12강

일반화선형모형 (2)

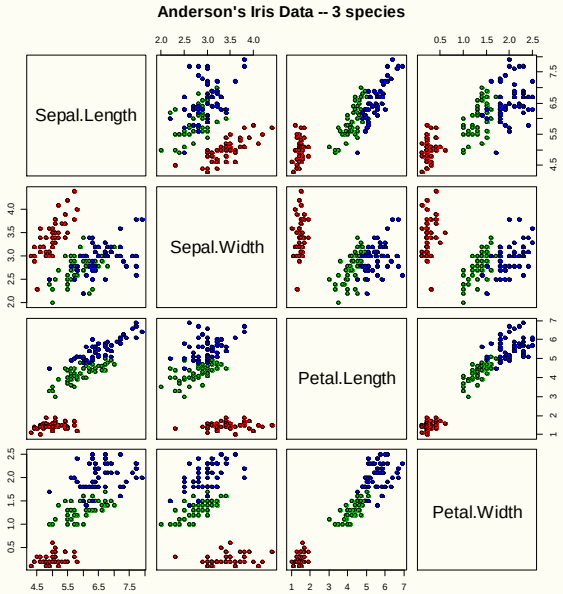
서강대학교 경영학과 이윤동교수

목 차

1 정규분포 경우의 예

정규분포 : Iris data

- E. Anderson(1936)에 의하여 수집
- 세 종류의 붓꽃(iris), Setosa, Versicolor, Virginica 각 50개 씩
- 5개의 변수: (꽃잎, 꽃받침) X (폭, 길이), 붓꽃의 종류



붓꽃자료

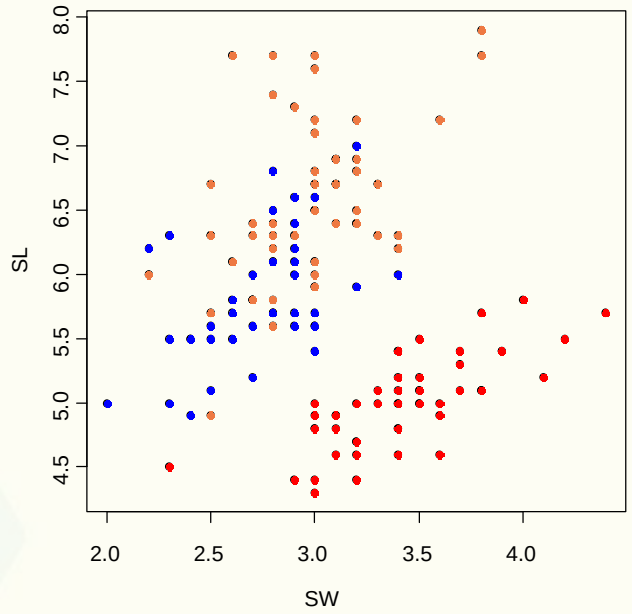
```
> names(iris)<- c("SL","SW","PL","PW","SP")
> levels(iris$SP)<-c("st","vc","vg")
> zip<- function(data,k=3) rbind(head(data,k), tail(data,k))
> zip(iris)
     SL  SW  PL  PW  SP
  1  5.1  3.5  1.4  0.2  st
  2  4.9  3.0  1.4  0.2  st
  3  4.7  3.2  1.3  0.2  st
148  6.5  3.0  5.2  2.0 vg
149  6.2  3.4  5.4  2.3 vg
150  5.9  3.0  5.1  1.8 vg
```

[R 6.1] 붓꽃자료

Sepal.Length → SL
Sepal.Width → SW
Species → SP

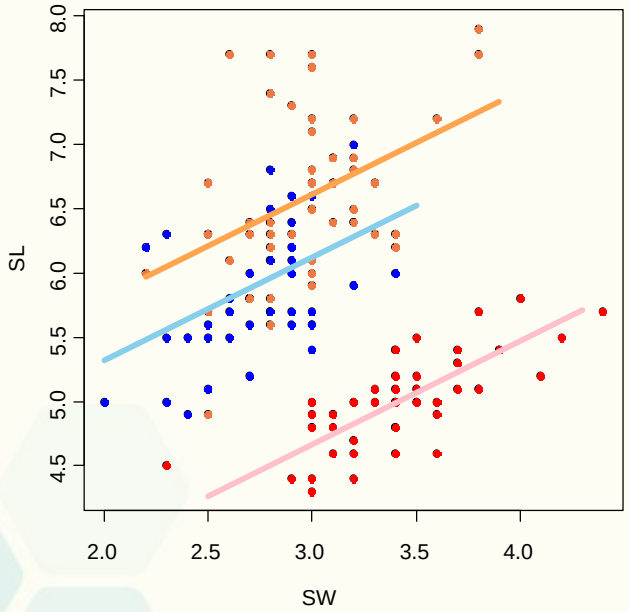
setosa → st
versicolor → vc
virginica → vg

꽃받침의 폭과 길이

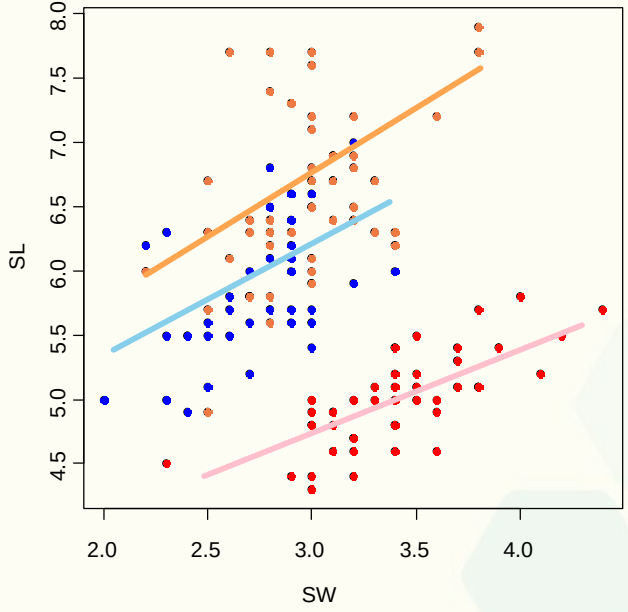


두 가지 모형

모형 A: 자유절편 **동일**기울기



모형 B: 자유절편 **자유**기울기



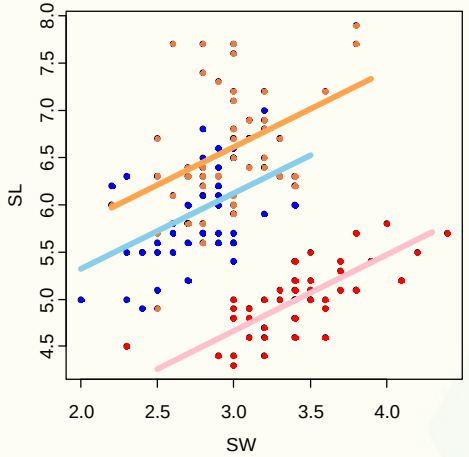
붓꽃 자료 : 모형 A

> lm(**SL~SP+SW-1** , data=iris)
glm(**SL~SP+SW-1** , data=iris) **동일한 표현**

....
Coefficients:

SPst	SPvc	SPvg	SW
2.2514	3.7101	4.1982	0.8036

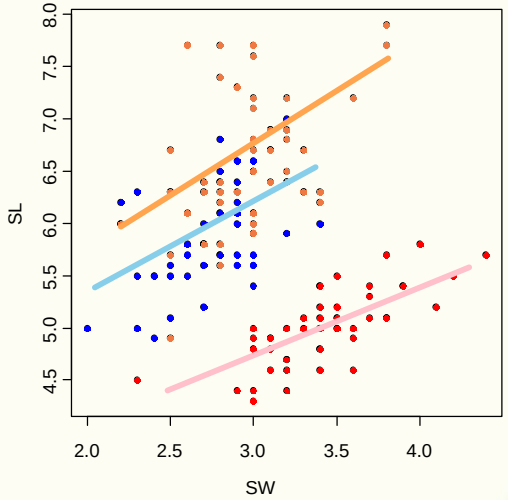
....
Residual Deviance: 28 **AIC: 183.9**



> lm(**SL~SP+SW** , data=iris)
....
Coefficients:

(Intercept)	SPvc	SPvg	SW
2.2514	1.4587	1.9468	0.8036

붓꽃 자료 : 모형 B



> lm(**SL~SP+SP/SW -1** , data=iris)
lm(**SL~ SP/SW -1**, data=iris)
glm(**SL~ SP/SW -1**, data=iris)
....
Coefficients:

SPst	SPvc	SPvg	SPst:SW	SPvc:SW	SPvg:SW
2.6390	3.5397	3.9068	0.6905	0.8651	0.9015

....
Residual Deviance: 28 **AIC: 187.09**

두 모형의 비교

	모형 A	모형 B
이탈도	28	27.8
2*음로그우도	173.93	173.09
잔차의 자유도	146	144
모형모수 개수	5	7
AIC	183.93	187.09

질문 & 답변

Q : **동일절편 자유기울기 모형 ?**

```
> summary( glm( SL~SP/SW -SP , data=iris) )
...
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  3.35789    0.33000      10.175    < 2e-16 ***
SPst:SW      0.48326    0.09683       4.991    1.69e-06 ***
SPvc:SW      0.92991    0.11976       7.765    1.32e-12 ***
SPvg:SW      1.08401    0.11166       9.708    < 2e-16 ***
....
AIC: 185.75
```

정규분포 : lm, glm 비교

```
> anova( glm( SL~SP/SW , data=iris))
Analysis of Deviance Table      이탈도분석표
Terms added sequentially (first to last)
....
         Df   Deviance    Resid. Df    Resid. Dev
NULL                          149     102.168
SP        2     63.212        147      38.956
SP:SW     3     11.110        144      27.846
```

```
> anova( lm( SL~SP/SW , data=iris))
Analysis of Variance Table      분산분석표
...
             Df  Sum Sq  Mean Sq  F value    Pr(>F)
SP            2  63.212  31.6061   163.44  < 2.2e-16 ***
SP:SW         3  11.110   3.7032    19.15   1.66e-10 ***
Residuals   144  27.846   0.1934
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

2 이항분포 경우의 예

이항분포 : 담배나방 자료

〈표 6.3〉 담배나방 독성실험 자료

	살충제 투여량					
암수	1	2	4	8	16	32
수컷	1	4	9	13	18	20
암컷	0	2	6	10	12	16

- Collett(1991,p75) 소개
- 담배나방에 대한 살충제의 살충효과 실험 자료
- 6개의 실험 챔버에, 각 챔버별로 암수 각각 20마리씩을 넣고
- 챔버별로 살충제 양을 달리하여 투여
- 사흘 후 **죽은 나방의 수**를 기록함

담배나방 자료 : 변수

〈표 6.3〉 담배나방 독성실험 자료

	살충제 투여량					
암수	1	2	4	8	16	32
수컷	1	4	9	13	18	20
암컷	0	2	6	10	12	16

- 종속변수 : 죽은나방의 수 (ndead)
  살아 있는 나방의 수(nalive)

  nalive = 20 − ndead

- 독립변수 : 로그투여량 (ldose)
  암수 (sex)

담배나방 자료의 준비

```
ldose <- rep(0:5, 2)
ndead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex <- factor(rep(c("M","F"),e=6))
nda<-cbind(ndead, alive=20-ndead)
xd<-data.frame(ldose, sex)
```

```
> xd
   ldose sex
1      0   M
2      1   M
3      2   M
4      3   M
5      4   M
6      5   M
7      0   F
8      1   F
9      2   F
10     3   F
11     4   F
12     5   F
```

```
> nda
      ndead alive
 [1,]     1    19
 [2,]     4    16
 [3,]     9    11
 [4,]    13     7
 [5,]    18     2
 [6,]    20     0
 [7,]     0    20
 [8,]     2    18
 [9,]     6    14
[10,]    10    10
[11,]    12     8
[12,]    16     4
```

담배나방 자료 : 두 모형

- 모형 A : 교호작용이 있는 모형

  (ndead, nalive) ~ log(dose) * sex

- 모형 B: 교호작용이 없는 모형

  (ndead, nalive) ~ log(dose) + sex

담배나방 자료 : 모형A

```
> ( bw.glm0<-glm( nda~sex*ldose , family=binomial(link=probit), data=xd ) )
...
> summary(bw.glm0)
```

모형 A : 교호작용 있는 모형

```
Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)  -1.80072    0.29832    -6.036   1.58e-09 ***
sexM          0.15479    0.41635     0.372   0.710
ldose         0.54523    0.09138     5.966   2.43e-09 ***
sexM:ldose    0.19165    0.14259     1.344   0.179
--- ....
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:   3.768  on  8  degrees of freedom
AIC: 41.878
```

담배나방 자료 : 모형B

```
> summary( update( bw.glm0, . ~sex+ldose ))
```

...

모형 B: 교호작용 없는 모형

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.06033	0.25087	-8.213	< 2e-16 ***
sexM	0.65364	0.20235	3.230	0.00124 **
ldose	0.63245	0.06975	9.068	< 2e-16 ***

--- ...

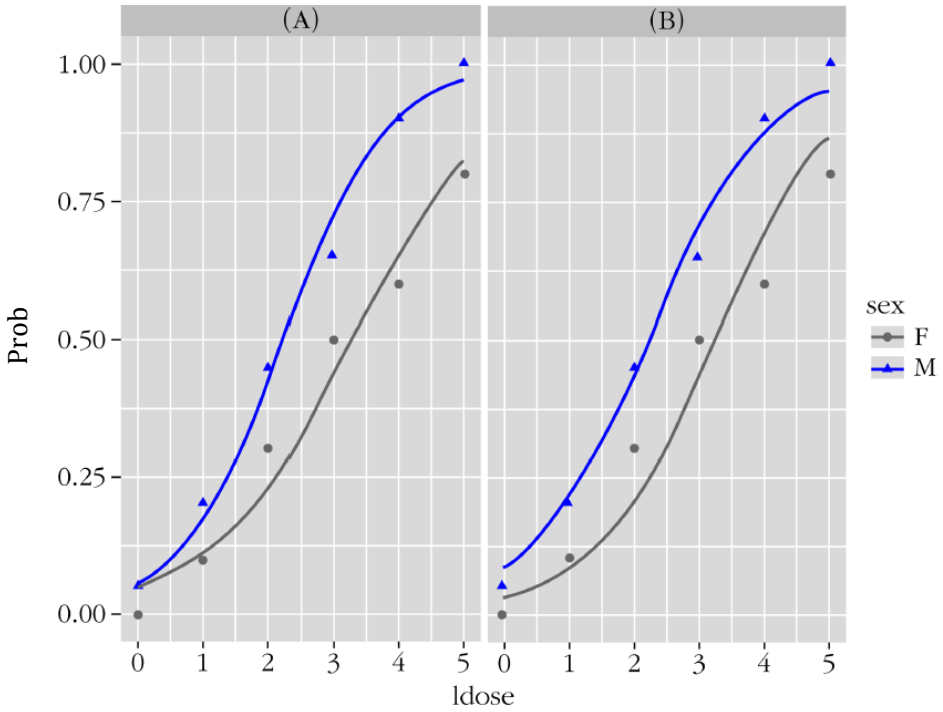
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.876 on 11 degrees of freedom

Residual deviance: 5.566 on 9 degrees of freedom

AIC: 41.676

두 모형의 비교



[그림 6.2] 담배나방의 살충효과에 관한 두 모형의 비교

	모형A	모형B
이탈도	3.768	5.566
AIC	41.878	41.676

3 포아송분포 경우의 예

갈라파고스 자료

Johnson & Raven (1973)에 소개된, 갈라파고스 제도에 속하는 30개의 섬들 대한 생태적 특성을 기록한 자료.

섬 이름, 다음과 같은 7개의 변수를 포함.

Species : 섬에서 발견된 식물 종의 수

Endemics : 토착 종의 수

Area : 섬의 면적(km^2)

Elevation : 섬의 최고점 해발고도(m)

Nearest : 가장 가까운 섬과의 거리(km)

Scruz : 산타크루즈(Santa Cruz) 섬과의 거리(km)

Adjacent : 인접한 섬의 면적(km^2)

갈라파고스 자료 읽기

```
> install.packages('faraway')
> library('faraway')
> gala
```

```
> rbind(gala[1:3,], gala[27:30,])
           Species Endemics   Area Elevation Nearest Scruz Adjacent
Baltra          58       23  25.09       346     0.6   0.6     1.84
Bartolome       31       21   1.24       109     0.6  26.3   572.33
Caldwell         3        3   0.21       114     2.8  58.7     0.78
SantaMaria     285       73 170.92       640     2.6  49.2     0.10
Seymour         44       16   1.84       147     0.6   9.6    25.09
Tortuga         16        8   1.24       186     6.8  50.9    17.95
Wolf            21       12   2.85       253    34.1 254.7     2.33
```

갈라파고스 자료

```
> galax<- gala[,-2]  # Endemics 제거
> galax0<- glm(Species~., family=poisson, data=galax)
> summary( galax0 )
....
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
....
    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68
....
```

Endemics 변수를 제외하고 data.frame 구성

Species 종속변수, 나머지 모든 변수를 독립변수로.

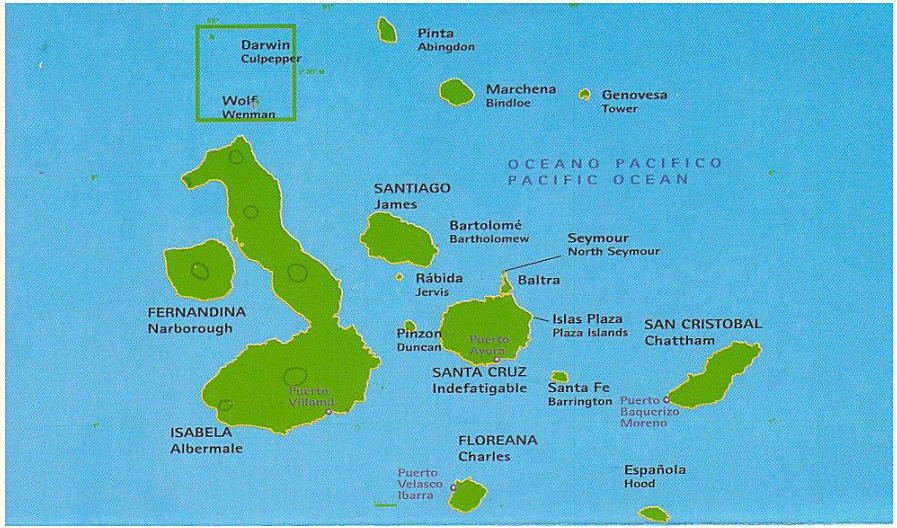
[R 6.16] 갈라파고스 자료 전체 변수를 이용한 포아송 회귀분석

갈라파고스 자료

```
> galax<- gala[,-2]
> gala.glm1<- stepAIC( glm(Species~log(Area)+log(Elevation)
+                  +Nearest+Scruz+log(Adjacent), family=poisson, data=galax ) )
Start:  AIC=519.45
....
Step:  AIC=516.91
Species ~ log(Area) + Scruz + log(Adjacent)

                Df Deviance    AIC
<none>              348.1   516.9
- Scruz          1   395.5   562.4
- log(Adjacent)  1   527.7   694.5
- log(Area)      1  3343.6  3510.4
```

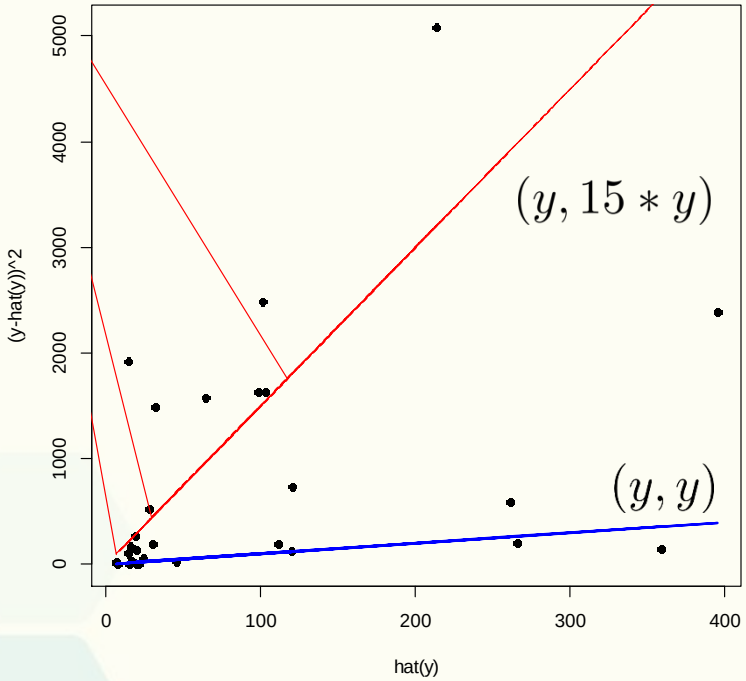
```
> summary( gala.glm1 )

Coefficients:
                Estimate  Std. Error z value  Pr(>|z|)
(Intercept)     3.4893    0.0515   67.745   < 2e-16 ***
log(Area)        0.3577    0.0080   44.429   < 2e-16 ***
Scruz           -0.0032    0.0004   -6.582  4.63e-11 ***
log(Adjacent)  -0.0835    0.0063  -13.166   < 2e-16 ***
---  ...
Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  348.08  on 26  degrees of freedom
AIC: 516.91
```

갈라파고스 제도의 배치


Darwin
Culpepper
Wolf
Wenman
Pinta
Abingdon
Marchena
Bindloe
Genovesa
Tower
OCEANO PACIFICO
PACIFIC OCEAN
SANTIAGO
James
Bartolomé
Bartholomew
Seymour
North Seymour
Rábida
Jervis
Baltra
FERNANDINA
Narborough
Pinzon
Duncan
Puerto
Ayora
Islas Plaza
Plaza Islands
SAN CRISTOBAL
Chattham
SANTA CRUZ
Indefatigable
Santa Fe
Barrington
Puerto
Baquerizo
Moreno
Puerto
Villamil
ISABELA
Albermale
FLOREANA
Charles
Puerto
Velasco
Ibarra
Española
Hood

과산포성의 확인

$$(\hat{y}_i, (y_i - \hat{y}_i)^2)$$



산포모수의 추정

```
> rd<- resid( gala.glm1 )
        #  기본설정은 "deviance" residual
> dpd<- sum(rd*rd)/gala.glm1$df.res
        #  13.38   =   이탈도 348.1  / 잔차 자유도 (30-4)

> rsp<-resid( gala.glm1 , "pearson" )
        # Pearson 잔차 (y-yhat)/sqrt(yhat)
> dpp<- sum(rsp*rsp)/gala.glm1$df.res
        #  15.07  = 피어슨 잔차 제곱합 391.9 / 26
```

의사포아송 분포

```
> gala.glm3<-glm(Species~log(Area)+Scruz+log(Adjacent),
+                family=quasipoisson(link=log), data=galax)
> summary(gala.glm3)
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   3.489344   0.199974    17.449   7.12e-16 ***
log(Area)     0.357777   0.031265    11.443   1.19e-11 ***
Scruz        -0.003186   0.001879    -1.695   0.10195
log(Adjacent) -0.083586  0.024648    -3.391   0.00223 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 15.07352)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  348.08  on 26  degrees of freedom
AIC: NA
```

4 다항분포 경우의 예

다항분포 로지스틱 회귀모형

$$(y_{(0)}, y_{(1)}, \ldots, y_{(K)}) \sim Multinom(n, p_0, \ldots, p_K)$$

$$\sum_{i=0}^{K} y_{(k)} = n \qquad \sum_{i=0}^{K} p_k = 1$$

$$\log(p_k/p_0) = \alpha_k + \beta_k x, \quad k = 0, 1, \ldots, K$$

$$\alpha_0 = \beta_0 = 0$$

다항분포 로짓 모형.

$$p_k = \frac{\exp(\alpha_k + \beta_k x)}{\sum_{k=0}^{K} \exp(\alpha_k + \beta_k x)}, \; k = 0, \ldots, K$$

교육프로그램 선택 자료

http://faculty.knou.ac.kr/~sskim/ progselection.txt

교육프로그램 선택 자료

- 외국 고등학교의 교육 프로그램 선택에 대한 자료.
- 교육 선택 대안 : 직업교육, 진학교육, 일반교육
- 200명 학생에 대하여 생성한 가상적 자료.

읽기(read), 쓰기(write), 수학 등의 성적과, 다음 변수들.

sex : 학생의 남녀 성별(여: f, 남: m)

ses : 학생의 사회경제적 수준. 상(high), 중(middle), 하(low).

type : 학생이 속한 학교가 공립(public)인지 사립(private)인지를 나타냄

prog : 학생이 선택한 고교 교육 프로그램
(직업교육: vocational, 진학교육: academic, 일반교육: general)

honors : 우등반에 등록 했는지의 여부

awards : 수상횟수

기초분석

```
> hdata<-read.table("c:/temp/progselection.txt")
> hdata$prog<-factor(hdata$prog,c("vocation","general","academic"))
> hdata$ses<-factor(hdata$ses,c("low","middle","high"))

> with(hdatax, table(ses,prog))
 prog
ses       vocation general academic
  low           15      14       20
  middle        30      24       40
  high           8      10       39
```

	prog		
ses			
	0.31	0.29	0.41
	0.32	0.26	0.43
	0.14	0.18	0.68

```
> with(hdata, do.call(rbind, tapply( read, prog,
+                     function(x) c(M = mean(x), SD = sd(x)))))
                M        SD
vocation 46.54717  8.830816
general  51.04167  9.118507
academic 56.37374  9.358600
```

세 모형

- 모형 A : prog ~ ses –1
- 모형 B : prog ~ ses + read –1
- 모형 C : prog ~ read

모형 A

```
> library(nnet)
> summary(multinom( prog~ ses-1 , data=hdata))
....
Coefficients:
            seslow     sesmiddle    seshigh
general   -0.06899311  -0.2231440   0.223136
academic   0.28768238   0.2876817   1.584115

Std. Errors:
            seslow     sesmiddle    seshigh
general    0.3716117   0.2738613   0.4743413
academic   0.3415650   0.2415229   0.3881243

Residual Deviance: 404.6719
AIC: 416.6719
```

모형 B

> summary(multinom(**prog~ ses+read-1**, data=hdata))

...

Coefficients:

	seslow	sesmiddle	seshigh	read
general	-2.867706	-3.134470	-2.825895	0.05960965
academic	-5.202425	-5.451592	-4.483538	0.11275868

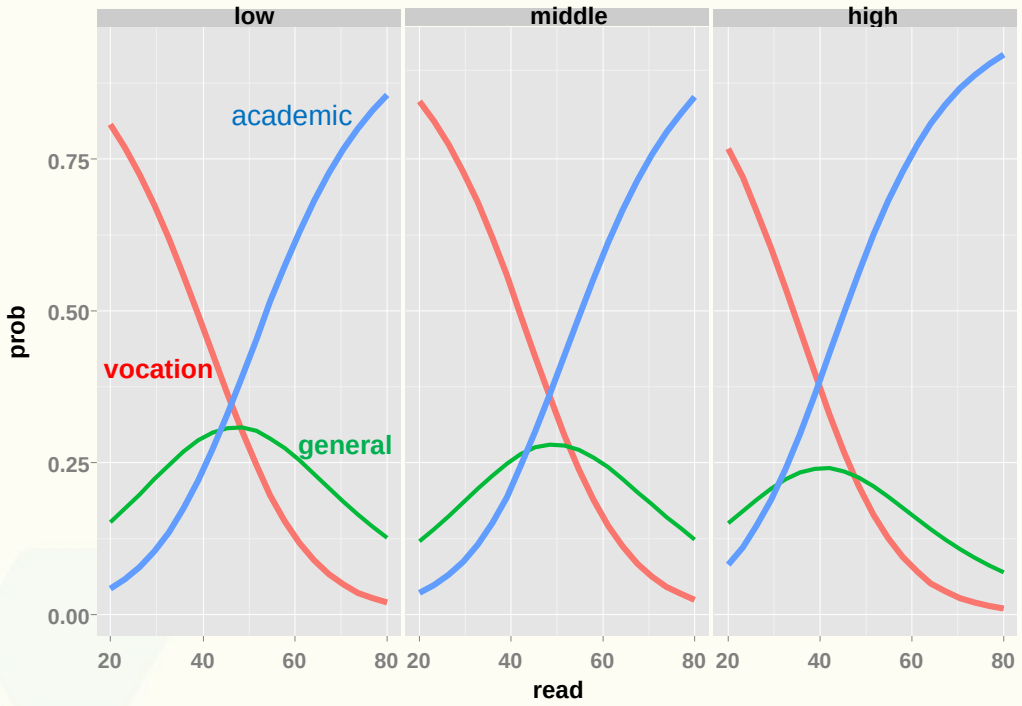
Std. Errors:

	seslow	sesmiddle	seshigh	read
general	1.195471	1.216225	1.323616	0.02419606
academic	1.155365	1.175912	1.242470	0.02290054

Residual Deviance: 373.9069

AIC: 389.9069

모형 B


low
middle
high
academic
vocation
general
prob
0.75
0.50
0.25
0.00
20
40
60
80
read

모형 C

> summary(multinom(**prog~ read** , data=hdata))

....

Coefficients:

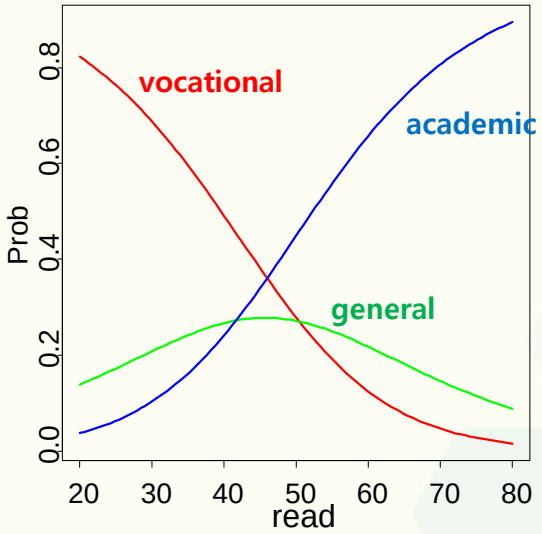
	(Intercept)	read
general	-2.951153	0.05854582
academic	-5.460551	0.11877769

Std. Errors:

	(Intercept)	read
general	1.166806	0.02359973
academic	1.129366	0.02236958

Residual Deviance: 379.0829

AIC: 387.0829



세 모형의 비교

	A	B	C
독립변수	ses	ses, read	read
이탈도	404.7	**373.9**	379.1
AIC	416.7	389.9	**387.1**

다음시간 안내

분류분석 (1)