8강

모형개발

정보통계학과 김성수교수

학습목차

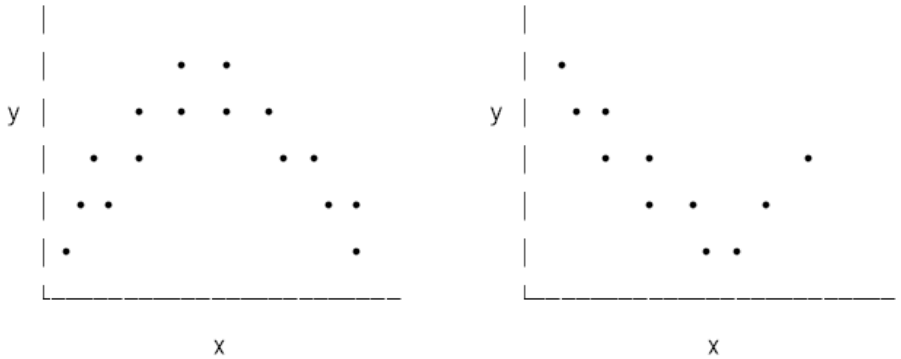
1. 다항회귀모형
2. 가변수 회귀모형

1 다항회귀모형

다항회귀모형

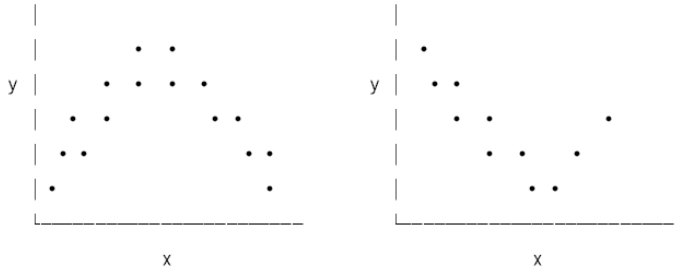
✓ 회귀모형

- 단순회귀모형 : 독립변수와 종속변수 간에 직선적인 관계가 있는 경우에 사용
- 다항회귀모형 : 독립변수와 종속변수 간에 직선적인 관계가 아니고, 다음과 같이 곡선관계가 있는 경우 사용



모형식

✓ 1) 독립변수가 하나인 경우의 이차다항회귀모형식



$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

R 적합
> lm(Y ~ X1 + I(X1^2), data=test)

⇒ 계수 β_2는 곡선의 오목, 볼록 상태를 결정

(β_2>0 :아래로 볼록, β_2<0 :위로 볼록)

※ 참고 : 산점도로부터 오목, 볼록 상태가 번갈아 나타나는 경우는
3차이상의 모형 사용

모형식

✓ 2) 독립변수가 두개인 경우의 이차다항회귀모형식

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \epsilon$$

* $X_1 X_2$ 항 : 교호작용 항 , β_{12} : 교호작용계수

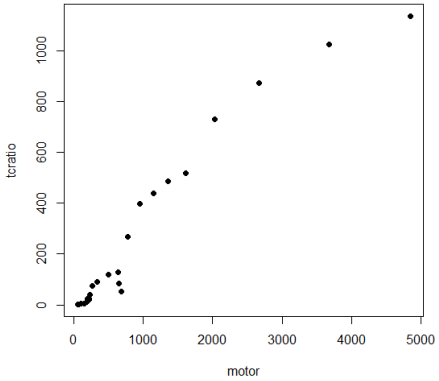
R 적합 :

```
> lm(Y~ X1+X2+I(X1^2) + I(X2^2) + X1:X2, data=test)
```

R 활용 예 1-①

〈 교통범죄자료 〉

연도	교통범죄 발생률	승용차 보급률	연도	교통범죄 발생률	승용차 보급률	연도	교통범죄 발생률	승용차 보급률
1966	0.9	59.5	1975	40.8	238.7	1984	438.6	1151.2
1967	2.0	77.1	1976	74.3	267.8	1985	487.2	1364.2
1968	4.1	107.4	1977	91.2	345.0	1986	519.1	1612.8
1969	5.6	159.5	1978	118.3	500.1	1987	730.0	2030.9
1970	12.0	188.2	1979	128.8	643.2	1988	874.1	2663.5
1971	23.6	205.5	1980	84.6	653.4	1989	1026.4	3677.8
1972	22.5	209.7	1981	51.1	691.1	1990	1138.1	4840.1
1973	24.2	229.7	1982	267.7	777.6			
1974	21.3	220.4	1983	398.1	954.6			




```
> tcrime = read.table("c:/data/reg/tcrime.txt", header=T)
> head(tcrime , 3)
  year tcratio motor
1 1966     0.9  59.5
2 1967     2.0  77.1
3 1968     4.1 107.4
> attach(tcrime)
> plot(motor, tcratio, pch=19)
>
```

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

⇒ β_1 : 선형효과계수 (linear effect coefficient),

β_2 : 곡선의 정도를 나타내는 2차효과계수

(quadratic effect coefficient)

(위로 볼록이므로 β_2<0 로 예상됨)

R 활용 예 1-②

```
> tcrime.lm = lm( tcratio ~ motor + I(motor^2), data=tcrime)
> summary(tcrime.lm)
…
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.450e+01  1.856e+01  -4.014 0.000583 ***
motor        4.539e-01  3.041e-02  14.930 5.40e-13 ***
I(motor^2)  -4.149e-05  6.873e-06  -6.036 4.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.8 on 22 degrees of freedom
Multiple R-squared:  0.9764,    Adjusted R-squared:  0.9742
F-statistic: 454.4 on 2 and 22 DF,  p-value: < 2.2e-16
```

이차다항회귀모형 : $\widehat{tcratio} = -74.5 + 0.4539 motor - 4.149 \times 10^{-5} motor^2$

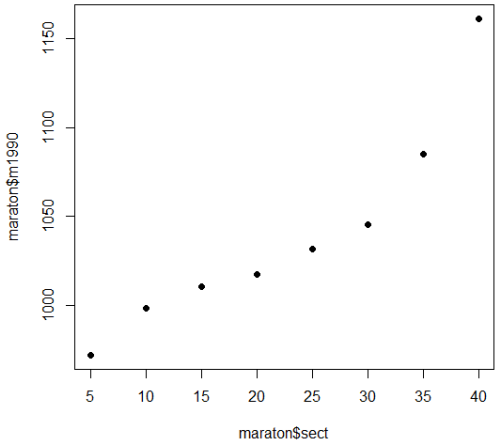
결정계수 R^2 = 0.9764

R 활용 예 2-①

〈1990-1992년 동경국제마라톤자료〉

연도	1990		1991		1992	
구간	평균	표준편차	평균	표준편차	평균	표준편차
5	971.9	47.6	971.9	38.9	966.1	40.0
10	998.3	55.6	974.8	47.2	979.7	56.5
15	1010.5	55.1	987.2	57.0	992.4	49.1
20	1017.7	44.8	977.6	60.0	996.8	50.5
25	1031.5	51.1	991.1	56.8	1009.8	61.6
30	1045.4	54.2	1006.0	66.2	1039.2	66.1
35	1085.2	79.8	1041.4	75.2	1073.0	83.4
40	1161.4	123.0	1136.0	142.1	1140.6	105.5

1990년부터 1992년까지의 동경국제마라톤대회에서 얻은 선수의 5km구간 별 평균속도(단위: 초)를 기록한 결과이다. 이 중 1990년 자료를 시간대 별로 그려 살펴보자.



$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

```
> maraton = read.table("c:/data/reg/maraton.txt", header=T)
> head(maraton,2)
  sect m1990 sd1990 m1991 sd1991 m1992 sd1992
1    5 971.9   47.6 971.9   38.9 966.1   40.0
2   10 998.3   55.6 974.8   47.2 979.7   56.5
> plot(maraton$sect, maraton$m1990, pch=19)
```

R 활용 예 2-②

```
> maraton.lm = lm(m1990 ~ sect+I(sect^2)+I(sect^3), data=maraton)
> summary(maraton.lm)
…
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 917.592857   8.083355 113.516 3.61e-08 ***
sect         13.785281   1.462847   9.424 0.000707 ***
I(sect^2)    -0.683225   0.073387  -9.310 0.000741 ***
I(sect^3)     0.012248   0.001077  11.375 0.000341 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.281 on 4 degrees of freedom
Multiple R-squared:  0.9983,    Adjusted R-squared:  0.9969
F-statistic: 761.4 on 3 and 4 DF,  p-value: 5.726e-06
```

적합된 3차 다항회귀모형식 :

$$\widehat{m1990} = 917.593 + 13.785 \times sect - 0.683 \times sect^2 + 0.012 \times sect^3$$

2 가변수 회귀모형

가변수를 이용한 회귀모형

✓ **독립변수에 이산형 변수가 포함되어 있는 경우에 사용**

(자료 예)

<비누생산공정에서 비누부스러기 부산물의 양과 공정속도>

1번 생산공정 ($D=1$)			2번 생산공정 ($D=0$)		
부산물의 양	공정속도		부산물의 양	공정속도	
Y	X	D	Y	X	D
218	100	1	140	105	0
248	125	1	277	215	0
360	220	1	384	270	0
351	205	1	341	255	0
470	300	1	215	175	0
394	255	1	180	135	0
332	225	1	260	200	0
321	175	1	361	275	0
410	270	1	252	155	0
260	170	1	422	320	0
241	155	1	273	190	0
331	190	1	410	295	0
275	140	1			
425	290	1			
367	265	1			

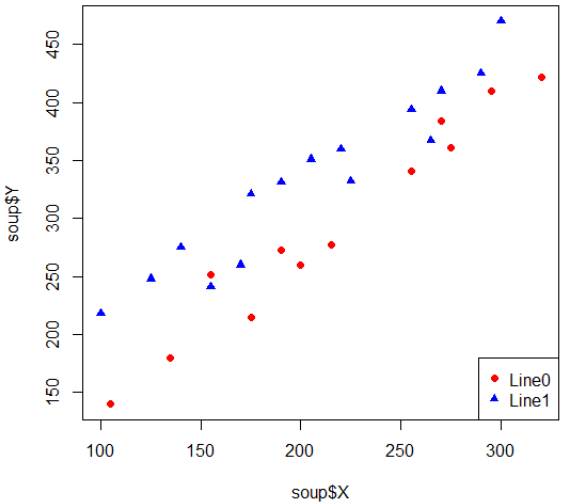
교호작용을 고려한 모형 :

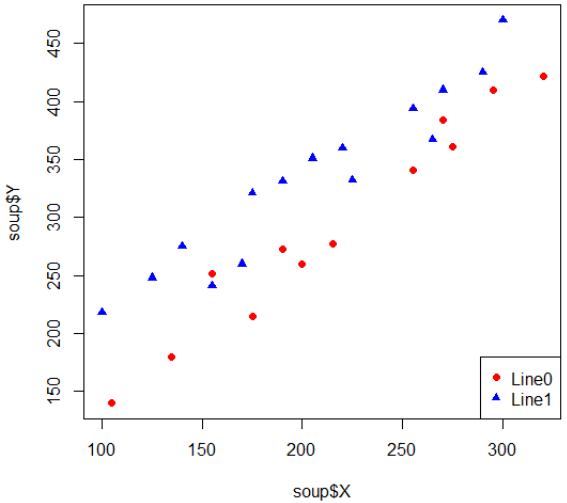
$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 XD + \epsilon$$

교호작용이 없는 모형 :

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \epsilon$$

```
> soup = read.table("c:/data/reg/soup.txt", header=T)
> soup[c(1,15,16,27),]
     Y   X D
1  218 100 1
15 367 265 1
16 140 105 0
27 410 295 0
> soup$D = factor(soup$D, levels=c(0,1), label=c("Line0", "Line1"))
> plot(soup$X, soup$Y, type="n")
> points(soup$X[soup$D=="Line1"], soup$Y[soup$D=="Line1"],
         pch=17, col="BLUE")
> points(soup$X[soup$D=="Line0"], soup$Y[soup$D=="Line0"],
         pch=19, col="RED")
> legend("bottomright", legend=levels(soup$D),
         pch=c(19,17), col=c("RED","BLUE"))
```





⇒ 두 생산라인은 기울기는 차이가 없고, 공정에 따라 차이가 있게 보임.

⇒ 교호작용이 없는 모형을 고려

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \epsilon$$

1번 생산라인 $(D=1)$: $Y = (\beta_0 + \beta_2) + \beta_1 X + \epsilon$

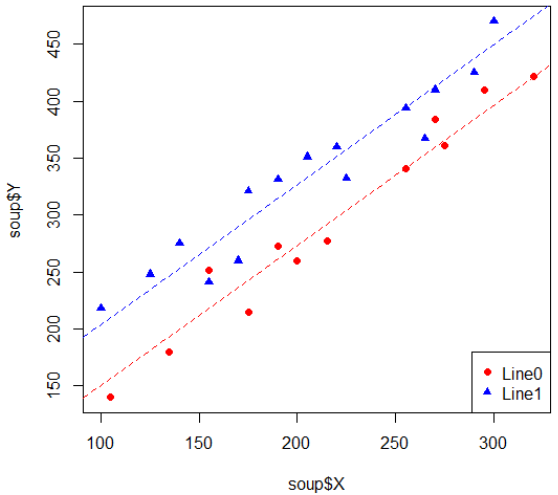
2번 생산라인 $(D=0)$: $Y = \beta_0 + \beta_1 X + \epsilon$

⇒ 두 생산라인은 β_2만큼 차이가 남.

⇒ $H_0 : \beta_2 = 0$을 기각하는 경우, 두 생산공정에 차이가 있다.

```
> soup.lm = lm(Y ~ X+D, data=soup)
> summary(soup.lm)
…
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.28179   15.40701   1.771   0.0893 .
X            1.23074    0.06555  18.775 7.48e-16 ***
DLine1      53.12920    8.21003   6.471 1.08e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
 ‘ ’ 1

Residual standard error: 21.13 on 24 degrees of freedom
Multiple R-squared:  0.9402,    Adjusted R-squared:  0.9352
F-statistic: 188.6 on 2 and 24 DF,  p-value: 2.104e-15
> abline(27.28179, 1.23074, lty=2, col="RED")
> abline(27.28179+53.1292, 1.23074, lty=2, col="BLUE")
```



적합된 회귀모형 : $\hat{Y} = 27.282 + 1.231X + 54.129D$

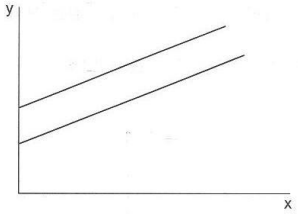
⇒ 따라서 기울기가 동일하다고 가정하는 경우의 회귀모형 적합에서 두 생산라인의 차이는 54.129 가 됨을 알 수 있음.

가변수를 이용한 회귀모형 : 교호작용을 고려한 경우

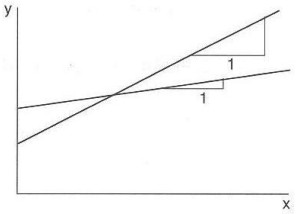
✓ **두 생산라인의 기울기가 다른 경우 : 교호작용을 고려한 모형을 이용**

교호작용을 고려한 모형 :

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 XD + \epsilon$$



기울기가 같은 모형



완전독립모형

⇒ 1번 공정과정 $(D=1)$: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \epsilon$

2번 공정과정 $(D=0)$: $Y = \beta_0 + \beta_1 X + \epsilon$

⇒ 기울기의 동질성여부 : H_0: $\beta_3 = 0$에 대한 F-검정을 통하여 수행.

가변수를 이용한 회귀모형 : 교호작용을 고려한 경우

```
> soup2.lm = lm(Y ~ X+D+X:D, data=soup)
> summary(soup2.lm)
…
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.57446   20.86970   0.363  0.71996
X            1.32205    0.09262  14.273 6.45e-13 ***
DLine1      90.39086   28.34573   3.189  0.00409 **
X:DLine1    -0.17666    0.12884  -1.371  0.18355
---
Residual standard error: 20.75 on 23 degrees of freedom
Multiple R-squared:  0.9447,    Adjusted R-squared:  0.9375
F-statistic: 130.9 on 3 and 23 DF,  p-value: 1.341e-14
```

⇒ 교호작용항의 경우, 회귀계수의 추정값은 -0.1767 이고, t_0 값에 대한 유의확률은 0.18355 로서 유의수준 0.05 보다 크므로 유의하지 않음을 알 수 있음.

⇒ 따라서 이 자료의 경우에는 교호작용을 고려하지 않은 모형으로 적합하는 것이 좋음.

가변수의 범주가 3개인 경우

✓ **가변수의 범주가 3개인 경우 : 두 개의 가변수를 고려**

- 범주의 수가 3개인 경우 가변수의 값을 지정하는 방법

범주	D_1	D_2	
A	1	0	
B	0	1	
C	0	0	기저범주(baseline category)

- 설명변수(X)의 개수가 하나이고 범주의 수가 세 개인 경우 (기저범주; base category = C 인 경우)

교호작용이 포함된 모형

$$Y = \beta_0 + \beta_1 X + \beta_2 D_1 + \beta_3 D_2 + \beta_4 X D_1 + \beta_5 X D_2 + \epsilon$$

교호작용이 없는 모형

$$Y = \beta_0 + \beta_1 X + \beta_2 D_1 + \beta_3 D_2 + \epsilon$$

〈교호작용이 포함된 모형의 해석〉

- 범주 A의 경우(D_1=1 , D_2=0)

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X + \epsilon$$

- 범주 B의 경우(D_1=0 , D_2=1)

$$Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X + \epsilon$$

- 범주 C의 경우(D_1=0 , D_2=0)

$$Y = \beta_0 + \beta_1 X + \epsilon$$

R 활용 예 : Multilevel factor–①

- Data: come from a study on the sexual activity and the life span of male by Partridge and Farquhar(1981); 125 fruitflies were divided randomly into five groups of 25 each. The response was the longevity of the fruitfly in days.
- Five groups :
  - solitary
  - kept individually with a virgin female each day
  - given eight virgin females per day
  - kept with one pregnant females per day
  - kept with eight pregnant females per day

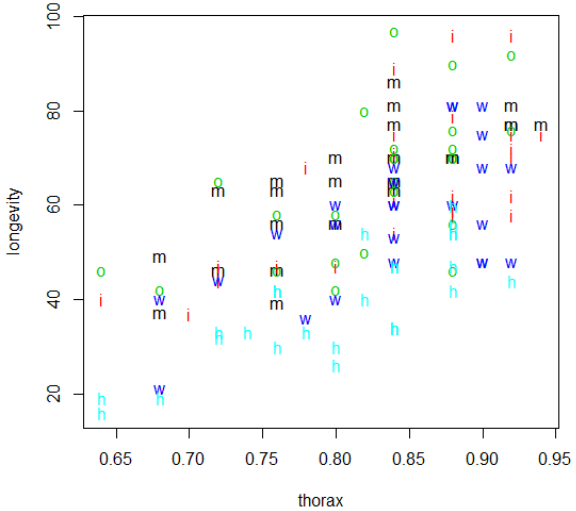
```
> library(faraway)
> data(fruitfly)
> fruitfly[c(1,26,51,75,101), ]
    thorax longevity activity
1     0.68        37     many
26    0.70        37 isolated
51    0.68        42      one
75    0.68        21      low
101   0.64        19     high
```

참고서적 : Faraway, J.J.(2002), Practical Regression and Anova Using R

R 활용 예 : Multilevel factor-②

◆ **Plot of Data**

```
> library(faraway)
> data(fruitfly)
> fruitfly[c(1,26,51,75,101), ]
    thorax longevity activity
1     0.68        37     many
26    0.70        37 isolated
51    0.68        42      one
75    0.68        21      low
101   0.64        19     high
> attach(fruitfly)
> plot(thorax, longevity, type="n")
> points(thorax[activity=="many"], longevity[activity=="many"], pch="m", col=1)
> points(thorax[activity=="isolated"], longevity[activity=="isolated"], pch="i", col=2)
> points(thorax[activity=="one"], longevity[activity=="one"], pch="o", col=3)
> points(thorax[activity=="low"], longevity[activity=="low"], pch="w", col=4)
> points(thorax[activity=="high"], longevity[activity=="high"], pch="h", col=5)
```



R 활용 예 : Multilevel factor–③

◆ **Fit general linear model considering interaction effects**

```
> g = lm(longevity ~ thorax * activity, data=fruitfly)
> summary(g)
…
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -50.2420    21.8012  -2.305    0.023 *
thorax             136.1268    25.9517   5.245 7.27e-07 ***
activityone          6.5172    33.8708   0.192    0.848
activitylow         -7.7501    33.9690  -0.228    0.820
activitymany        -1.1394    32.5298  -0.035    0.972
activityhigh       -11.0380    31.2866  -0.353    0.725
thorax:activityone  -4.6771    40.6518  -0.115    0.909
thorax:activitylow   0.8743    40.4253   0.022    0.983
thorax:activitymany  6.5478    39.3600   0.166    0.868
thorax:activityhigh -11.1268    38.1200  -0.292    0.771
---
Residual standard error: 10.71 on 114 degrees of freedom
Multiple R-squared:  0.6534,    Adjusted R-squared:  0.626
F-statistic: 23.88 on 9 and 114 DF,  p-value: < 2.2e-16
```

- **Reference level : isolated**
- **Fitted regression line for "isolated",**

$$longevity = -50.2 + 136.1 * thorax$$

- **For "one"**

$$longevity = (-50.2 + 6.5) + (136.1 - 4.7) * thorax$$

R 활용 예 : Multilevel factor–④

◆ **ANOVA table**

```
> anova(g)
Analysis of Variance Table

Response: longevity
                Df  Sum Sq Mean Sq F value    Pr(>F)
thorax           1 15003.3 15003.3 130.733 < 2.2e-16 ***
activity         4  9634.6  2408.6  20.988 5.503e-13 ***
thorax:activity  4    24.3     6.1   0.053    0.9947
Residuals      114 13083.0   114.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Interaction term thorax:activity is not significant indicating that we can fit the same slope within each group.**
- **We now refit without interaction term.**

R 활용 예 : Multilevel factor–⑤

◆ **Refit model without interaction term**

```
> gb = lm(longevity~thorax+activity, data=fruitfly)
> summary(gb)
…
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -48.749     10.850  -4.493 1.65e-05 ***
thorax        134.341     12.731  10.552  < 2e-16 ***
activityone     2.637      2.984   0.884   0.3786
activitylow    -7.015      2.981  -2.353   0.0203 *
activitymany    4.139      3.027   1.367   0.1741
activityhigh  -20.004      3.016  -6.632 1.05e-09 ***
---
Residual standard error: 10.54 on 118 degrees of freedom
Multiple R-squared:  0.6527,    Adjusted R-squared:  0.638
F-statistic: 44.36 on 5 and 118 DF,  p-value: < 2.2e-16
```

- "isolated" is the reference level.
- We see that "one" and "many" are not significantly different from "isolated"(see p-value).
- "low" group survives about seven days less.
- "high" group has a life span 20days less than the reference group.

다음시간 안내

9강. 자료진단