분산분석
(One-way ANOVA)

Recall: 평균 비교

- 한 집단의 평균과 특정한 수와의 비교

  ➔ One-sample t-test

  $$H_0: \mu = \mu_0$$

- 독립적인 두 집단의 평균 비교

  $$H_0: \mu_1 - \mu_2 = 0$$

  ➔ Two-sample t-test

  t.test(종속변수~그룹변수)

  t.test(자료1, 자료2)

분산분석

- 세 그룹 이상의 평균이 같은지 검정

- $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$
- H_a: 적어도 하나의 μ_i가 나머지와 다르다.

분산분석 vs. 회귀분석

- **<u>설명변수가 이산형인 회귀분석과 동일</u>**

- 회귀식:$y = \beta_0 + \beta_1 x + \epsilon$
  - 만일 x가 0 또는 1을 가지는 이산형 변수라면?
    - $x = 0 \Rightarrow y = \beta_0 + \epsilon$
    - $x = 1 \Rightarrow y = \beta_0 + \beta_1 + \epsilon$
  - $\beta_1 = 0$이라면 x=0인 그룹과 x=1인 그룹 사이의 평균이 같다.
  - $H_0: \mu_1 = \mu_2 \Leftrightarrow H_0: \beta_1 = 0$

분산분석 vs. 회귀분석

- 그룹이 3 개 이상이라면?
- X가 3개의 그룹을 정의하는 질적변수라면? (예, 서울, 대전, 대구)
- 더미 변수 (k-1)개를 만든다.
  - $x_1 = 1\ if\ x =$ 서울$, x_1 = 0\ elsewhere$
  - $x_2 = 1\ if\ x =$ 대전$, x_2 = 0\ elsewhere$
  - 그럼 대구는?

분산분석 vs. 회귀분석

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
  - X=서울: $y = \beta_0 + \beta_1 + \epsilon$
  - X=대전: $y = \beta_0 + \beta_2 + \epsilon$
  - X=대구: $y = \beta_0 + \epsilon$

$$H_0: \mu_1 = \mu_2 = \mu_3 \Leftrightarrow H_0: \beta_1 = \beta_2 = 0$$

$$\Leftrightarrow H_0: \text{회귀식이 유의하지 않다.}$$

분산분석 in R

- 회귀분석과 마찬가지로 lm 명령어를 사용
- 단, 설명변수가 그룹을 정의하는 질적변수
- Factor 함수를 사용하여 설명변수가 질적변수라고 정의

등급별 영화흥행

- 영화 등급 (전체관람가. 12세 이상 관람가, 15세 이상 관람가, 청소년 관람불가)이 각 영화의 총관객수에 영향이 있는가?

```
> levels(data$등급)
[1] "12세이상관람가" "15세이상관람가" "전체관람가"     "청소년관람불가"

> describeBy(data$총관객수, group=data$등급,mat=TRUE)
item         group1 var  n      mean        sd  median   trimmed       mad    min
11    1 12세이상관람가   1 43 1774489.7 2069031.5  893027 1390505.8 1082776.9 101351
12    2 15세이상관람가   1 94 2095732.6 2824207.5 1045561 1477272.3 1139430.7 101425
13    3     전체관람가   1 48  638541.3  532817.5  343360  571803.5  269161.6 106432
14    4 청소년관람불가   1 42 1015156.6 1133648.8  493634  803595.4  492685.0 113848

        max    range     skew    kurtosis        se
11  9001312  8899961 1.663481  2.32459613 315524.35
12 12983330 12881905 2.211749  4.75169040 291294.76
13  2080445  1974013 1.014964 -0.04105884  76905.58
14  4720050  4606202 1.735021  2.78575581 174925.82
>
```

등급 별 영화흥행

```
> out=lm(log(총관객수)~등급,data)
> summary(out)

Call:
lm(formula = log(총관객수) ~ 등급, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.24526 -0.86293 -0.01617  0.87955  2.60684

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        13.70416    0.17979  76.221  < 2e-16 ***
등급15세이상관람가  0.06818    0.21706   0.314  0.75375
등급전체관람가     -0.68294    0.24756  -2.759  0.00628 **
등급청소년관람불가 -0.43123    0.25578  -1.686  0.09320 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.179 on 223 degrees of freedom
Multiple R-squared:  0.06604,     Adjusted R-squared:  0.05347
F-statistic: 5.256 on 3 and 223 DF,  p-value: 0.001601
```

3개의 더미변수
- X1=1 for 15세이상 관람가
- X2=1 for 전체관람가
- X3=1 for 청소년관람불가

등급별 영화흥행

- H0: 네 영화등급 별 총관객수의 차이가 없다.

⇔H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

⇔H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

⇔H0: 회귀식이 유의하지 않다.

- F-test를 사용하여 검정!

```
F-statistic: 5.256 on 3 and 223 DF,  p-value: 0.001601
```

- P-value<0.05➔ 네 영화등급별 총관객수의 차이가 있다.

다중비교

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       13.70416    0.17979  76.221  < 2e-16 ***
등급15세이상관람가  0.06818    0.21706   0.314  0.75375
등급전체관람가     -0.68294    0.24756  -2.759  0.00628 **
등급청소년관람불가 -0.43123    0.25578  -1.686  0.09320 .
```

- F- test: 회귀계수 전체가 0인지 test ➔ “등급”이란 변수가 유의한지 test
- T-test: 각 회귀계수가 0이 아닌지 test
- “12세이상관람가”와 다른 3개 그룹을 각각 비교한 3개의 test결과 ➔실제로 유의하지 않은데 유의하게 결론이 나올 수 있음.
- **위의 t-test 결과 대신 dunnett 방법 사용!**

다중비교

```
> dunnett=glht(out,linfct=mcp(등급="Dunnett"))
> summary(dunnett)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = log(총관객수) ~ 등급, data = data)

Linear Hypotheses:
                                          Estimate Std. Error t value Pr(>|t|)
15세이상관람가 - 12세이상관람가 == 0  0.06818    0.21706   0.314   0.9769
전체관람가 - 12세이상관람가 == 0      -0.68294    0.24756  -2.759   0.0168 *
청소년관람불가 - 12세이상관람가 == 0 -0.43123    0.25578  -1.686   0.2127
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

전체관람가와 12세 이상 관람가 사이에 유의한 차이가 있다.

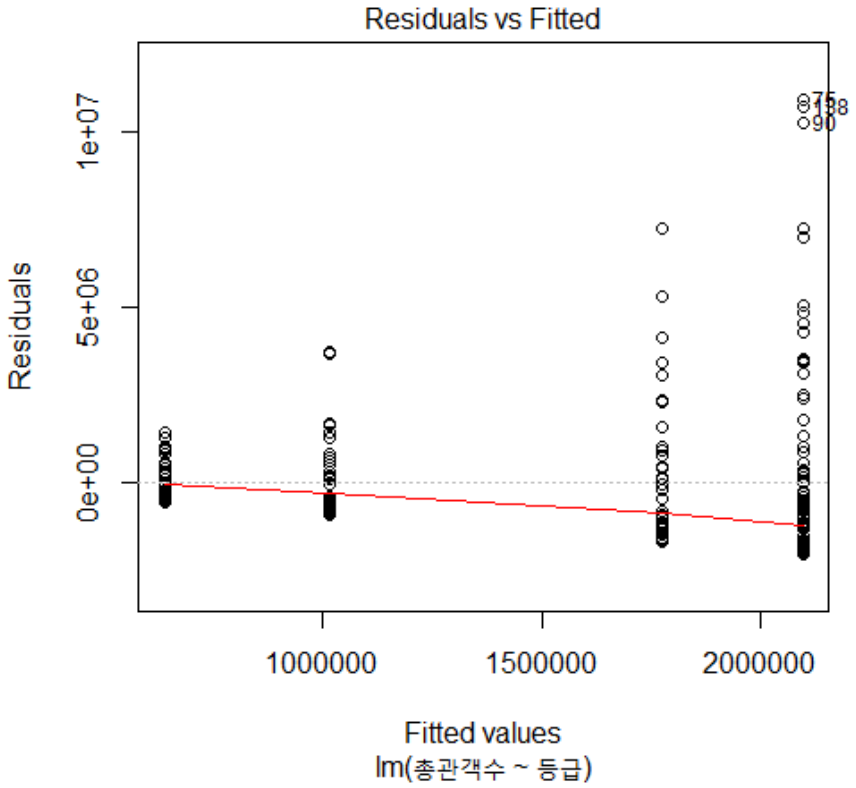
등급별 영화흥행

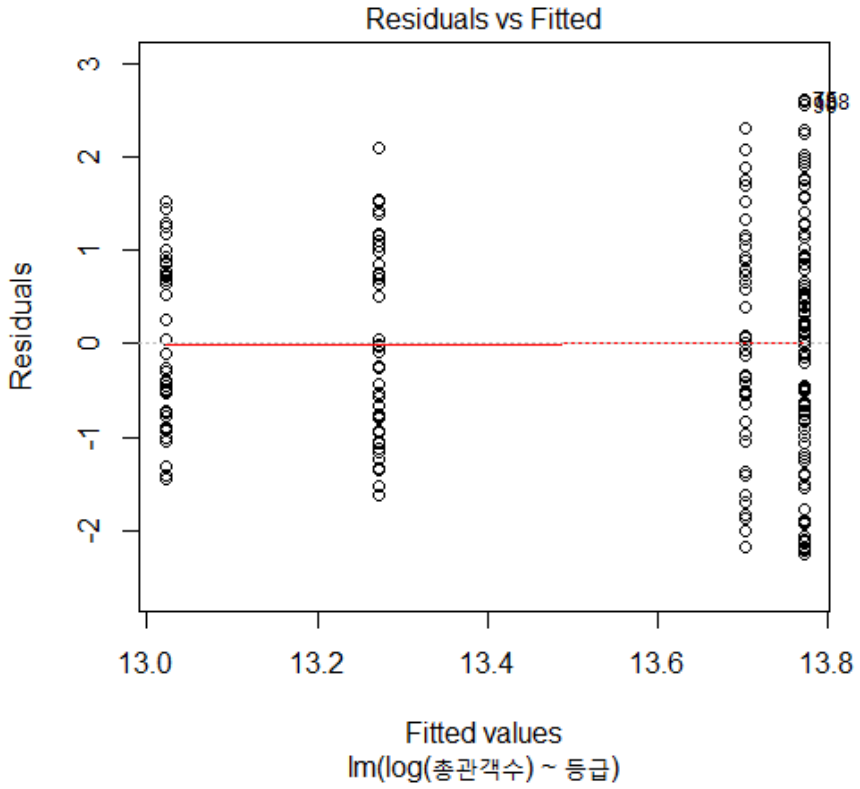
- 회귀추정식

$$\hat{y} = 13.70 + 0.068x_1 - 0.68x_2 - 0.43x_3$$

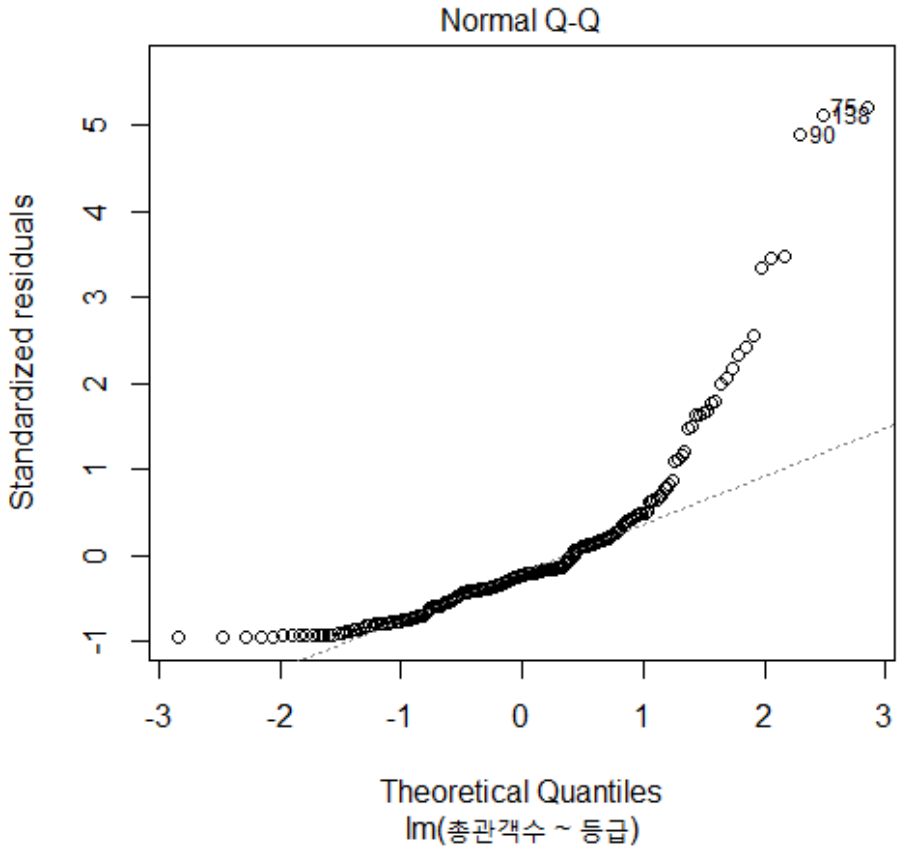
  - 13.70=12세이상 관람가의 평균 log(총관객수)
  - 13.70+0.068=15세이상 관람가의 평균 log(총관객수)
  - 13.70-0.68=전체관람가의 평균 log(총관객수)
  - 13.70-0.43=청소년관람불가의 평균 log(총관객수)

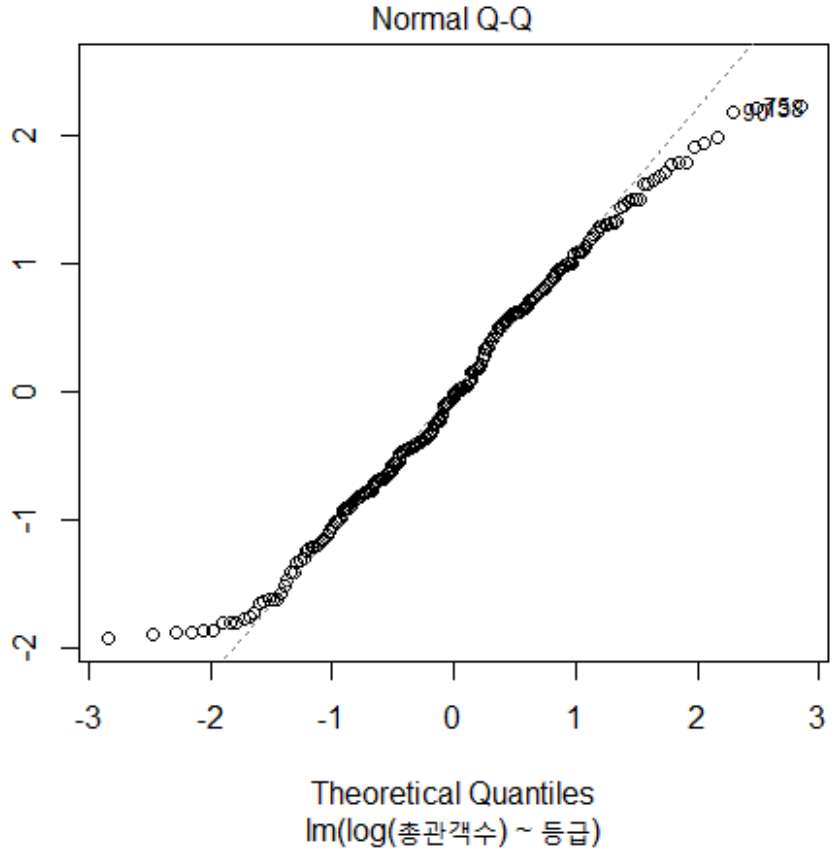
회귀진단


Residuals vs Fitted
Residuals
0e+00
5e+06
1e+07
1000000
1500000
2000000
Fitted values
lm(총관객수 ~ 등급)


Residuals vs Fitted
Residuals
-2
-1
0
1
2
3
13.0
13.2
13.4
13.6
13.8
Fitted values
lm(log(총관객수) ~ 등급)

회귀진단


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(총관객수 ~ 등급)


Normal Q-Q
Theoretical Quantiles
lm(log(총관객수) ~ 등급)

숫자형 질적변수

- 네 개의 등급에 모두 관심이 없고
  (청소년관람불가=1, 전체관람가=2, 나머지=3)의 세 그룹의 차이에 관심이 있다면?

```
> describeBy(data$총관객수, group=data$등급2,mat=TRUE)
   item group1 var   n      mean        sd median   trimmed       mad
11    1      1   1  42 1015156.6 1133648.8 493634  803595.4  492685.0
12    2      2   1  48  638541.3  532817.5 343360  571803.5  269161.6
13    3      3   1 137 1994904.5 2607432.6 978413 1446102.9 1116799.6

      min      max    range     skew    kurtosis        se
11 113848  4720050  4606202 1.735021  2.78575581 174925.82
12 106432  2080445  1974013 1.014964 -0.04105884  76905.58
13 101351 12983330 12881979 2.232542  5.20855569 222768.01
```

숫자형 질적변수

```
> data$등급2=factor(data$등급2)
> out2=lm(log(총관객수)~등급2,data)
> summary(out2)
Call:
lm(formula = log(총관객수) ~ 등급2, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.22459 -0.88038 -0.04003  0.86960  2.62824

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.2729     0.1816  73.107   <2e-16 ***
등급22       -0.2517     0.2486  -1.012   0.3124
등급23        0.4780     0.2075   2.303   0.0222 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 1.177 on 224 degrees of freedom
Multiple R-squared:  0.06562,      Adjusted R-squared:  0.05728
F-statistic: 7.866 on 2 and 224 DF,  p-value: 0.0004994
```

1,2,3을 숫자로 인식하지 않고 그룹을 정의하는 factor로 인식
➔ 2개의 더미변수 생성

등급간에 유의한 차이가 있다.

공분산분석
(ANCOVA)

공분산분석

- 종속변수의 변동을 설명하는데 그룹 변수 이외의 다른 변인이 있을 때 그 효과를 통제
- 공분산분석=분산분석+회귀분석
- 설명변수가 질적변수와 양적변수가 함께 있음

공분산분석: 거식증 치료제

- 거식증에 대한 임상실험으로 CBT, FT, Control 세가지 치료방법을 적용하였다.
  - 종속변수: 치료전후 몸무게 차이 (postwt-prewt)
  - 설명변수: 치료 전 몸무게, 치료방법

공변량
(covariate)

- 분산분석: 치료전후 몸무게 변화가 치료방법 간에 차이가 있는가?
- 공분산분석: 치료 전 몸무게가 무거울수록 몸무게 변화가 크지 않을까? 이것이 치료방법 간 차이를 보는데 방해가 될 수도…

```
> data$Treat=factor(data$Treat,levels=c("Cont","CBT","FT"))

> out=lm(Postwt-Prewt~Prewt+Treat,data)

> anova(out)
Analysis of Variance Table

Response: Postwt - Prewt
          Df Sum Sq Mean Sq F value    Pr(>F)
Prewt      1  447.9  447.85  9.1970 0.0034297 **
Treat      2  766.3  383.14  7.8681 0.0008438 ***
Residuals 68 3311.3   48.70
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

더미변수 생성시 Cont그룹을 레퍼런스로 하기 위해 level의 순서를 바꿔줌

Treat 변수가 설명해주는 Y의 변동성에 대한 Test

P-value<0.05 ➔ 치료효과의 차이가 있다.

```
> summary(out)

Call:
lm(formula = Postwt - Prewt ~ Prewt + Treat, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1083  -4.2773  -0.5484   5.4838  15.2922

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.6740    13.2167   3.456 0.000950 ***
Prewt        -0.5655     0.1612  -3.509 0.000803 ***
TreatCBT      4.0971     1.8935   2.164 0.033999 *
TreatFT       8.6601     2.1931   3.949 0.000189 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.978 on 68 degrees of freedom
Multiple R-squared:  0.2683,      Adjusted R-squared:  0.236
F-statistic: 8.311 on 3 and 68 DF,  p-value: 8.725e-05
```

그룹 간 비교는 t-test 결과 대신 Dunnett test 를 통해 다중비교

다중비교

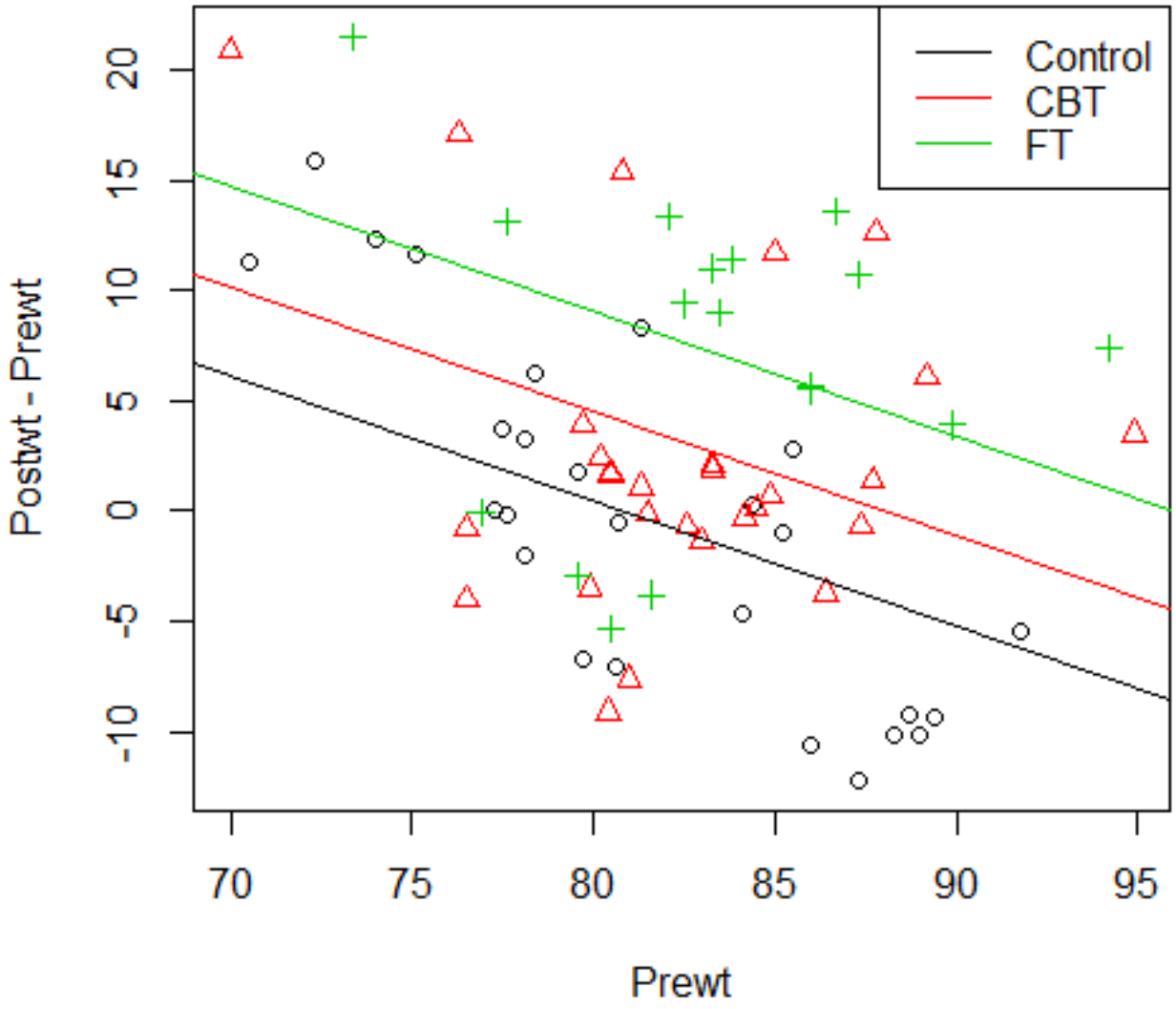
```
> dunnett=glht(out,linfct=mcp(Treat="Dunnett"))
> summary(dunnett)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = Postwt - Prewt ~ Prewt + Treat, data = data)

Linear Hypotheses:
                 Estimate Std. Error t value Pr(>|t|)
Cont - CBT == 0   -4.097      1.893  -2.164   0.0637 .
FT - CBT == 0      4.563      2.133   2.139   0.0674 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```


Control
CBT
FT
Postwt - Prewt
20
15
10
5
0
-5
-10
70
75
80
85
90
95
Prewt

공분산분석: 거식증 치료제

- $\hat{y} = 45.67 - 0.57\ Prewt + 4.10\ x_{CBT} + 8.66\ x_{FT}$
  - Control: $\hat{y} = 45.67 - 0.57\ Prewt$
  - CBT: $\hat{y} = 45.67 + 4.10 - 0.57\ Prewt$
  - FT: $\hat{y} = 45.67 + 8.66 - 0.57\ Prewt$

  - **Prewt이 평균이었던 사람에 대해** CBT는 control 그룹보다 4.10 만큼 더 몸무게 변화를 주었다.
  - **Prewt이 평균이었던 사람에 대해** FT는 control 그룹보다 8.66만큼 더 몸무게 변화를 주었다.

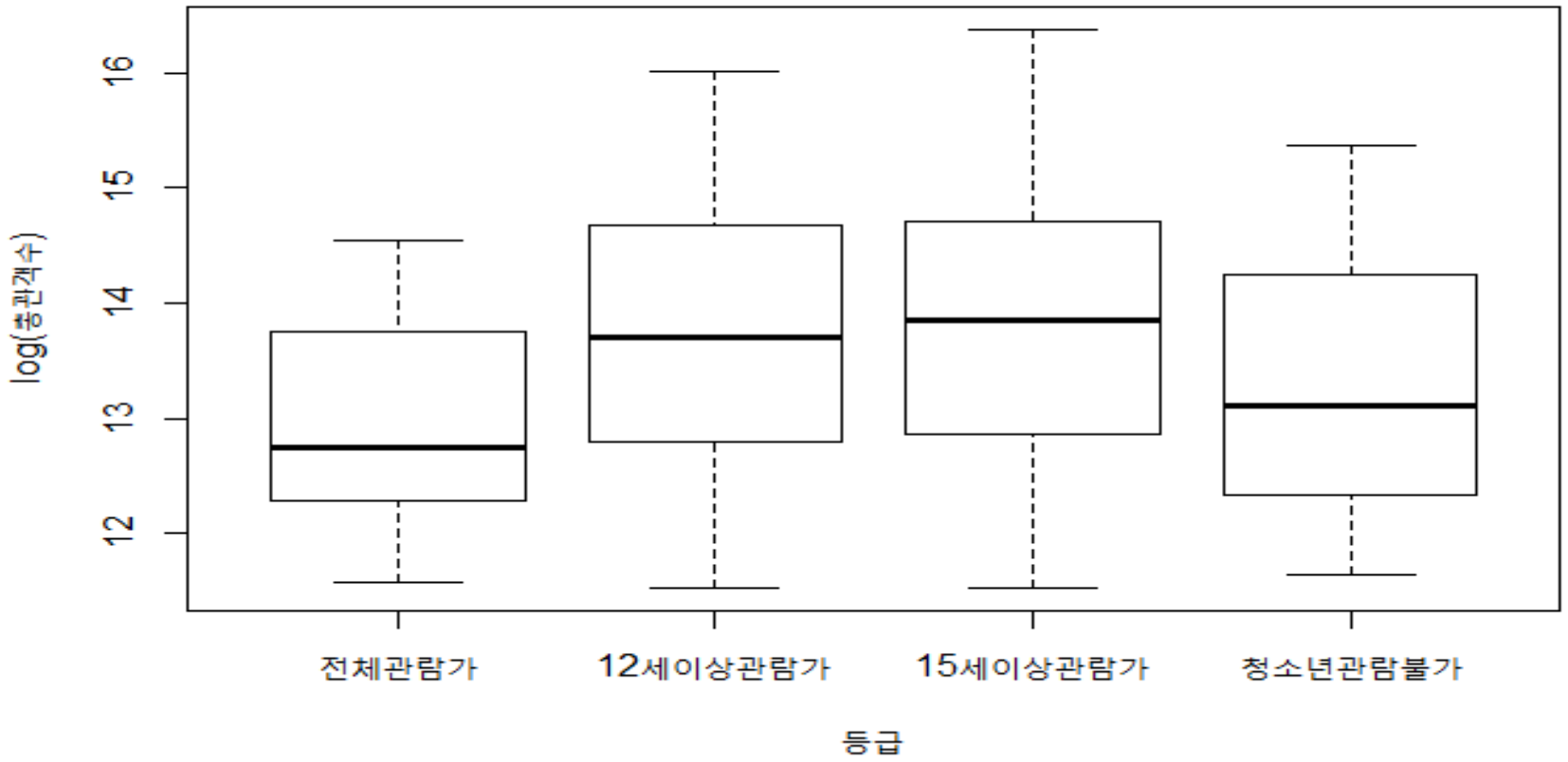
공분산분석: 영화흥행

- 공분산 분석=회귀절편의 집단 간 차이에 대한 검정

```
> out=lm(log(총관객수)~log(첫주관객수)+등급,data)
> anova(out)
Analysis of Variance Table

Response: log(총관객수)
                  Df  Sum Sq Mean Sq   F value  Pr(>F)
log(첫주관객수)    1 307.697 307.697 2907.3595 < 2e-16 ***
등급               3   0.700   0.233    2.2035 0.08854 .
Residuals        222  23.495   0.106
---
```

Log(첫주관객수)의 영향을 통제한 후 "등급"에 따라 log(총관객수)의 유의한 차이가 있다. (alpha=0.1)



- ‘전체관람가’면 관객층이 더 넓은데 15세이상 관람가보다 더 관객이 적다?
- ‘등급’에 `흥행’이 내포되어 있나? 전체관람가 영화는 흥행이 저조한 어린이영화?

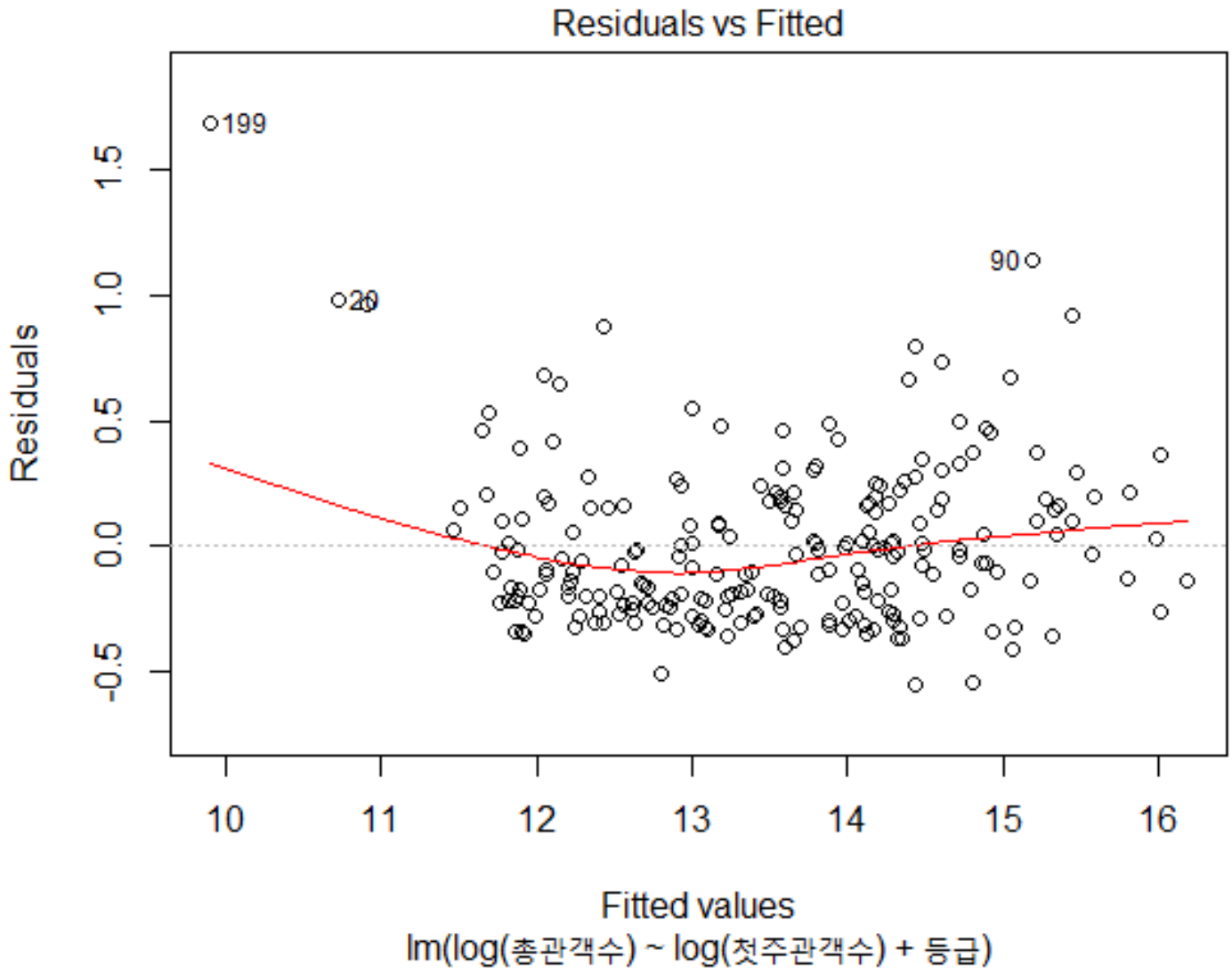
공분산 분석:영화흥행

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.00272    0.27361  -3.665  0.00031 ***
log(첫주관객수)      1.12916    0.02170  52.028  < 2e-16 ***
등급12세이상관람가 -0.11673    0.07002  -1.667  0.09688 .
등급15세이상관람가 -0.11046    0.06004  -1.840  0.06715 .
등급청소년관람불가 -0.17260    0.06922  -2.494  0.01338 *
```

- 전체관람가: log(총관객수)=-1.003 + 1.13*log(첫주관객수)
- 12세이상관람가: log(총관객수)=-1.003-0.12 + 1.13*log(첫주관객수)
- 15세이상관람가: log(총관객수)=-1.003 -0.11+ 1.13*log(첫주관객수)
- 청소년관람불가: log(총관객수)=-1.003 -0.17+ 1.13*log(첫주관객수)

- log(첫주관객수)가 평균 수준인 영화에 대해 log(총관객수)가
  - 12세이상관람가 가 전체관람가보다 -0.12만큼 적다.
  - 15세이상관람가가 전체관람가보다 -0.11만큼 적다.
  - 청소년관람불가가 전체관람가보다 -0.17만큼 적다.

회ㄱ


Residuals vs Fitted
199
20
90
Residuals
1.5
1.0
0.5
0.0
-0.5
10
11
12
13
14
15
16
Fitted values
lm(log(총관객수) ~ log(첫주관객수) + 등급)

회귀진단

```
> data[c(199,20,90),]
                영화명     개봉일  첫주매출액 첫주관객수 대표국적
199       마지막 4중주 2013-07-25   130860500      17214     미국
20            아티스트 2012-02-16   274149000      35829     미국
90  광해, 왕이 된 남자 2012-09-13 13480586000    1854694     한국
                                        제작사                 배급사            등급
199                                                    (주)티캐스트 15세이상관람가
20                                                  (주)영화사 진진 12세이상관람가
90  리얼라이즈픽쳐스(주),씨제이이앤엠 주식회사 씨제이이앤엠 주식회사 15세이상관람가
                           장르              감독
199                      드라마       야론 질버만
20  멜로/로맨스,코미디,드라마 미셸 아자나비슈스
90                 사극,드라마            추창민
                                                                                  배우
199                      필립 세이무어 호프먼,크리스토퍼 월큰,캐서린 키너,마크 이바니르
20                                  진 두자르딘,베레니스 베조,존 굿맨,제임스 크롬웰
90  이병헌,류승룡,한효주,장광,김인권,심은경,김명곤,서진원,장재현,정창국,조혜정,김남준
    총관객수    총매출액 등급2
199   107480   806877000     3
20    120434   921795100     3
90  12323291 88907726769     3
```