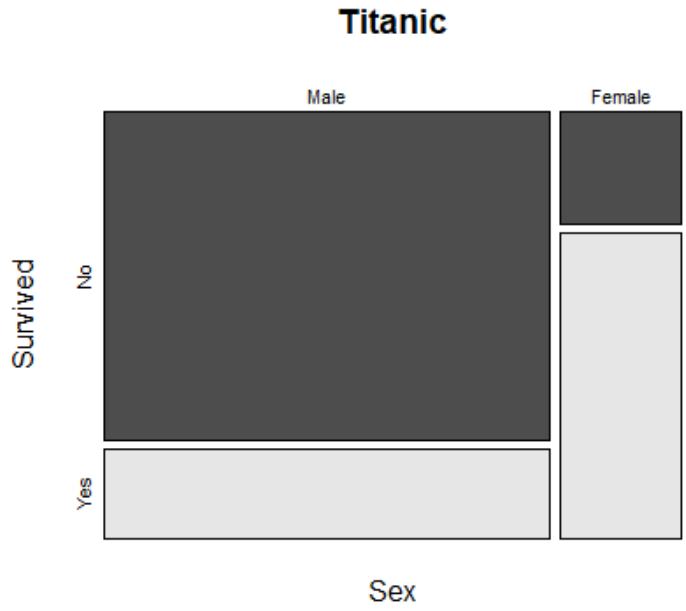
3. 타이타닉 자료를 모자이크플랏으로 그리고 각 그룹별 또는 그룹 조합별의 생존율을 비교 분석하여라.

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



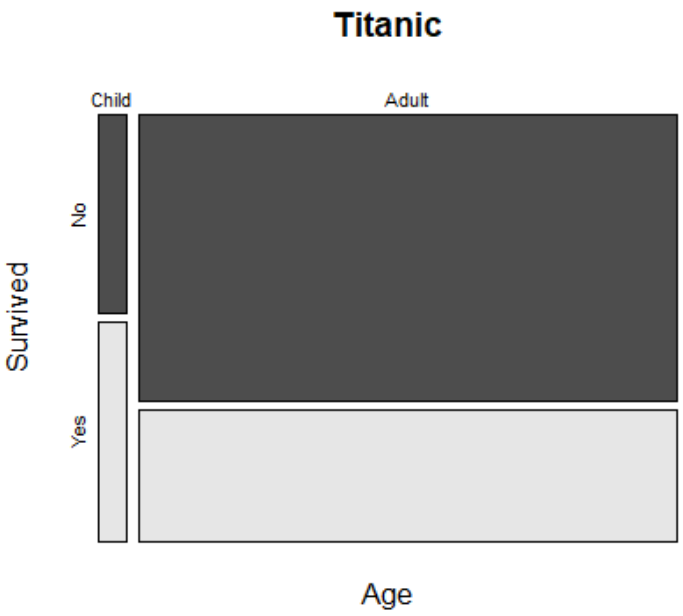
Class에 따른 생존을 분석한 모자이크플랏이다. Class가 높은 경우 생존율이 높은 것을 확인할 수 있다. Crew는 3rd와 비슷한 수준으로 생존했다.

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



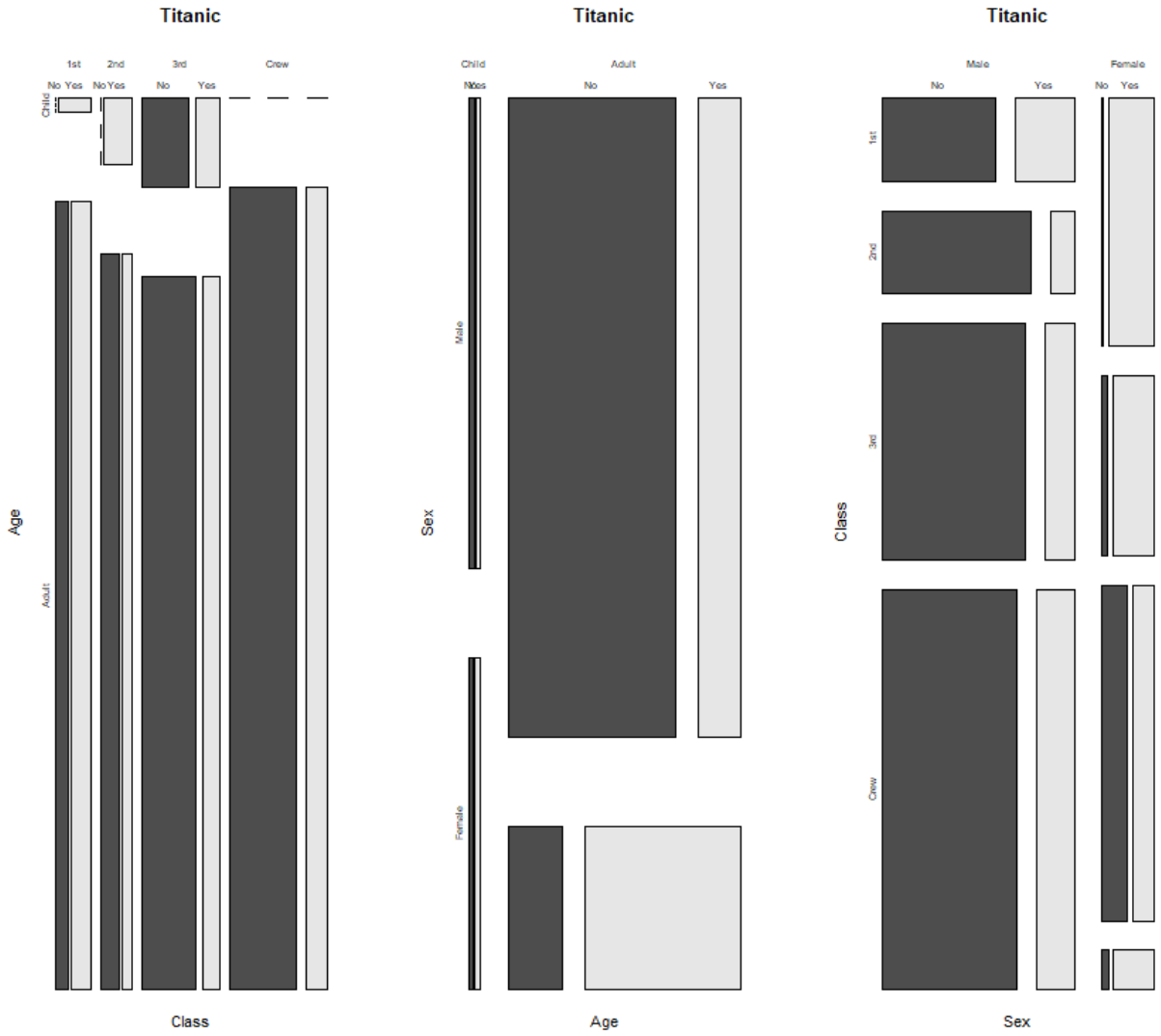
Sex에 따른 모자이크플랏이다. 절대적인 총수는 Male이 3배이상 높다. 그 중에서도 생존한 비율은 여성일 경우 더 높은 것으로 확인된다.

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



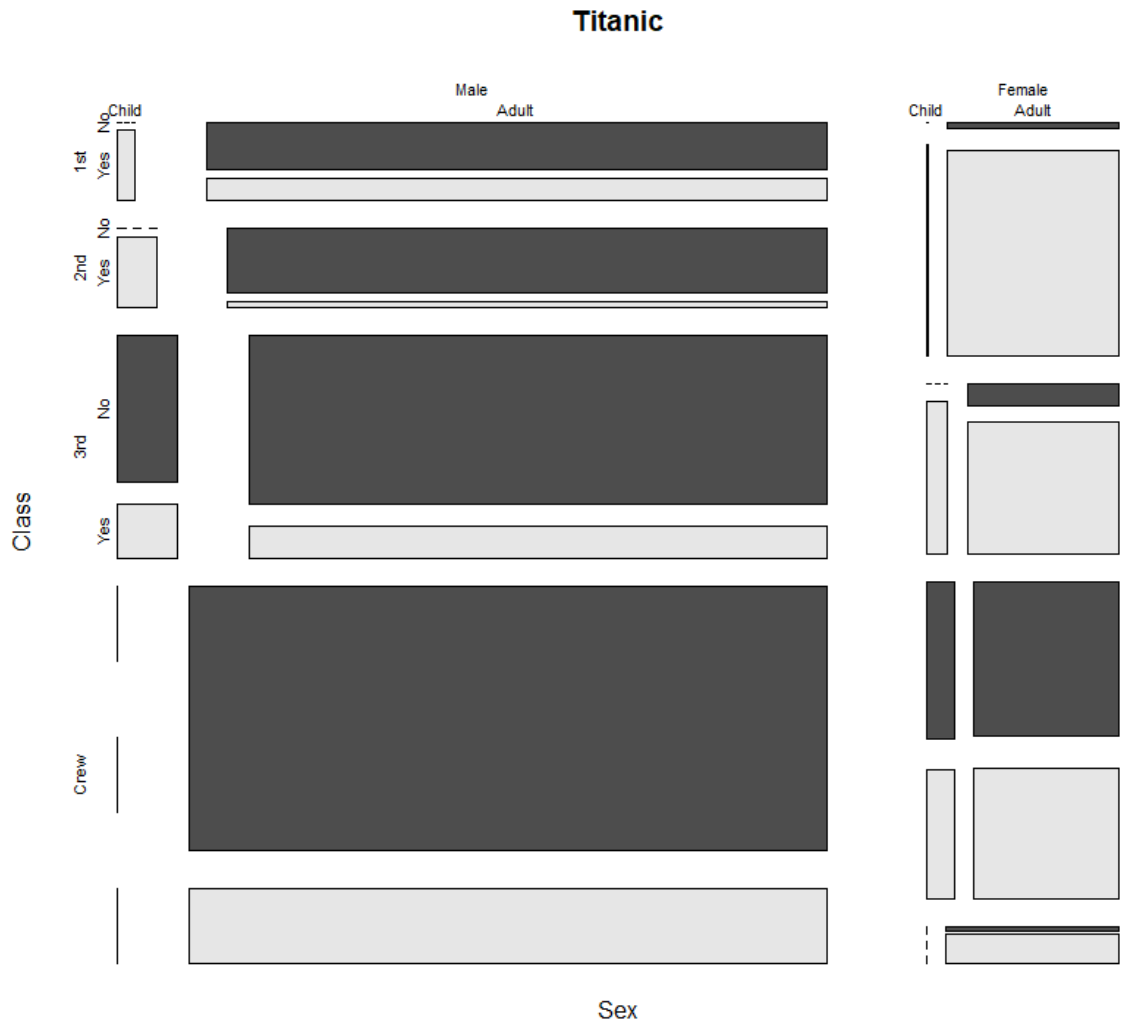
Age에 따른 생존 모자이크플랏이다. Child일 경우 생존한 비율이 더 높았다.

```
par(mfrow=c(1,3))
mosaicplot(~ Class+Age+Survived, data = Titanic, color = TRUE)
mosaicplot(~
+Survived, data = Titanic, color = TRUE)
mosaicplot(~ Sex+Class+Survived, data = Titanic, color = TRUE)
```



Class+Age로 생존여부를 분석하면 어린이일 경우 1st, 2nd Class일 경우 생존하는 비율이 높았지만 어른일 경우 1st일 경우만 생존하는 비율이 높았다.
Age+Sex로 생존여부를 분석하면 어린이와 어른 모두 남자인 경우 생존하는 비율이 적었다. 그러나 절대적인 생존 비율은 어른이 모든 성별에서 높다.
Sex+Class로 생존여부를 분석하면 남성의 경우 1st에서 생존 비율이 높지만 2nd부터 Crew는 오히려 Class가 낮아질수록 생존 비율이 높다. 여성의 경우 1st와 2nd에서 대부분이 생존하며 3rd에서 사망 비율이 높지만 오히려 Crew에서 생존 비율이 높음이 확인된다.

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

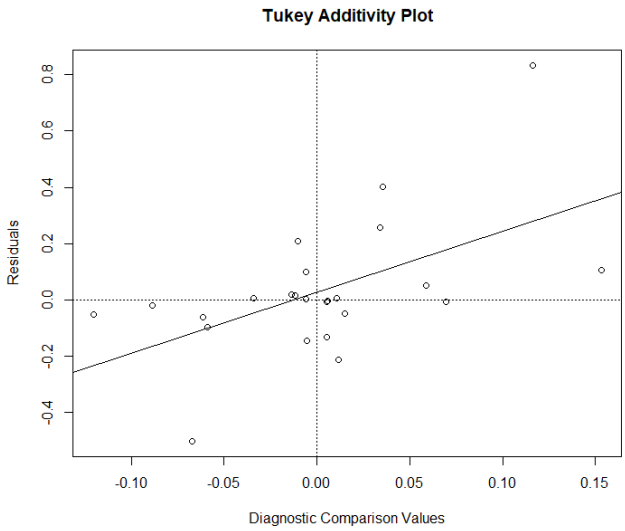


Sex + Class + Age로 생존 여부를 분석한다. 남성 어린이일 경우 Class가 높을수록 생존비율이 높으며 3rd에서 급격히 생존 비율이 적어진다. 남성 어른일 경우 1st에서 생존 비율이 높은 편이지만 2nd부터 Crew까지는 오히려 Class가 낮아질수록 생존 비율이 높아진다. 여성은 1st와 2nd 에서 어른과 아이 모두 생존 비율이 높으나 3rd에서 생존과 사망 비율이 비슷하다. Crew에서는 어른 여성 생존 비율이 높다.

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.

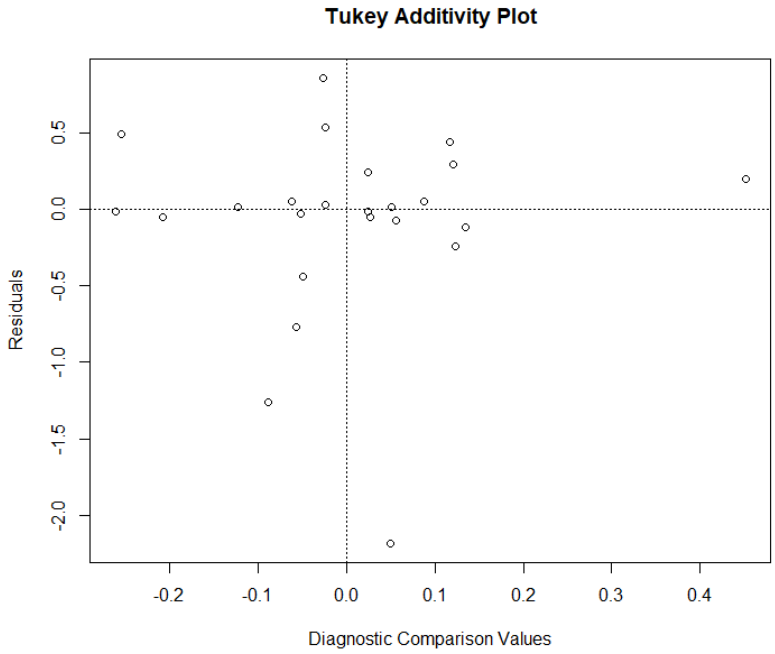
```
DeathRate <- rbind(c(0.07, 0.47, 0.86, 1.66),
                   c(0.00, 0.13, 0.09, 0.21),
                   c(0.41, 0.36, 0.10, 0.31),
                   c(0.44, 0.54, 0.37, 0.74),
                   c(0.55, 0.26, 0.22, 0.34),
                   c(0.64, 0.72, 0.76, 1.02))
colnames(DeathRate)=c("None","1-14","15-24","25+")
```

```
rownames(DeathRate)=c("Lung","Upper respiratory","Stomach","Colon and rectum",
                      "Prostate","Other")
(med.d <- medpolish(DeathRate))
plot(med.d)
abline(lm(as.vector(med.d$residuals) ~
          as.vector(outer(med.d$row,med.d$col, "*")/med.d$overall)))
```



비교값과 Residual의 산점도가 직선의 경향을 보이므로 log를 취해 다시 중간값 고르기를 시행한다.
자료의 [2,1]번 데이터가 0이므로 0.03을 더해 로그변환이 가능하도록 수정한다.

```
DeathRate[2,1] <- 0.03
(med.d.log=medpolish(log(DeathRate)))
plot(med.d.log)
```



직선 경향이 사라졌으므로 로그변환 모형이 적절하다고 생각된다.
격자모양 그래프의 각 지점 수치를 파악하기 위해 선형으로 fitting된 경우를 가정한다. 이때 각 부분의 추

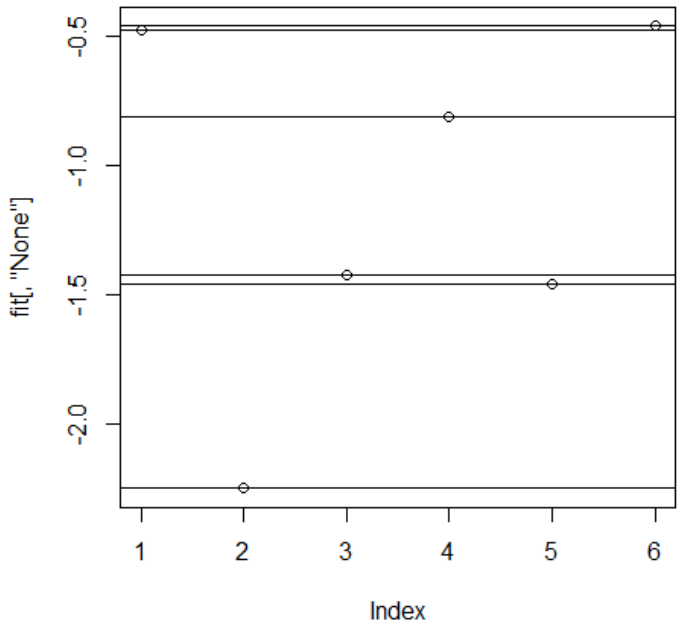
정값은 log를 취한 원 자료에 residual을 차감하는 방식으로 구할 수 있다.

```
log(DeathRate) - med.d.log$residuals
                          None       1-14       15-24         25+
Lung                -0.4770483 -0.3157596 -0.5866175   0.01593427
Upper respiratory   -2.2482088 -2.0869201 -2.3577779 -1.75522621
Stomach             -1.4235528 -1.2622640 -1.5331219 -0.93057018
Colon and rectum    -0.8075341 -0.6462454 -0.9171033 -0.31455153
Prostate            -1.4547263 -1.2934376 -1.5642954 -0.96174368
Other               -0.4597335 -0.2984448 -0.5693027  0.03324906
```

각 지점을 선으로 연장한 격자모양 그래프는 반드시 선들이 평행하므로 대표할 행과 열을 각각 1개씩만 선택해 선을 그리고 이를 연장해 격자모양 그래프를 그린다.

```
fit <- log(DeathRate) - med.d.log$residuals

plot(fit[,"None"])
for(i in 1:6){
  abline(fit[,"None"][i],0)
}
```

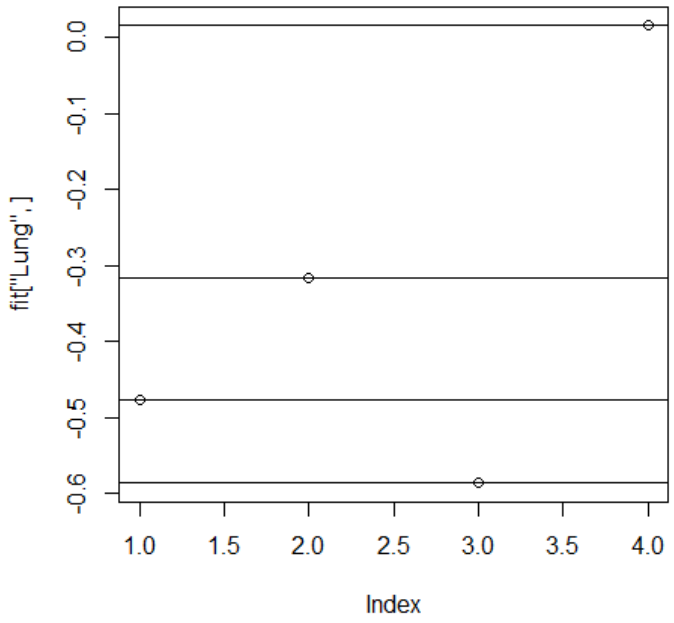


... (1)

(1)은 None열을 선택해 그린 점들이므로 위에서 구한 각 부분의 추정값과 비교하면 그려진 선들은 위에서부터 Other, Lung, Colon and rectum, Stomach, Prostate, Upper respiratory에 대응함을 확인할 수 있다.

```
plot(fit["Lung",])
```

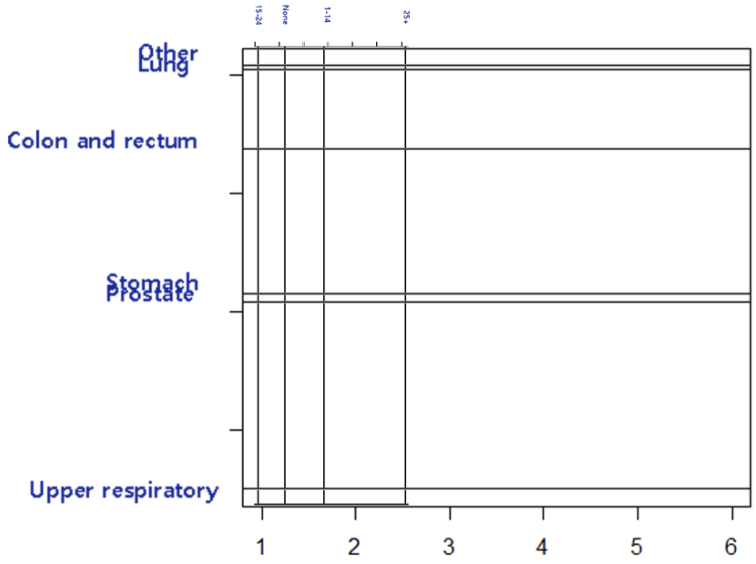
```
for(i in 1:4){
  abline(fit["Lung",][i],0)
}
```



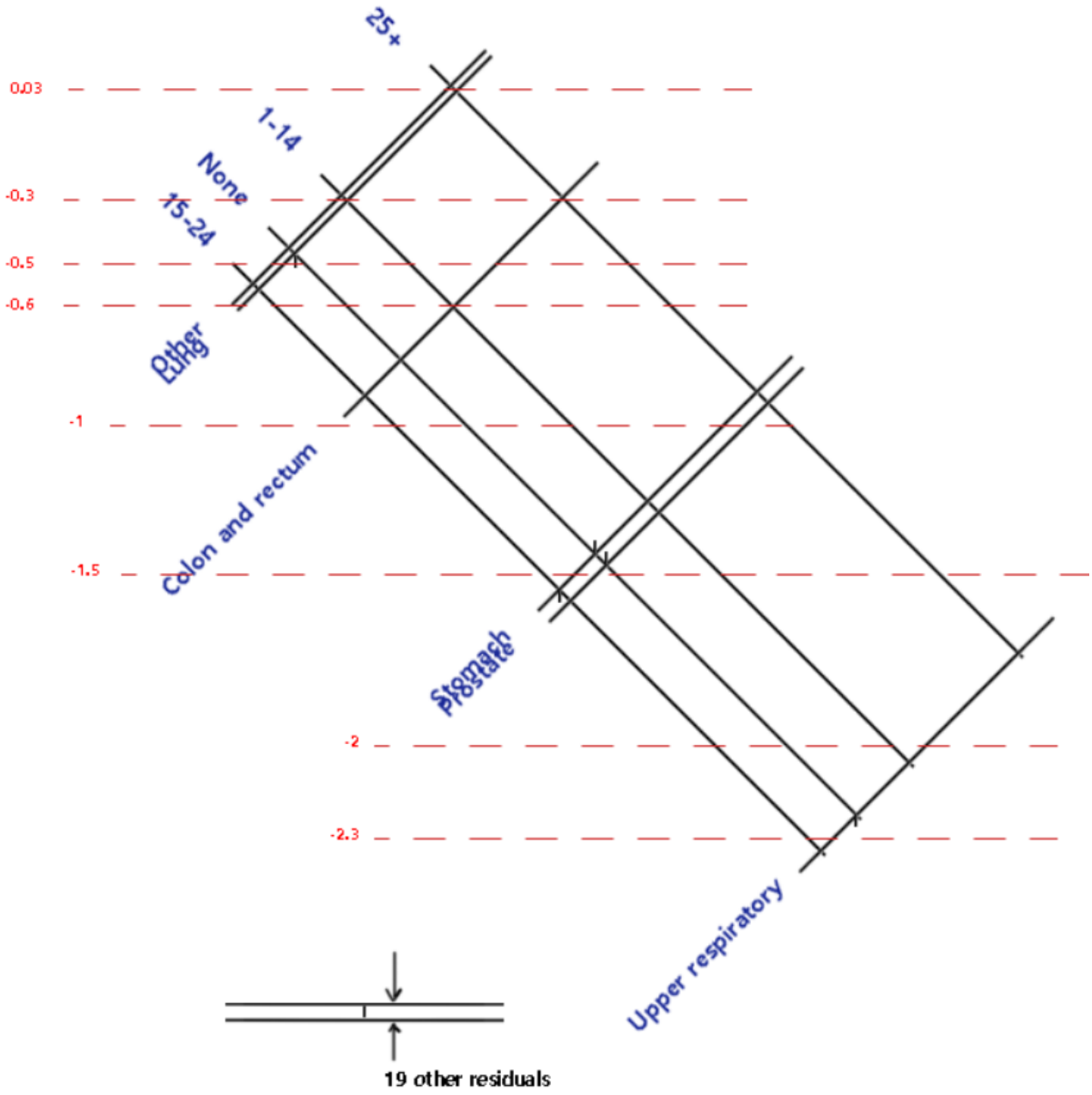
...(2)

(2)은 Lung행을 선택해 그린 점들이므로 위에서 구한 각 부분의 추정값과 비교하면 그려진 선들은 위에서부터 25+, 1-14, None, 15-24임을 확인할 수 있다.

(1)과 (2)를 Photoshop을 이용해 포개어 다음과 같은 격자를 얻는다.



그림을 수정하여 눈금을 표시하고 격자모양 그래프를 완성한다.



sort(fit)

```
 [1] -2.35777793 -2.24820880 -2.08692008 -1.75522621 -1.56429541 -1.53312191 -1.45472627
 [8] -1.42355277 -1.29343755 -1.26226405 -0.96174368 -0.93057018 -0.91710325 -0.80753412
[15] -0.64624540 -0.58661745 -0.56930267 -0.47704832 -0.45973353 -0.31575960 -0.31455153
[22] -0.29844481  0.01593427  0.03324906
```

그래프를 그릴때 fit를 sort함수로 정렬시키고 각 지점의 높이관계를 sort결과와 일치되도록 확인하며 격자 그림을 완성했다. 표시한 5개의 residual은 절댓값 0.5를 기준으로 표시했다.

참고: residuals

	None	1-14	15-24	25+
Lung	-2.18221172	-0.43926299	0.43579456	0.49088333
Upper respiratory	-1.25834909	0.04669925	-0.05016767	0.19457846
Stomach	0.53195465	0.24061280	-0.76946319	-0.24061280
Colon and rectum	-0.01344643	0.03005926	-0.07714902	0.01344643
Prostate	0.85688927	-0.05363610	0.05016767	-0.11706598
Other	0.01344643	-0.03005926	0.29486582	-0.01344643