1. **다항회귀모형**

!!그래프가 곡선이지 비선형 회귀가 아님

비선형 회귀 예: $y = w_1 x^{w2}$

- 다항식을 사용한 선형회귀(항을 늘려 가짜 데이터 생성)

- 다항회귀모형에서 다중공선성 주의!!
  1개의 독립변수로 여러개의 변수를 회귀모델에 부여하므로
  (원독립변수에서 파생)

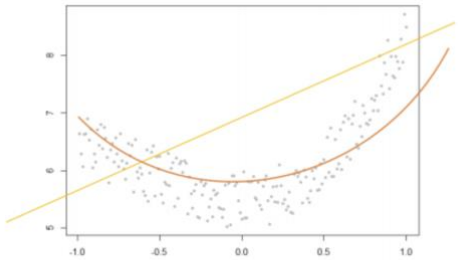
- **산점도**를 그려
  직선의 관계 => 단순회귀모형
  곡선의 관계 => 다항회귀모형(2차, 3차)
  n-1개의 굴절 관찰 시 n차 다항식
  $\beta^2 < 0$: 위로 볼록
  $\beta^2 > 0$: 아래로 볼록

- **Summary**를 통한 모델 선정 :
  -유의하지 않을 때까지 차수 높이는 방법

- **차수**를 높이면: 정확도 증가, 과적합 가능성

⇨ 일반적으로 3차항 초과 모델링하지 x

$$\hat{y} = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_p x_p^p$$



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_1^{\,2} + \cdots$$

포뮬러 심볼. 다양한 심볼을 이용하여 모델에 포함될 변수를 지정한다.

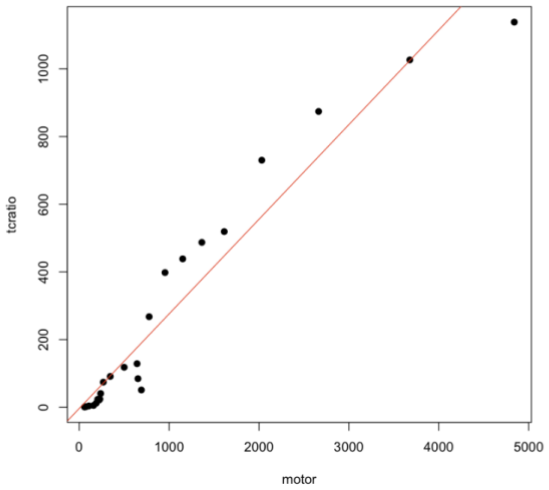
심볼	설명	예
:	독립변수 간의 상호작용을 나타낸다.	y ~ x + w + x:w에서 x:w는 x와 w 간의 상호작용항을 나타낸다.
*	독립변수 간의 모든 가능한 상호작용을 나타낸다.	y ~ x * w * z는 y ~ x + w + z + x:w + x:z + w:z + x:w:z와 같다.
^	지정한 차수까지의 상호작용을 나타낸다.	y ~ (x + w + z)^2은 y ~ x + w + z + x:w + x:z + w:z와 같다.
.	데이터셋에 포함된 종속변수를 제외한 다른 모든 변수를 나타낸다.	데이터셋에 x, y, w, z 변수가 포함되어 있을 때 y ~ .는 y ~ x + w + z와 같다.
-	변수를 제외한다.	y ~ (x + w + z)^2 - w:z는 y ~ x + w + z + x:w + x:z와 같다.
I()	괄호 안의 수식을 산술적으로 해석한다.	y ~ x + I((z + w)^2)은 y ~ x + u(여기에서 u는 z와 w의 합을 제곱한 새로운 변수)와 같다.

출처: https://youtu.be/53N4NQ1bgCA

In [10]:
```
#선형회귀로 적합시
tcrime1.lm=lm(tcratio~motor, data=tcrime)
summary(tcrime1.lm)
```

```
Call:
lm(formula = tcratio ~ motor, data = tcrime)

Residuals:
    Min      1Q  Median      3Q     Max
-211.36  -36.51  -17.99   53.84  165.80

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.50235   22.88871  -0.153     0.88
motor        0.27953    0.01508  18.531 2.53e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.95 on 23 degrees of freedom
Multiple R-squared:  0.9372,    Adjusted R-squared:  0.9345
F-statistic: 343.4 on 1 and 23 DF,  p-value: 2.533e-15
```

In [13]:
```
plot(motor,tcratio,pch=19)
abline(tcrime1.lm, col="red")
```



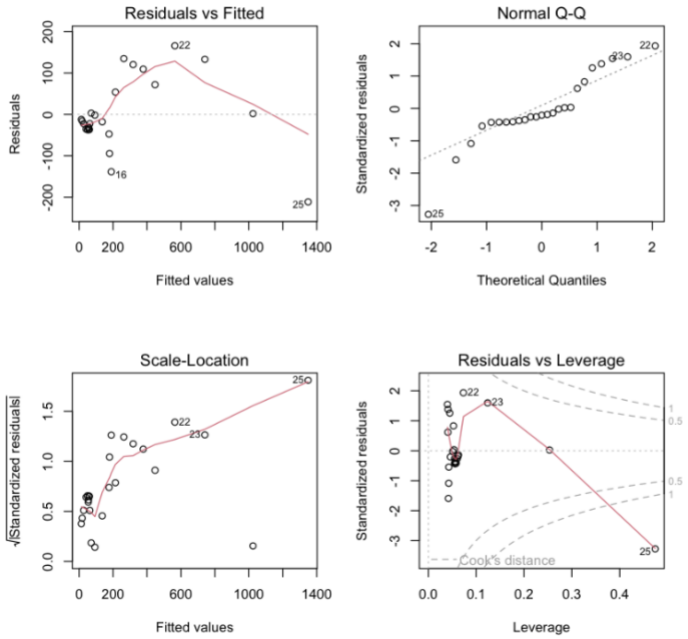

```
#다항회귀로 적합시
tcrime2.lm=lm(tcratio~motor+I(motor^2), data=tcrime)
summary(tcrime2.lm)
```

```
Call:
lm(formula = tcratio ~ motor + I(motor^2), data = tcrime)

Residuals:
     Min       1Q   Median       3Q      Max
-168.304   -7.443    8.754   30.328   77.072

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: trcatio = -74.5 + 0.4539motor - 4.149 x 10^-5 motor^2
97.64%의 설명력 유의하다.

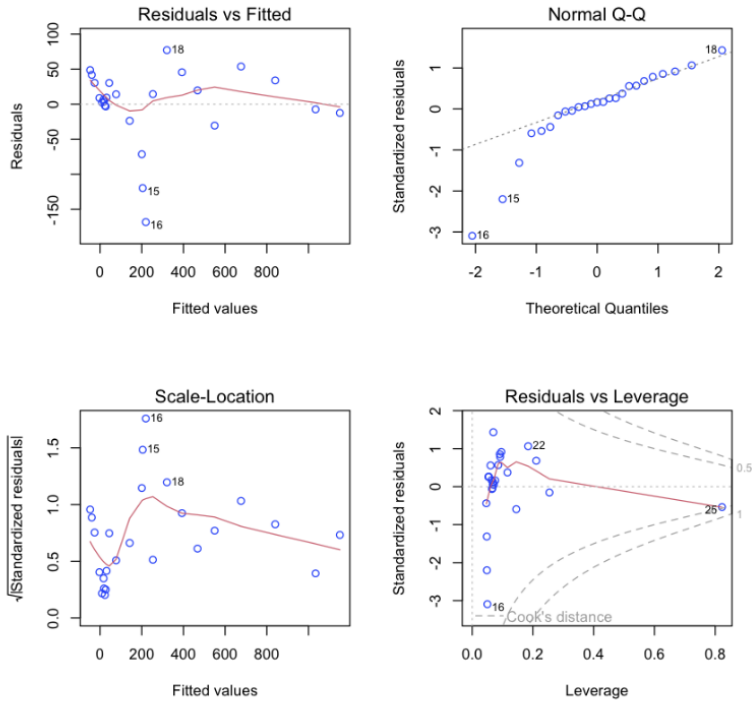
잔차진단 등분산성: 좌측 상하/정규성: Normal Q-Q(정규분포와 잔차의 분포 비교)/ 영향점: 우측 하단

```
#선형회귀
par(mfrow=c(2,2))
plot(tcrime1.lm)
```

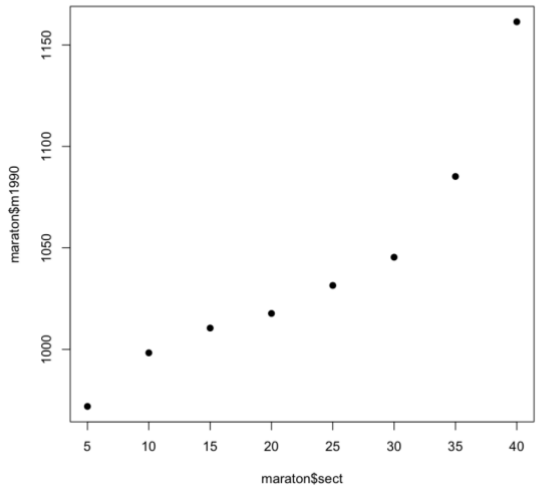


(참고: https://rstudio-pubs-static.s3.amazonaws.com/190997_40fa09db8e344b19b14a687ea5de914b.html)

```
#다항회귀
par(mfrow=c(2,2))
plot(tcrime2.lm, col="blue")
```



```r
#3차 다항회귀모형 적합
plot(maraton$sect, maraton$m1990, pch=19)
```



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

```r
maraton.lm = lm(m1990~sect+I(sect^2)+I(sect^3), data=maraton)
summary(maraton.lm)
```

```
Call:
lm(formula = m1990 ~ sect + I(sect^2) + I(sect^3), data = maraton)

Residuals:
      1       2       3       4       5       6       7       8
 0.9303 -1.0716 -1.4851 -0.2963  4.9082 -1.5578 -3.0807  1.6530

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

m1990 = 917.593 + (13.785 x sect) - (0.683 x sect^2)+(0.012 x sect^3)

p-value 모두 유의, 설명력 99.83

2. 가변수(= 지시변수 = dummy 변수) 를 이용한 회귀모형

- 범주형 변수를 연속형처럼 숫자로 변환 :
  분석기법이 연속형 변수로만 사용 가능할 때
  (ex) 선형회귀, 로지스틱회귀

- 0 또는 1의 값

- 더미변수의 수 = 범주의 개수 -1

	Gender	Age	Income	Spending_Score
	<chr>	<int>	<int>	<int>
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76

0

]:
```
summary(customer)
```

```
    Gender              Age             Income       Spending_Score
 Length:200         Min.   :18.00   Min.   : 15.00   Min.   : 1.00
 Class :character   1st Qu.:28.75   1st Qu.: 41.50   1st Qu.:34.75
 Mode  :character   Median :36.00   Median : 61.50   Median :50.00
                    Mean   :38.85   Mean   : 60.56   Mean   :50.20
                    3rd Qu.:49.00   3rd Qu.: 78.00   3rd Qu.:73.00
                    Max.   :70.00   Max.   :137.00   Max.   :99.00
```

]:
```
#Gender를 가변수 처리
customer$Gender <- ifelse(customer$Gender == "Male", 1, 0)
head(customer)
```

A data.frame: 6 × 4

	Gender	Age	Income	Spending_Score
	<dbl>	<int>	<int>	<int>
1	1	19	15	39
2	1	21	15	81
3	0	20	16	6
4	0	23	16	77
5	0	31	17	40
6	0	22	17	76

기준범주 : 생략되는 범주 (하기 표의 C) 더미변수값이 모두 0인 범주
회귀식 결과로 회귀계수 해석하는 기준이 됨

(참고_더미설정: https://blog.naver.com/statstorm/222012419025)

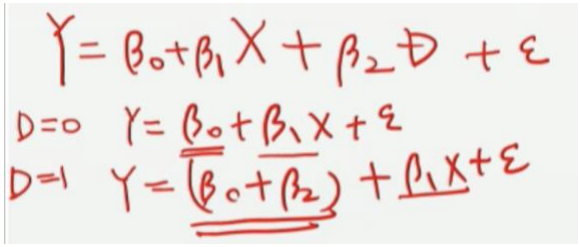
예) 계절별 평균 매출액

계절변수	D1	D2	D3
봄	0	0	0
여름	1	0	1
가을	0	1	0
겨울	0	0	1

$$\text{매출액} = \beta_0 + \beta_1 \times \text{계절}$$ ☹

$$\text{매출액} = \beta_0 + \beta_1 \times D_1 + \beta_2 \times D_2 + \beta_3 \times D_3$$ ☺

X : 공정속도, Y: 부산물의 양/ D=1 1번생산공정, D=0 2번생산공정
*교호작용이 없는 경우: 기울기 차이x, 절편의 차이 O




```
#비누 생산공정에서 부산물인 비누 부스러기 양과 공정속도

soap = read.table("/Users/sson/Desktop/Practice/Regression model/reg2020/soap.txt",header=T)
soap[c(1,15,16,27),]
```

A data.frame: 4 × 3

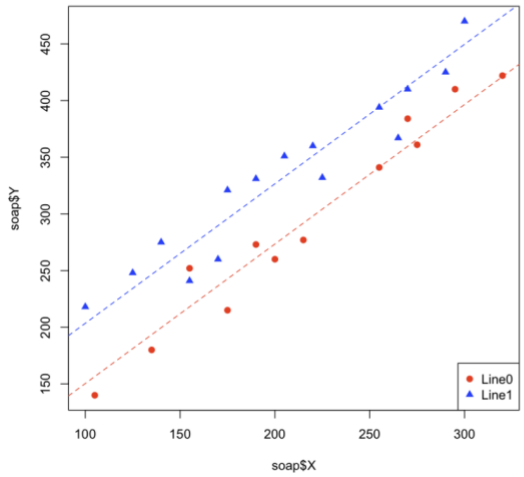
	Y	X	D
	<int>	<int>	<int>
1	218	100	1
15	367	265	1
16	140	105	0
27	410	295	0

```
soap$D = factor(soap$D, levels=c(0,1), label=c("Line0","Line1")) #D변수를 인자변수로 인지시키기
soap[c(1,15,16,27),]
```

A data.frame: 4 × 3

	Y	X	D
	<int>	<int>	<fct>
1	218	100	Line1
15	367	265	Line1
16	140	105	Line0
27	410	295	Line0

```r
plot(soap$X, soap$Y, type="n")     #type=n은 아직 그리지 말고 틀만 가지고 있어라
points(soap$X[soap$D =="Line1"], soap$Y[soap$D == "Line1"], pch=17, col="BLUE")
points(soap$X[soap$D =="Line0"], soap$Y[soap$D == "Line0"], pch=19, col="RED")
legend("bottomright", legend=levels(soap$D), pch=c(19,17), col=c("RED","BLUE"))


abline(27.28179, 1.23074, lty=2, col="RED")
abline(27.28179+53.1292, 1.23074, lty=2, col="BLUE")
```



두 생산라인의 차이(부산물의 양) 즉, β_2 = 53.129
생산라인 공정속도에 따라 부산물의 양에 차이가 있다.

```r
soap.lm=lm(Y~ X+D, data=soap)
summary(soap.lm)
```

```
Call:
lm(formula = Y ~ X + D, data = soap)

Residuals:
    Min      1Q  Median      3Q     Max
-39.557 -14.161  -0.121  17.518  33.953

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
```

#적합된 회귀모형 Y = 27.282 + 1.231X + 53.129
기울기가 동일하다고 가정하는 경우 회귀모형 적합에서 두 생산라인의 차이는 53.129

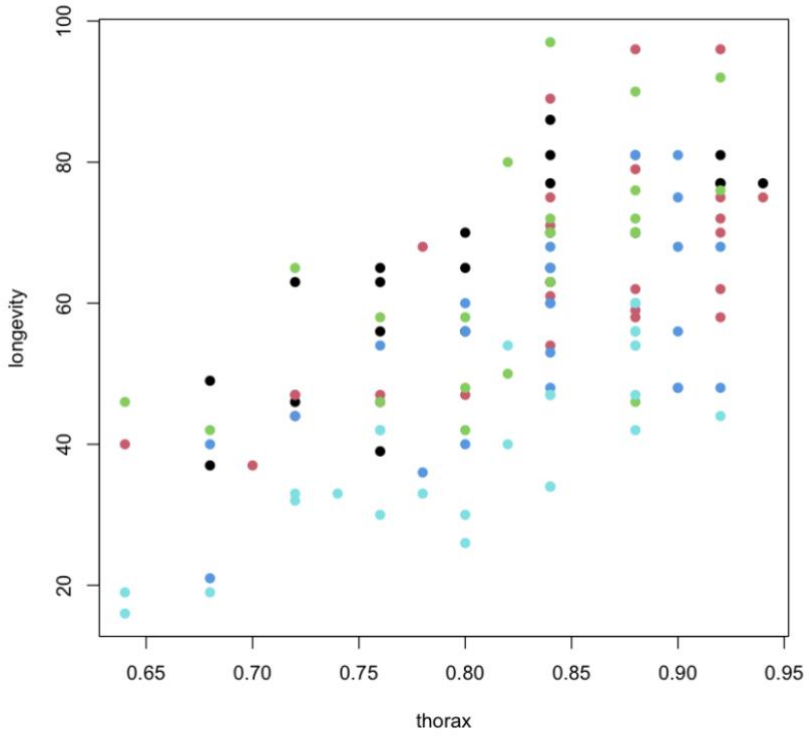
교호작용 고려한 모형 => 기울기와 절편 모두 차이 O

```
#교호작용 고려한 경우(X+D+X:D)
soap2.lm=lm(Y~X+D+X:D, data=soap)
summary(soap2.lm)
```

```
Call:
lm(formula = Y ~ X + D + X:D, data = soap)

Residuals:
   Min     1Q Median     3Q    Max
-34.50 -11.06   2.78  14.82  39.51

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 ***
D           90.39086   28.34573   3.189  0.00409 **
X:D         -0.17666    0.12884  -1.371  0.18355
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

```
# 교호작용항 x:DLine1의 경우, 회귀계수 추정값은 -0.1767이고 t0값에 대한 유의확률은 0.18355로 유의하지 않음
  교호작용 고려하지 않은 모형으로 적합해야 함
```

*가변수의 범주가 3개이상인 경우



```
g=lm(longevity~thorax*activity, data= fruitfly)
summary(g)
```

```
Call:
lm(formula = longevity ~ thorax * activity, data = fruitfly)

Residuals:
     Min       1Q   Median       3Q      Max 
-25.9509  -6.7296  -0.9103   6.1854  30.3071 

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    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
isolate : longevity = -50.2+136.1*thorax
one : longevity = (-50.2+6.5) + (136.1-5.7)*thorax
```

```
anova(g)
```

A anova: 4 × 5

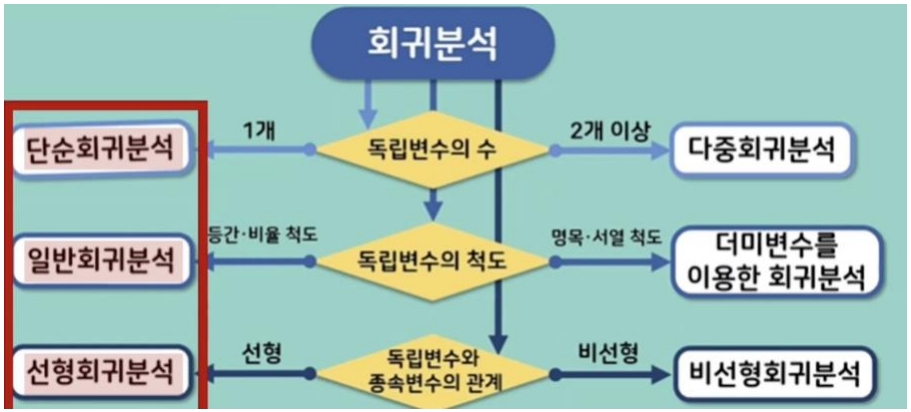
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
thorax	1	15003.30045	15003.300454	130.73288661	1.240497e-20
activity	4	9634.58753	2408.646883	20.98800599	5.503126e-13
thorax:activity	4	24.31359	6.078398	0.05296478	9.946914e-01
Residuals	114	13082.98391	114.763017	NA	NA

thorax와 activity 유의하지만 교호작용항은 유의하지 않다.
교호작용항 없이 모형 적합해야함

```
#교호작용 제외 재적합
gb=lm(longevity ~ thorax+activity, data=fruitfly)
summary(gb)
```

```
Call:
lm(formula = longevity ~ thorax + activity, data = fruitfly)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.108  -7.014  -1.101   6.234  30.265 

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

```
isolate = -48.749 + 134.341thorax
low = -7 일 덜 산다.
high = 20일 덜 산다.
```


회귀분석
독립변수의 수
1개
2개 이상
단순회귀분석
다중회귀분석
독립변수의 척도
등간·비율 척도
명목·서열 척도
일반회귀분석
더미변수를
이용한 회귀분석
독립변수와
종속변수의 관계
선형
비선형
선형회귀분석
비선형회귀분석