분산분석
(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 범주형인 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.
  - $H_0: \mu_1 = \mu_2 \Leftrightarrow H_0: \beta_1 = 0$

- 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?

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

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

$\Leftrightarrow H_0$: 회귀식이 유의하지 않다.

분산분석 in R

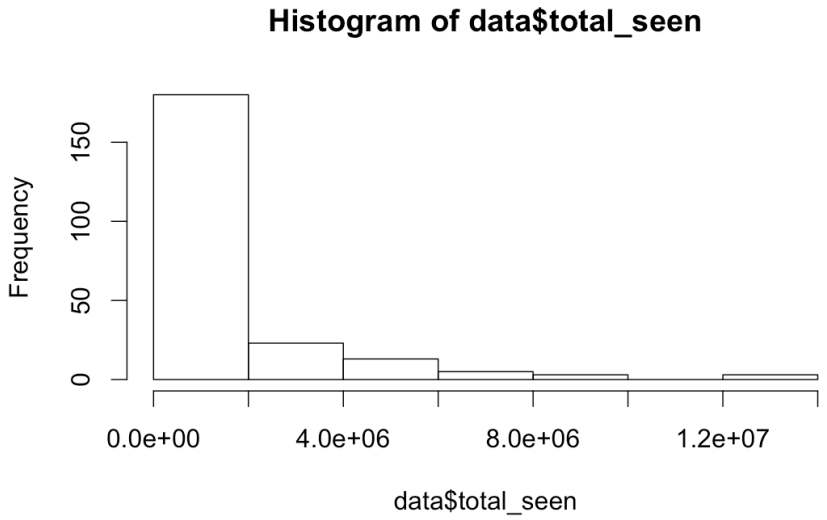
- 회귀분석과 마찬가지로 lm 명령어를 사용
- 단, 설명변수가 그룹을 정의하는 범주형변수
- Factor 함수를 사용하여 설명변수가 범주형변수라고 정의

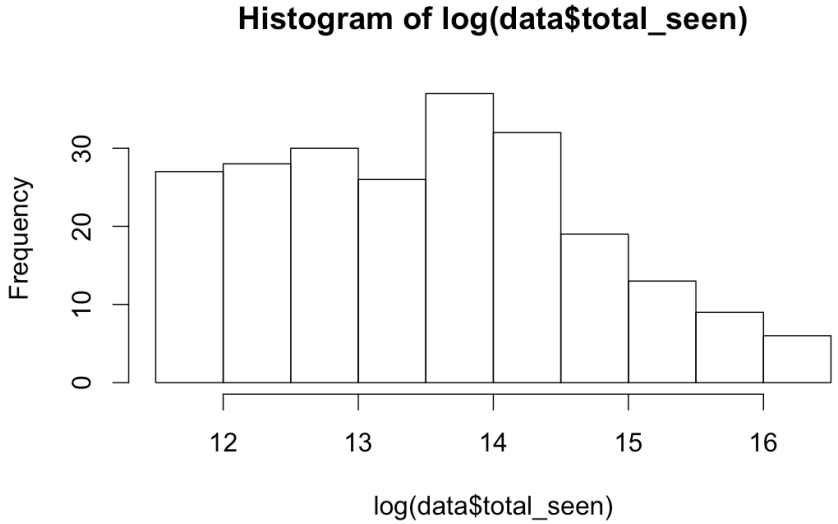
예: 등급별 영화흥행

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

```
> levels(data$rating)
[1] "12세이상관람가" "15세이상관람가" "전체관람가"      "청소년관람불가"
> library(psych)
> describeBy(data$total_seen,group=data$rating, mat=TRUE)
   item         group1 vars  n      mean        sd  median   trimmed       mad    min      max    range     skew
11    1 12세이상관람가    1 43 1774489.7 2069031.5  893027 1390505.8 1082776.9 101351  9001312  8899961 1.663481
12    2 15세이상관람가    1 94 2095732.6 2824207.5 1045561 1477272.3 1139430.7 101425 12983330 12881905 2.211749
13    3     전체관람가    1 48  638541.3  532817.5  343360  571803.5  269161.6 106432  2080445  1974013 1.014964
14    4 청소년관람불가    1 42 1015156.6 1133648.8  493634  803595.4  492685.0 113848  4720050  4606202 1.735021
      kurtosis        se
11  2.32459613 315524.35
12  4.75169040 291294.76
13 -0.04105884  76905.58
14  2.78575581 174925.82
```

- 종속변수 변환의 필요성


Histogram of data$total_seen
Frequency
0
50
100
150
0.0e+00
4.0e+06
8.0e+06
1.2e+07
data$total_seen


Histogram of log(data$total_seen)
Frequency
0
10
20
30
12
13
14
15
16
log(data$total_seen)

```
> out=lm(log(total_seen)~rating,data)
> summary(out)

Call:
lm(formula = log(total_seen) ~ rating, 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 ***
rating15세이상관람가   0.06818    0.21706   0.314  0.75375
rating전체관람가      -0.68294    0.24756  -2.759  0.00628 **
rating청소년관람불가  -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

F-test: 평균 차이의 검정

- 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

- 영화 등급별로 관객수의 유의한 차이가 있다 (F-test 의 결론)

➔ 그렇다면 어떤 등급 간의 차이가 있나?

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           13.70416    0.17979  76.221  < 2e-16 ***
rating15세이상관람가   0.06818    0.21706   0.314  0.75375
rating전체관람가      -0.68294    0.24756  -2.759  0.00628 **
rating청소년관람불가  -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 또는 Tukey 방법 사용!**

다중비교: Dunnett Method

- Reference level과의 각 범주의 평균 차이 검정

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

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = log(total_seen) ~ rating, 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   0.0167 *
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.

다중비교: Tukey Method

- 모든 쌍의 범주에 대해 평균 차이 검정

(전체관람가, 12세 이상 관람가)
(전체관람가, 15세 이상 관람가)
사이에 유의한 차이가 있다.

```
> Tukey=glht(out,linfct=mcp(rating="Tukey"))
> summary(Tukey)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = log(total_seen) ~ rating, data = data)

Linear Hypotheses:
                                          Estimate Std. Error t value Pr(>|t|)
15세이상관람가 - 12세이상관람가 == 0  0.06818    0.21706   0.314  0.98911
전체관람가 - 12세이상관람가 == 0      -0.68294    0.24756  -2.759  0.03096 *
청소년관람불가 - 12세이상관람가 == 0 -0.43123    0.25578  -1.686  0.33066
전체관람가 - 15세이상관람가 == 0      -0.75111    0.20916  -3.591  0.00223 **
청소년관람불가 - 15세이상관람가 == 0 -0.49941    0.21882  -2.282  0.10384
청소년관람불가 - 전체관람가 == 0       0.25170    0.24911   1.010  0.74125
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

추정된 회귀식

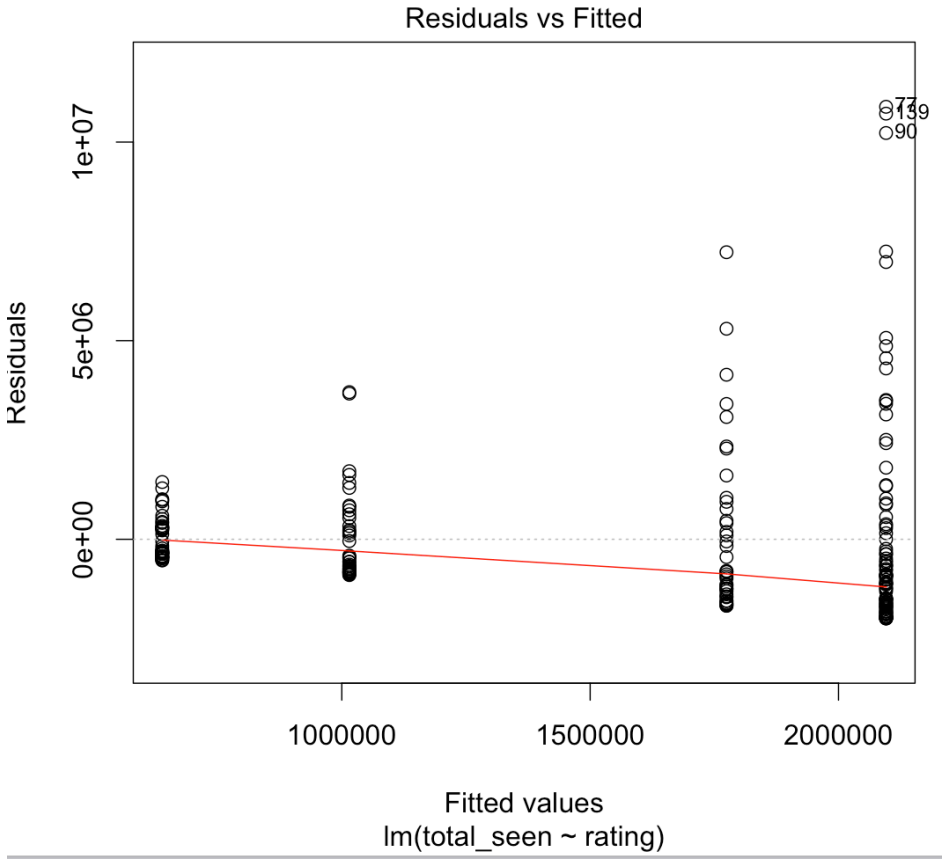
- 회귀추정식

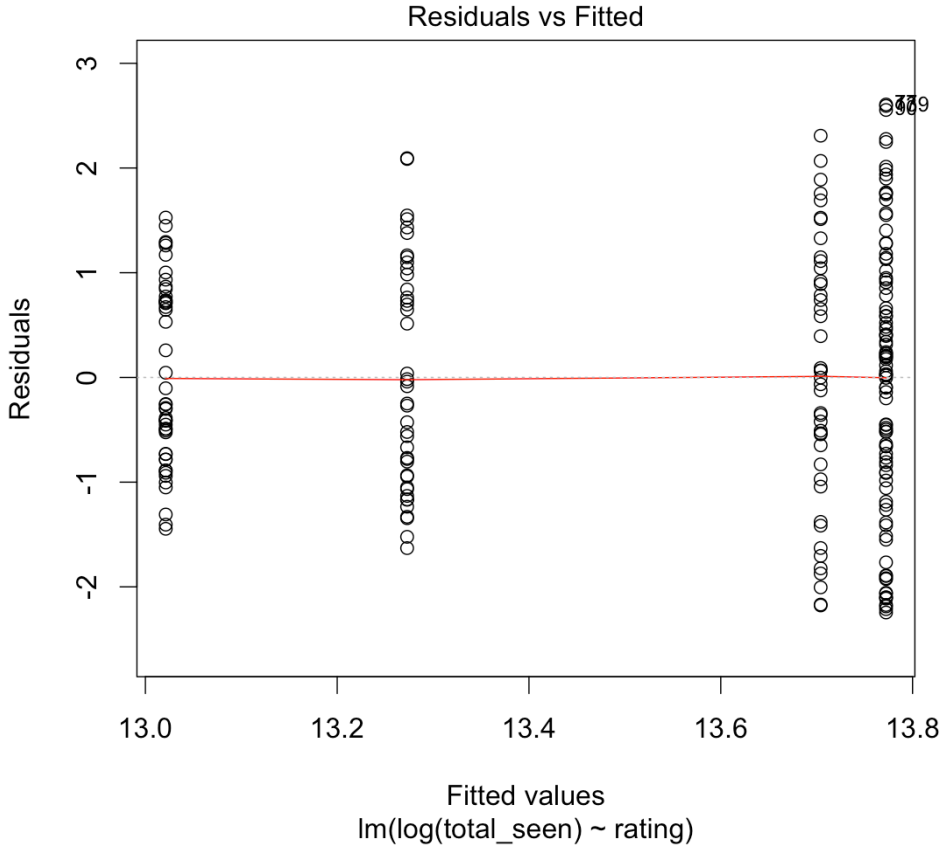
$$\hat{y} = 13.70 + 0.068x_1 - 0.68x_2 - 0.43x_3$$

  - 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)

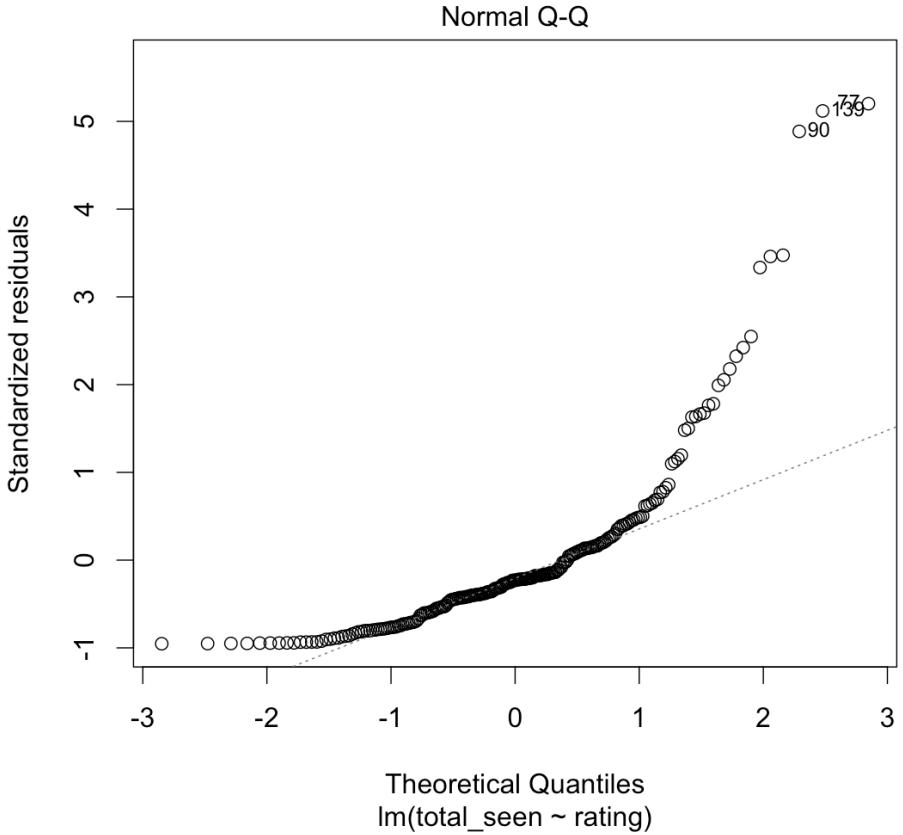
회귀진단

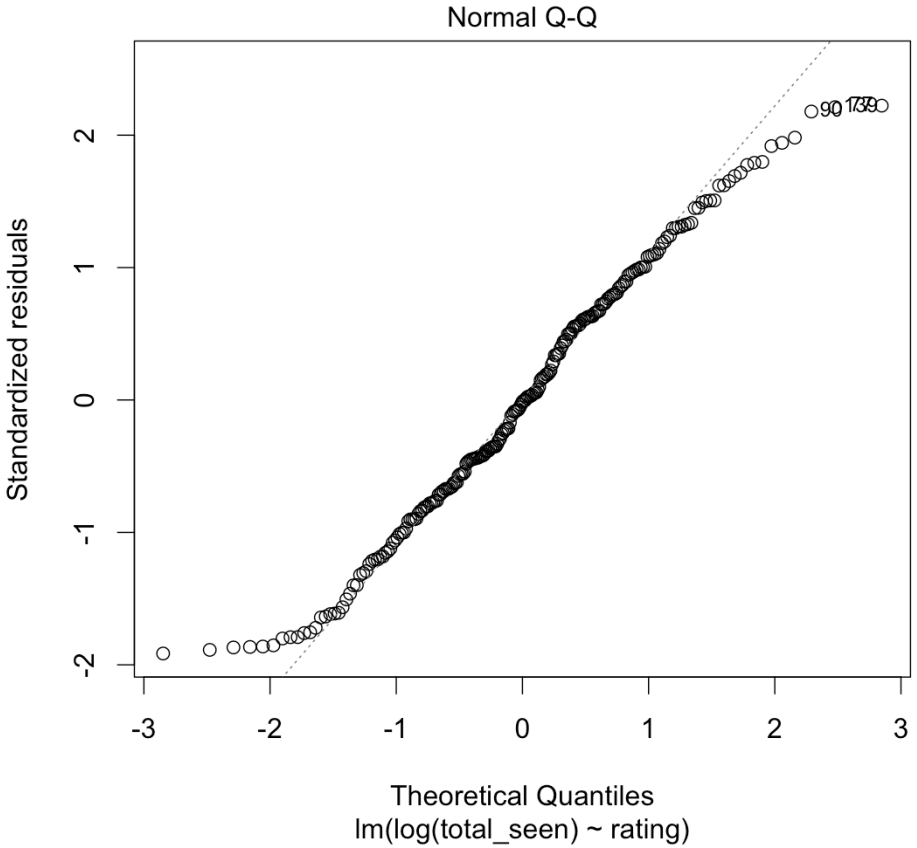
- 종속변수의 변환 전 보다 후에 잔차도가 안정됨





- 종속변수의 변환 전 보다 후에 잔차가 정규분포에 더 가까움





범주형 변수의 변환

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

```
> data$rating2=data$rating
> levels(data$rating2)
[1] "12세이상관람가" "15세이상관람가" "전체관람가"      "청소년관람불가"
> levels(data$rating2)=c(2,2,1,3)
> describeBy(data$total_seen,group=data$rating2,mat=TRUE)
   item group1 vars   n      mean        sd median   trimmed       mad    min      max    range     skew    kurtosis
11    1      2    1 137 1994904.5 2607432.6 978413 1446102.9 1116799.6 101351 12983330 12881979 2.232542  5.20855569
12    2      1    1  48  638541.3  532817.5 343360  571803.5  269161.6 106432  2080445  1974013 1.014964 -0.04105884
13    3      3    1  42 1015156.6 1133648.8 493634  803595.4  492685.0 113848  4720050  4606202 1.735021  2.78575581
          se
11 222768.01
12  76905.58
13 174925.82
```

```
> out2=lm(log(total_seen)~rating2,data )
> summary(out2)

Call:
lm(formula = log(total_seen) ~ rating2, 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)  13.7509     0.1005 136.791  < 2e-16 ***
rating21     -0.7297     0.1974  -3.698 0.000274 ***
rating23     -0.4780     0.2075  -2.303 0.022176 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 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
```

Reference level을 rating=1로 변환

```
> data$rating2=relevel(data$rating2,ref="1")
> summary(lm(log(total_seen)~rating2,data ))

Call:
lm(formula = log(total_seen) ~ rating2, 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)  13.0212     0.1698  76.672  < 2e-16 ***
rating22      0.7297     0.1974   3.698 0.000274 ***
rating23      0.2517     0.2486   1.012 0.312411
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 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
```

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)
:통제할 변수

주요 관심
설명변수

- 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$Treat=relevel(data$Treat,ref="Cont")
> summary(data)
  Treat        Prewt           Postwt
 Cont:26   Min.   :70.00   Min.   : 71.30
 CBT :29   1st Qu.:79.60   1st Qu.: 79.33
 FT  :17   Median :82.30   Median : 84.05
           Mean   :82.41   Mean   : 85.17
           3rd Qu.:86.00   3rd Qu.: 91.55
           Max.   :94.90   Max.   :103.60
```

더미변수 생성시 Cont그룹을 레퍼런스로 하기위해 reference를 지정한다.

```
> 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 ‘ ’ 1
```

Prewt의 효과를 통제한 후
Treat 변수가 설명해주는 Y의 변동성에 대한 Test
P-value<0.05 ➔ 치료효과의 차이가 있다.

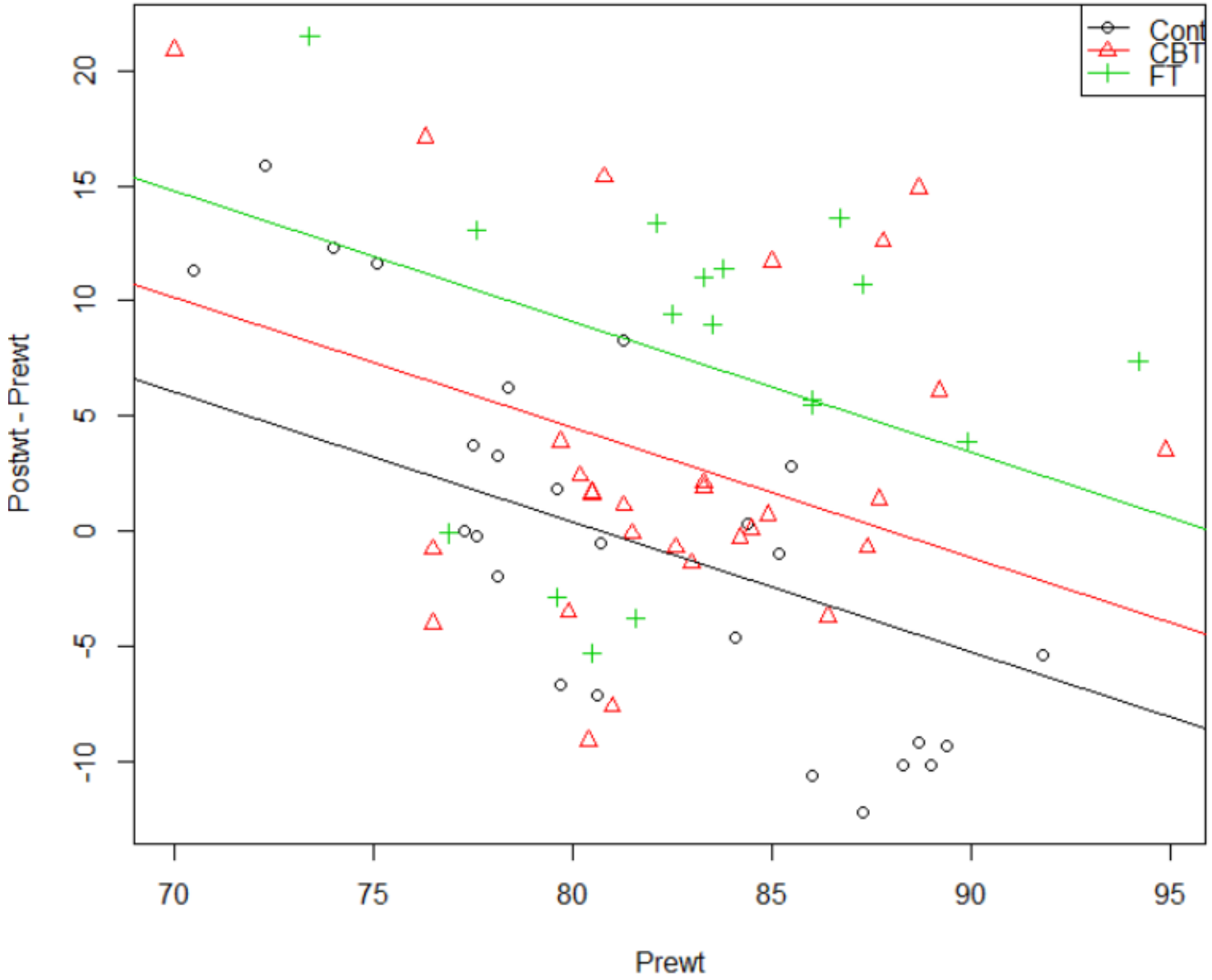
```
> summary(out)

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

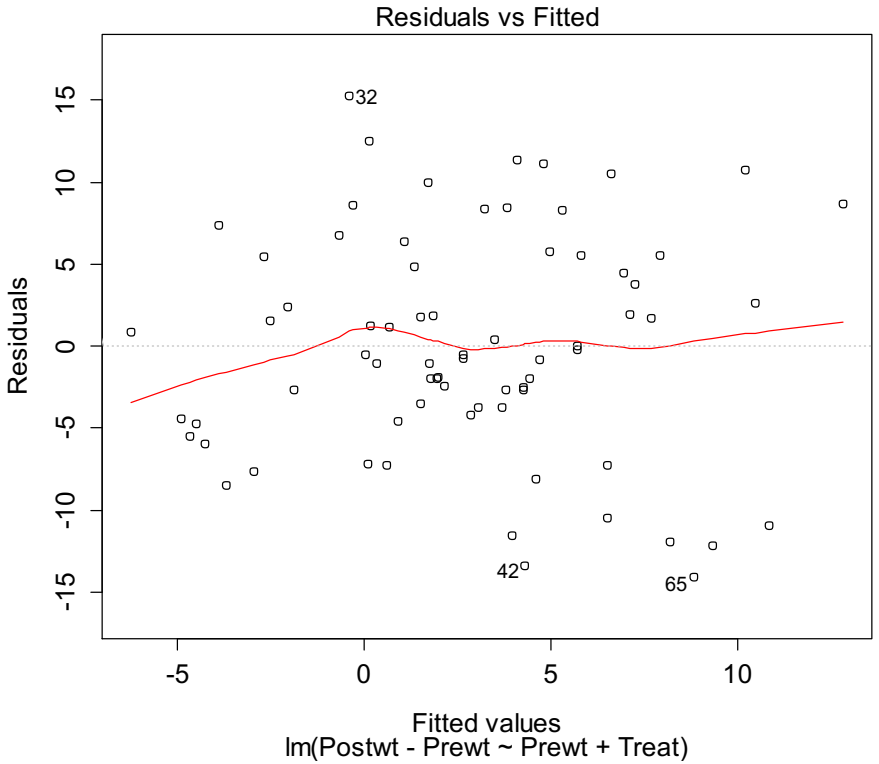
다중비교

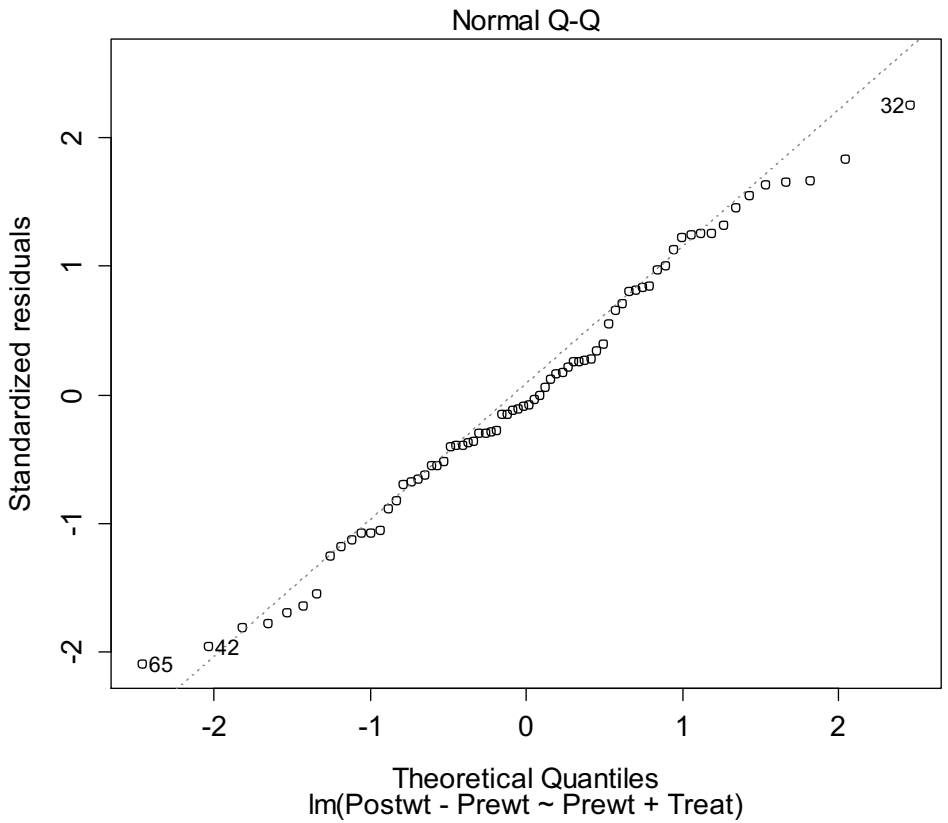
```
> dunnett=glht(out,linfct=mcp(Treat="Dunnett"))
> summary(dunnett)

         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)
```

```
> plot(Postwt-Prewt~Prewt,data,col=as.numeric(Treat),pch=as.numeric(Treat))
> abline(45.6740,-0.5655)
> abline(45.6740+4.0971,-0.5655,col=2)
> abline(45.6740+8.6601,-0.5655,col=3)
> legend("topright",c("Cont","CBT","FT"),col=1:3, pch=1:3,lty=1)
```



잔차분석


Residuals vs Fitted
Residuals
Fitted values
lm(Postwt - Prewt ~ Prewt + Treat)
32
42
65
-15
-10
-5
0
5
10
15
-5
0
5
10


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Postwt - Prewt ~ Prewt + Treat)
32
42
65
-2
-1
0
1
2

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: $\hat{y} = 45.67 + 8.66 - 0.57\ Prewt$

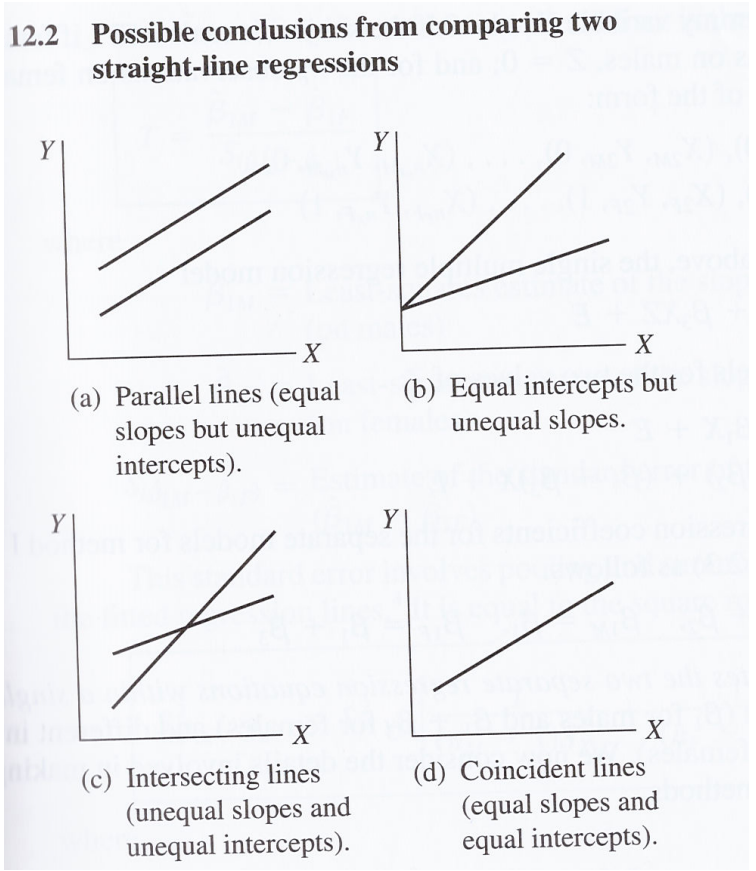
  - **Prewt이 평균이었던 사람에 대해** CBT는 control 그룹보다 4.10 만큼 더 몸무게 변화를 주었다.
  - **Prewt이 평균이었던 사람에 대해** FT는 control 그룹보다 8.66만큼 더 몸무게 변화를 주었다.

더미변수와 회귀분석

- 수축기혈압(Systolic blood pressure; SBP)과 연령 (age)을 40명의 남자와 29명의 여자로부터 기록 (SBP.csv)
- 연령이 높을 수록 수축기혈압이 높은 경향
- 연령과 혈압 사이의 관계가 남녀 간에 상이한가?

  – Y=SBP, X=AGE, Z=1 if female, 0 if male

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i$$



- 두 회귀선이 평행한가?
  - $H_0: \beta_3 = 0$

```
> model1=lm(SBP~AGE+SEX+AGE*SEX,SBP)
> summary(model1)

Call:
lm(formula = SBP ~ AGE + SEX + AGE * SEX, data = SBP)

Residuals:
    Min      1Q  Median      3Q     Max
-20.647  -3.410   1.254   4.314  21.153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 110.03853    4.73610  23.234  < 2e-16 ***
AGE           0.96135    0.09632   9.980 9.63e-15 ***
SEX         -12.96144    7.01172  -1.849   0.0691 .
AGE:SEX      -0.01203    0.14519  -0.083   0.9342
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.946 on 65 degrees of freedom
Multiple R-squared:  0.7759, Adjusted R-squared:  0.7656
F-statistic: 75.02 on 3 and 65 DF,  p-value: < 2.2e-16
```

- 두 회귀선이 동일한가?
  - $H_0: \beta_2 = \beta_3 = 0$

```
> summary(model3)

Call:
lm(formula = SBP ~ AGE, data = SBP)

Residuals:
    Min      1Q  Median      3Q     Max
-26.782  -7.632   1.968   8.201  22.651

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.34905    4.33190   23.86   <2e-16 ***
AGE           0.98333    0.08929   11.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.1 on 67 degrees of freedom
Multiple R-squared:  0.6441, Adjusted R-squared:  0.6388
F-statistic: 121.3 on 1 and 67 DF,  p-value: < 2.2e-16

> anova(model3)
Analysis of Variance Table

Response: SBP
          Df  Sum Sq Mean Sq F value    Pr(>F)
AGE        1 14951.3 14951.3  121.27 < 2.2e-16 ***
Residuals 67  8260.5   123.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> anova(model3,model1)
Analysis of Variance Table

Model 1: SBP ~ AGE
Model 2: SBP ~ AGE + SEX + AGE * SEX
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     67 8260.5
2     65 5201.4  2    3059.1 19.114 2.96e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANCOVA vs. 더미변수를 포함한 회귀분석

- ANCOVA
  - 주요 관심사는 집단 간 종속변수의 평균 차이
  - 공변량의 효과를 통제한 후 집단 간 차이를 파악하는 것이 목적
  - 공변량이 전체 평균인 수준에서 종속변수의 평균치를 비교
  - 각 집단의 회귀식이 평행하지 않으면 의미 없음

    ➔회귀선들이 평행한지에 대한 검정 후 귀무가설(회귀선이 평행하다)이 기각되지 않으면 ANCOVA 실시

- 더미변수를 포함한 회귀분석
  - 범주형 변수 뿐만 아니라 공변량도 관심대상
  - 만일 회귀선이 평행하지 않다면 해당 설명변수들 간에 **교호작용 (interaction effect)** 존재
  - 예) 나이가 어릴 때는 여자의 혈압이 더 높지만 나이가 들면 남자의 혈압이 높다