

12강. 로지스틱 회귀분석

- 로지스틱 회귀모형의 이해
- 로지스틱 함수
- R 로지스틱 회귀분석
- 파이썬 로지스틱 회귀분석

1. 로지스틱 회귀모형의 이해

◆ 로지스틱 회귀분석

독립변수들이 다변량 정규분포의 가정을 따르지 않는 경우,
예를 들어 독립변수들이 이산형과 연속형 변수들로 이루어져 있는 경우,
효과적으로 분류에 이용되는 모형

◆ 종속변수 Y가 두 가지 값(0 또는 1)을 갖고, 독립변수 X가 하나인 경우, 로지스틱 함수(logistic function)

$$P(Y=1 | X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

: X가 주어질 때 그룹 1에 속할 확률을 의미

1. 로지스틱 회귀모형의 이해

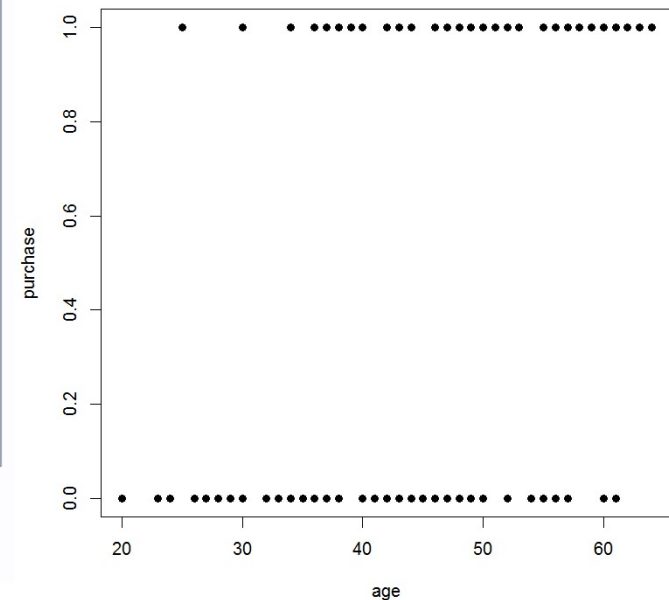
- ◆ 예) 어느 제약제품의 약 구매 여부를 조사하였다.
반응변수 purchase 0=구매 안함, 1=구매함이다. 케이스 수는 100개이다.
이 자료에서 변수 나이(age)와 구매여부(purchase)의 관계를 알아보도록 하자.

	A	B	C
1	id	age	purchase
2	1	20	0
3	2	23	0
4	3	24	0
5	4	25	1
6	5	26	0
7	6	27	0
8	7	27	0
9	8	28	0
10	9	29	0
11	10	29	0
12	11	30	0
13	12	30	0
14	13	30	0
15	14	30	1
16	15	32	0

1. 로지스틱 회귀모형의 이해

(1) 산점도 그리기

```
> drug = read.csv("c:/data/mva/drug.csv")
> head(drug.data)
  id age purchase
1  1  20        0
2  2  23        0
3  3  24        0
4  4  25        1
5  5  26        0
6  6  27        0
> plot(drug$age, drug$purchase, pch=19)
```



y-축을 살펴보면 $y=0$ 과 $y=1$ 에서 값이 위치해 있고, age가 높을수록 $y=1$ 의 값을 취하는 경향이 있음을 짐작할 수 있으나, 두 변수 age와 purchase의 관계를 뚜렷하게 밝히기가 어렵다.

1. 로지스틱 회귀모형의 이해

(2) 나이 그룹화

```
> #Recoding
> agr = age
> agr[agr <= 29 ] = 1
> agr[agr >= 30 & agr <= 34 ] = 2
> agr[agr >= 35 & agr <= 39 ] = 3
> agr[agr >= 40 & agr <= 44 ] = 4
> agr[agr >= 45 & agr <= 49 ] = 5
> agr[agr >= 50 & agr <= 54 ] = 6
> agr[agr >= 55 & agr <= 59 ] = 7
> agr[agr >= 60 & agr <= 64 ] = 8
```

=>

```
> library(car)
> drug$agr = recode(drug$age, "lo:29=1;
  30:34=2; 35:39=3; 40:44=4; 45:49=5;
  50:54=6; 55:59=7; 60:hi=8")
```

```
> drug[c(1,20,40, 60, 80, 100),]
      id age purchase agr
1      1  20         0   1
20     20  34         0   2
40     40  41         0   4
60     60  48         0   5
80     80  56         0   7
100    100  64         1   8
>
```

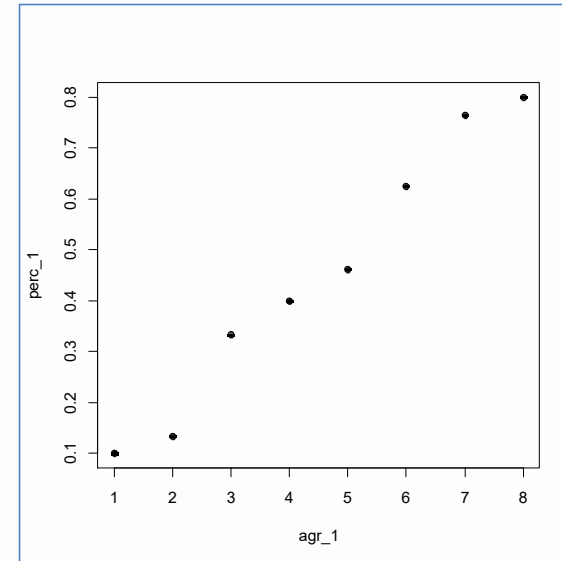
1. 로지스틱 회귀모형의 이해

(3) 그룹화된 변수 agr 과 purchase 관계

```
> purchase_table = table(drug$agr, drug$purchase)
> percent_table = prop.table(purchase_table, 1)
> percent_table
```

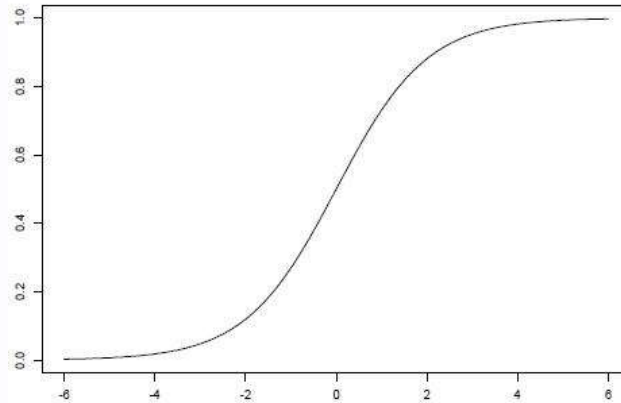
	0	1
1	0.9000000	0.1000000
2	0.8666667	0.1333333
3	0.6666667	0.3333333
4	0.6000000	0.4000000
5	0.5384615	0.4615385
6	0.3750000	0.6250000
7	0.2352941	0.7647059
8	0.2000000	0.8000000

```
> perc_1 = percent_table[,2]
> agr_1 = rownames(percent_table)
> plot(agr_1, perc_1, pch=19)
```



- 나이가 많아질수록 구입(purchase)할 확률은 증가하며, 이 형태는 S-shape 형태를 지니는 것을 알 수 있음.
- 로지스틱 함수는 이와 같이 S-shape 형태를 지닌 함수를 적합할 때 이용됨.

2. 로지스틱 함수



< 로지스틱 함수 >

- ◆ 종속변수 Y : 0 또는 1 의 두가지 값을 취하는 변수
독립변수 X 가 하나인 경우
 $P(Y=1|X)$: 주어진 X 에서 $Y=1$ 일 확률

로지스틱 함수 : 그림과 같이 S-shape 곡선으로 X 가 증가함에 따라 1에 수렴하고,
 X 가 감소할 때 0으로 수렴하는 함수.

$$P(Y = 1 | X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

2. 로지스틱 함수

◆ <로지스틱 함수 변환>

로지스틱 함수 $P(Y = 1 | X) = p$ 라 하면,

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

여기서, $\frac{p}{1-p}$ 를 오즈(odds) 라 함.

$$\text{오즈} = \frac{p}{1-p} \quad (\Leftrightarrow p = \frac{\text{odds}}{1 + \text{odds}})$$

◆ 오즈의 의미

예) 스포츠게임에서 A팀이 B팀을 이길 오즈가 4 : A팀이 B팀을 이길 확률이 4배라는 의미

$$\text{이길 확률} \quad p = \frac{4}{1+4} = 0.8, \quad \text{즉,} \quad \frac{0.8}{1-0.8} = 4$$

R 분석

- 예) 풀 깎는 기계를 생산하는 어느 회사에서 자사제품의 판매를 강화하기 위한 마케팅전략을 세우기 위하여 기계를 소유한 사람과 소유하지 않은 사람들의 두 그룹에 대하여 재산정도, 땅의 크기를 조사한 결과이다. 로지스틱 회귀분석을 이용하여 기계를 소유할 가능성이 있는 사람을 판별하기 위한 분석을 해보도록 하자.

소유자(1)		비소유자(0)	
재산	크기	재산	크기
(income)	(size)	(income)	(size)
20.0	9.2	25.0	9.8
28.5	8.4	17.6	10.4
21.6	10.8	21.6	8.6
20.5	10.4	14.4	10.2
29.0	11.8	28.0	8.8
36.7	9.6	16.4	8.8
36.0	8.8	19.8	8.0
27.6	11.2	22.0	9.2
23.0	10.0	15.8	8.2
31.0	10.4	11.0	9.4
17.0	11.0	17.0	7.0
27.0	10.0	21.0	7.4

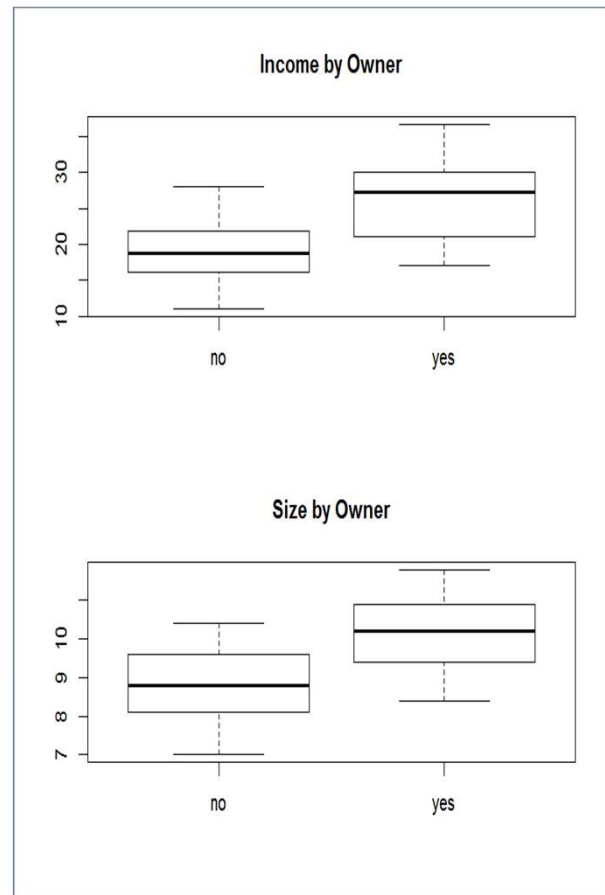


	A	B	C
1	owner	income	size
2	yes	20	9.2
3	yes	28.5	8.4
4	yes	21.6	10.8
5	yes	20.5	10.4
6	yes	29	11.8
7	yes	36.7	9.6
8	yes	36	8.8
9	yes	27.6	11.2
10	yes	23	10
11	yes	31	10.4
12	yes	17	11
13	yes	27	10
14	no	25	9.8
15	no	17.6	10.4
16	no	21.6	8.6
17	no	14.4	10.2
18	no	28	8.8
19	no	16.4	8.8
20	no	19.8	8
21	no	22	9.2
22	no	15.8	8.2
23	no	11	9.4
24	no	17	7
25	no	21	7.4

R 분석

(1) 데이터 읽기

```
> mower = read.csv("c:/data/mva/mower.csv")
> head(mower)
  owner income size
1  yes   20.0  9.2
2  yes   28.5  8.4
3  yes   21.6 10.8
4  yes   20.5 10.4
5  yes   29.0 11.8
6  yes   36.7  9.6
> par(mfrow=c(2:1))
> boxplot(income ~ owner, data=mower)
> title("Income by Owner")
> boxplot(size ~ owner, data=mower)
> title("Size by Owner")
```



R 분석

(2) 반응변수 변환

```
> mower_logit = glm(owner~ . , family=binomial, data=mower)
eval(family$initialize)에서 다음과 같은 에러가 발생했습니다:
y values must be 0 <= y <= 1
```

```
> mower$owner[mower$owner=='yes'] = 1
> mower$owner[mower$owner=='no'] = 0
```

```
> # library(car)
> # mower$owner = recode(mower$owner, "'yes'=1; 'no'=0")

> mower$owner = as.factor(mower$owner)
```

R 분석 예

(3) 로지스틱 회귀모형

```
> mower_logit = glm(owner ~ ., family=binomial, data=mower)
> summary(mower_logit)
```

Call:

```
glm(formula = owner ~ ., family = binomial, data = mower.data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.74044	-0.29685	0.00439	0.44750	1.86821

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-25.9382	11.4871	-2.258	0.0239 *
income	0.3326	0.1629	2.042	0.0412 *
size	1.9276	0.9256	2.083	0.0373 *

$$\Rightarrow \ln\left(\frac{p}{1-p}\right) = -25.938 + 0.333income + 1.928size$$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 33.271 on 23 degrees of freedom
Residual deviance: 15.323 on 21 degrees of freedom
AIC: 21.323
Number of Fisher Scoring iterations: 6

(income, size) 두 변수 모두

유의수준 0.05에서 유의

, $p = \text{Pr}(\text{owner} = \text{'yes'})$

R 분석 예

(4) 오즈비(Odds Ratio)

$$\ln\left(\frac{p}{1-p}\right) = -25.938 + 0.333income + 1.928size$$

$$\Rightarrow \frac{p}{1-p} = e^{-25.938 + 0.333 \times income + 1.928 \times size}$$

```
> exp(mower_logit$coef)
(Intercept)      income      size
5.434620e-12 1.394556e+00 6.872708e+00
```

$\exp(0.333) = 1.39$: income 이 1 단위 증가할 때, 오즈비의 증가율은 1.39배

즉, income 이 1 단위 증가할 때 39% 증가

$\exp(1.928) = 6.87$: size 가 1 단위 증가하면 오즈비의 증가는 6.87배

size 가 income 에 비해서 상대적으로 영향력이 큼.

참고: 계수가 음수이면 오즈비는 감소.

R 분석 예

(5) 로지스틱 회귀모형 평가

```
> mower_logit = glm(owner~ . ,family=binomial, data=mower)
> summary(mower_logit)
```

...

```
Null deviance: 33.271 on 23 degrees of freedom
Residual deviance: 15.323 on 21 degrees of freedom
AIC: 21.323
Number of Fisher Scoring iterations: 6
```

$$p - value = P(\chi^2 > 15.323) \\ = 1 - P(\chi^2 \leq 15.323)$$

```
> 1-pchisq(15.323, 21)
[1] 0.8064027
```

Null deviance : 상수항만이 포함된 모형으로 적합할 때 모형추정값과 관찰값의 차이에 관한 통계량

Residual deviance : 독립변수가 포함될 때의 차이에 대한 통계량

귀무가설 H_0 : 모형이 적합하다(Model is correct).

대립가설 H_1 : 모형이 적합하지 않다(Model is not correct).

유의확률 p-값 = 0.806 이므로 귀무가설, 즉 모형이 적합하다는 가설을 받아들임.

R 분석 예

(6) 새로운 자료의 분류

```
> mower_predict=predict(mower_logit, newdata=mower,
                        type="response")
> round(mower_predict,3)
  1    2    3    4    5    6    7    8    9   10   11   12   13
0.175 0.433 0.887 0.716 0.998 0.992 0.952 0.992 0.728 0.988 0.715 0.910 0.780
 14   15   16   17   18   19   20   21   22   23   24
0.490 0.102 0.184 0.583 0.029 0.019 0.292 0.008 0.015 0.001 0.009
> pred = ifelse(mower_predict < 0.5, "no", "yes")
> pred = factor(pred)
> pred
  1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19
no   no yes yes yes yes yes yes yes yes yes yes yes  no   no   no yes  no   no
 20   21   22   23   24
no   no   no   no   no
Levels: no yes
```

새로운 자료를 분류하기 위해서는 predict() 함수를 이용.

Mower_predict : 각 케이스별로 owner="yes" 에 속할 확률.

"pred = ifelse(mower_predict < 0.5, "no", "yes")" : 확률이 0.5 보다 작으면 "no" ,
0.5 이상이면 "yes" 로 분류

R 분석 예

(7) 분류표 결과

```
> cm = table(mower$owner, pred)
> cm
      pred
      no  yes
no    10    2
yes    2   10
> prop.table(cm,1)
      pred
      no      yes
0 0.8333333 0.1666667
1 0.1666667 0.8333333
> error = 1 - (sum(diag(cm)/sum(cm)))
> error
[1] 0.1666667
```

- 분류표 결과에서 보면, 그룹 yes, no 모두 12 케이스에서 각각 2개씩 잘못 분류됨.
- 전체 오류율 = 0.1667

R 분석 예 2

- 예) R 패키지 MASS 에는 “menarche” 자료 (Miller, H. and Szczotka, F., 1966, Age at Menarche in Warsaw girls in 1965, Human Biology, 38, 199-203)가 있다.
변수는 Age(그룹의 평균 연령), Total(그룹의 총 수), Menarche(menarche에 접한 수)이다.
이 자료를 이용하여 로지스틱 회귀분석을 해보자.

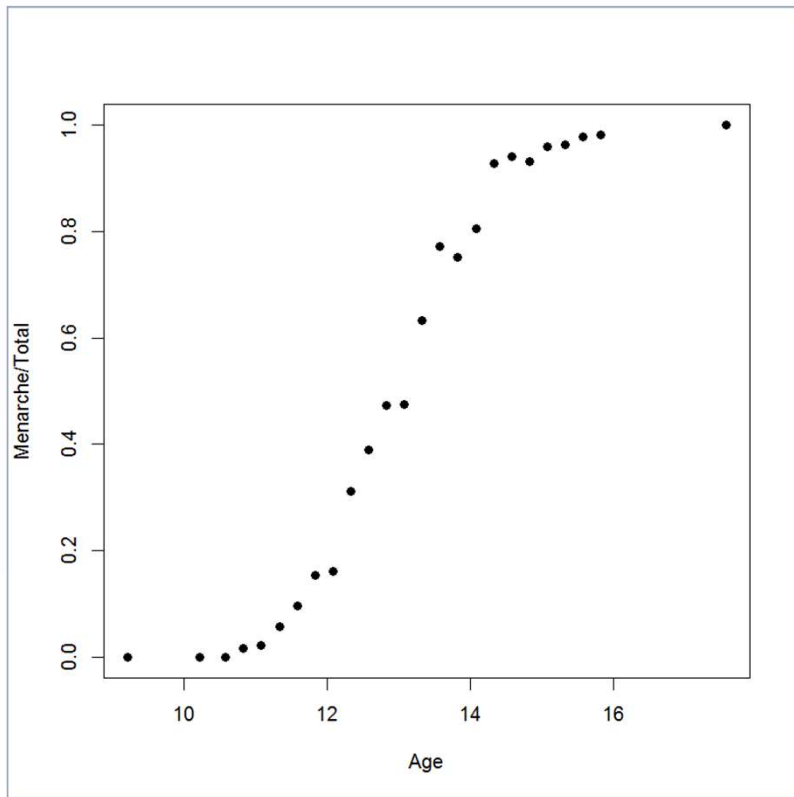
```
> library(MASS)
> data(menarche)
> head(menarche)
```

	Age	Total	Menarche
1	9.21	376	0
2	10.21	200	0
3	10.58	93	0
4	10.83	120	2
5	11.08	90	2
6	11.33	88	5

R 분석 예 2

(1) 그래프 보기

```
> plot(Menarche/Total ~ Age, data=menarche, pch=19)
```



그림을 보면 S-shape의 형태를 지니고 있으므로 로지스틱 회귀모형을 적합하는 것이 타당해 보임.

R 분석 예 2

(2) 로지스틱 회귀모형 적합

```
> menr_out = glm(cbind(Menarche, Total-Menarche)~Age, family=binomial,  
+ data=menarche)  
> summary(menr_out)
```

...

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-21.22639	0.77068	-27.54	<2e-16 ***
Age	1.63197	0.05895	27.68	<2e-16 ***

$$\Rightarrow \ln\left(\frac{p}{1-p}\right) = -21.226 + 1.632 \times \text{Age}$$

Null deviance: 3693.884 on 24 degrees of freedom
Residual deviance: 26.703 on 23 degrees of freedom
AIC: 114.76

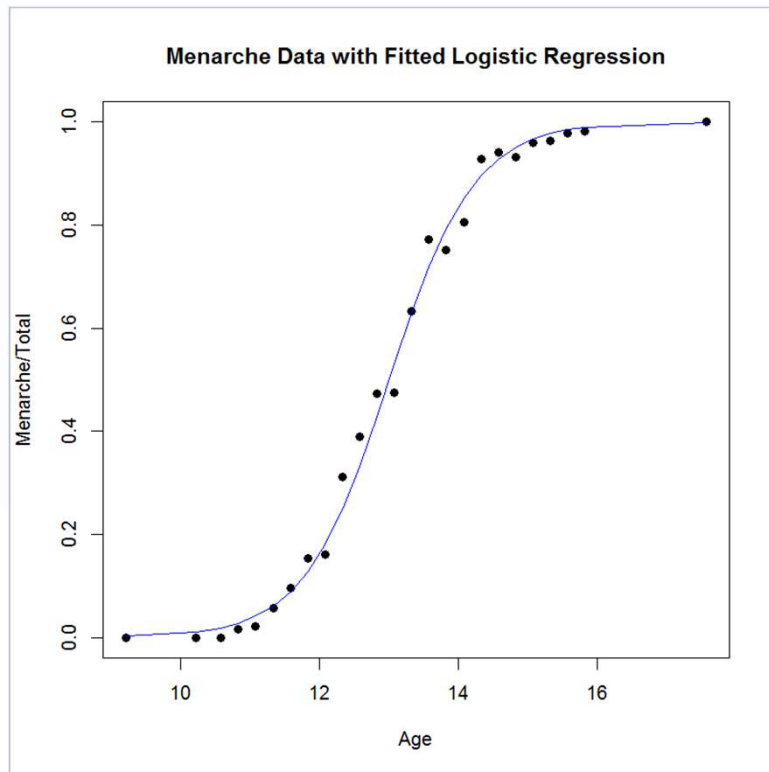
- Age 계수에 대한 p-값=0.000 이므로 변수 Age가 유의
- Age의 오즈비 : $\exp(1.63197) = 5.11$, Age가 1년 증가할수록 오즈비는 5.11배
- Residual deviance"에 대한 p-값 = 0.269 이므로
모형이 적합하다는 귀무가설을 받아들임.

```
> 1-pchisq(26.703, 23)  
[1] 0.2688152
```

R 분석 예 2

(3) 로지스틱 회귀선 적합

```
> plot(Menarche/Total ~ Age, data=menarche, pch=19)  
> lines(menarche$Age, menr.out$fitted, type="l", col="blue")  
> title("Menarche Data with Fitted Logistic Regression")
```



파이썬 분석

- 예) 풀 깎는 기계를 생산하는 어느 회사에서 자사제품의 판매를 강화하기 위한 마케팅 전략을 세우기 위하여 기계를 소유한 사람과 소유하지 않은 사람들의 두 그룹에 대하여 재산 정도, 땅의 크기를 조사한 결과이다. 로지스틱 회귀분석을 이용하여 기계를 소유할 가능성이 있는 사람을 판별하기 위한 분석을 해보도록 하자.

소유자(1)		비소유자(0)	
재산	크기	재산	크기
(income)	(size)	(income)	(size)
20.0	9.2	25.0	9.8
28.5	8.4	17.6	10.4
21.6	10.8	21.6	8.6
20.5	10.4	14.4	10.2
29.0	11.8	28.0	8.8
36.7	9.6	16.4	8.8
36.0	8.8	19.8	8.0
27.6	11.2	22.0	9.2
23.0	10.0	15.8	8.2
31.0	10.4	11.0	9.4
17.0	11.0	17.0	7.0
27.0	10.0	21.0	7.4



	A	B	C
1	owner	income	size
2	yes	20	9.2
3	yes	28.5	8.4
4	yes	21.6	10.8
5	yes	20.5	10.4
6	yes	29	11.8
7	yes	36.7	9.6
8	yes	36	8.8
9	yes	27.6	11.2
10	yes	23	10
11	yes	31	10.4
12	yes	17	11
13	yes	27	10
14	no	25	9.8
15	no	17.6	10.4
16	no	21.6	8.6
17	no	14.4	10.2
18	no	28	8.8
19	no	16.4	8.8
20	no	19.8	8
21	no	22	9.2
22	no	15.8	8.2
23	no	11	9.4
24	no	17	7
25	no	21	7.4

파이썬 분석 : sklearn

(1) 데이터 읽기

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# 데이터 읽기
mower = pd.read_csv("c:/data/mva/mower.csv") mower.head()
Out[1]:
```

	owner	income	size
0	yes	20.0	9.2
1	yes	28.5	8.4
2	yes	21.6	10.8
3	yes	20.5	10.4
4	yes	29.0	11.8

```
# 변수선택
y = mower["owner"]
X = mower[["income", "size"]]
```

파이썬 분석 : sklearn

(2) 로지스틱 회귀모형

```
# 로지스틱 회귀분석 실행
from sklearn.linear_model import LogisticRegression
mower_clf = LogisticRegression()
mower_clf.fit(X,y)
Out[4]: LogisticRegression()
```

```
# 로지스틱 회귀모형 절편
mower_clf.intercept_
Out[5]: array([-17.42159563])
```

```
# 로지스틱 회귀모형 계수
mower_clf.coef_
Out[6]: array([[0.26596415, 1.19925326]])
```

```
# 분류 클래스
mower_clf.classes_
Out[7]: array(['no', 'yes'], dtype=object)
```

```
# 로지스틱 회귀분석 실행2
from sklearn.linear_model import LogisticRegression
mower_clf2 = LogisticRegression(penalty= 'none' )
mower_clf2.fit(X,y)
Out[4]: LogisticRegression(penalty= 'none' )
```

```
# 로지스틱 회귀모형 절편2
mower_clf2.intercept_
Out[5]: array([-25.93824588])
```

```
# 로지스틱 회귀모형 계수2
mower_clf2.coef_
Out[6]: array([[0.33257614, 1.92755858]])
```

```
# 분류 클래스2
mower_clf2.classes_
Out[7]: array(['no', 'yes'], dtype=object)
```

파이썬 분석 : sklearn

(3) 새로운 자료의 분류

```
mower_clf.predict_proba(X) [0:6]
```

```
Out[20]:
```

```
array([[0.74444143, 0.25555857],  
       [0.44223159, 0.55776841],  
       [0.21837432, 0.78162568],  
       [0.3768586 , 0.6231414 ],  
       [0.01162903, 0.98837097],  
       [0.02079313, 0.97920687]])
```

```
# 분류 결과
```

```
mower_clf.predict(X)[0:6]
```

```
Out[21]: array(['no', 'yes', 'yes', 'yes', 'yes', 'yes'],  
               dtype=object)
```

```
mower_clf2.predict_proba(X) [0:6]
```

```
Out[20]:
```

```
array([[0.82537314, 0.17462686],  
       [0.5666838 , 0.4333162 ],  
       [0.11274223, 0.88725777],  
       [0.28370152, 0.71629848],  
       [0.00157543, 0.99842457],  
       [0.00839357, 0.99160643]])
```

```
# 분류 결과2
```

```
mower_clf2.predict(X)[0:6]
```

```
Out[21]: array(['no', 'no', 'yes', 'yes', 'yes', 'yes'],  
               dtype=object)
```


파이썬 분석 : sklearn

(4) 분류표

분류표 구하기

```
from sklearn.metrics import classification_report,  
confusion_matrix  
from sklearn.metrics import accuracy_score
```

```
cm = confusion_matrix(y, mower_clf.predict(X))
```

cm

Out[23]:

```
array([[10, 2],  
       [ 1, 11]], dtype=int64)
```

accuracy 계산하기

```
from sklearn.metrics import accuracy_score  
pred_class = mower_clf.predict(X)  
print('Accuracy = '+str(accuracy_score(y, pred_class)))  
Accuracy = 0.875
```

분류표 구하기2

```
from sklearn.metrics import classification_report, confusion_matrix  
from sklearn.metrics import accuracy_score
```

```
cm2 = confusion_matrix(y, mower_clf2.predict(X))
```

cm2

Out[23]:

```
array([[10, 2],  
       [ 2, 10]], dtype=int64)
```

accuracy 계산하기2

```
from sklearn.metrics import accuracy_score  
pred_class2 = mower_clf2.predict(X)  
print('Accuracy = '+str(accuracy_score(y, pred_class2)))  
Accuracy = 0.833
```

파이썬 분석 : sklearn

(4) 분류표

세분화된 분류표

```
cm_report = classification_report(y, mower_clf.predict(X))  
print(cm_report)
```

	precision	recall	f1-score	support
no	0.91	0.83	0.87	12
yes	0.85	0.92	0.88	12
accuracy			0.88	24
macro avg	0.88	0.88	0.87	24
weighted avg	0.88	0.88	0.87	24

세분화된 분류표2

```
cm_report2 = classification_report(y, mower_clf2.predict(X))  
print(cm_report2)
```

	precision	recall	f1-score	support
no	0.83	0.83	0.83	12
yes	0.83	0.83	0.83	12
accuracy			0.83	24
macro avg	0.83	0.83	0.83	24
weighted avg	0.83	0.83	0.83	24

파이썬 분석 : statsmodels

(1) 데이터 읽기

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
# 데이터 읽기
mower = pd.read_csv("c:/data/mva/mower.csv")
# 변수 선택
y = mower["owner"]
aX = mower[["income", "size"]]

import statsmodels.api as sm
# 상수 더하기
aX = sm.add_constant(aX)
# array 변환
ay = y.to_numpy()
iy = [0]*len(ay)
```

```
for i in range(0, len(ay)) :
    if(ay[i] == 'yes') :
        iy[i] = 1
    else :
        iy[i] = 0
```

파이썬 분석 : statsmodels

(2) 로지스틱 회귀모형 적합

```
# 로지스틱 회귀모형 적합하기
mower_sm = sm.Logit(iy, aX)
mower_logit = mower_sm.fit()
mower_logit.params
```

```
Out[32]:
const    -25.938232
income    0.332576
size      1.927558
dtype: float64
```

```
# 자료의 분류
mower_logit.predict(aX)
```

```
Out[33]:
0    0.174627
1    0.433316
2    0.887258
3    0.716299
4    0.998425
5    0.991606
...
```

```
mower_pred = (mower_logit.predict(aX) >= 0.5).astype(int)
```

```
mower_pred
```

```
Out[35]:
```

```
0    0
1    0
2    1
3    1
4    1
5    1
```

파이썬 분석 : statsmodels

(3) 로지스틱 회귀 적합 결과

로지스틱 회귀모형 적합 결과

mower_logit.summary()

Out[37]:

<class 'statsmodels.iolib.summary.Summary'>

"""

Logit Regression Results

Dep. Variable:	y	No. Observations:	24
Model:	Logit	Df Residuals:	21
Method:	MLE	Df Model:	2
Date:	Wed, 22 Dec 2021	Pseudo R-squ.:	0.5394
Time:	13:55:00	Log-Likelihood:	-7.6616
converged:	True	LL-Null:	-16.636
Covariance Type:	nonrobust	LLR p-value:	0.0001267

	coef	std err	z	P> z	[0.025	0.975]
const	-25.9382	11.487	-2.258	0.024	-48.453	-3.423
income	0.3326	0.163	2.042	0.041	0.013	0.652
size	1.9276	0.926	2.083	0.037	0.113	3.742

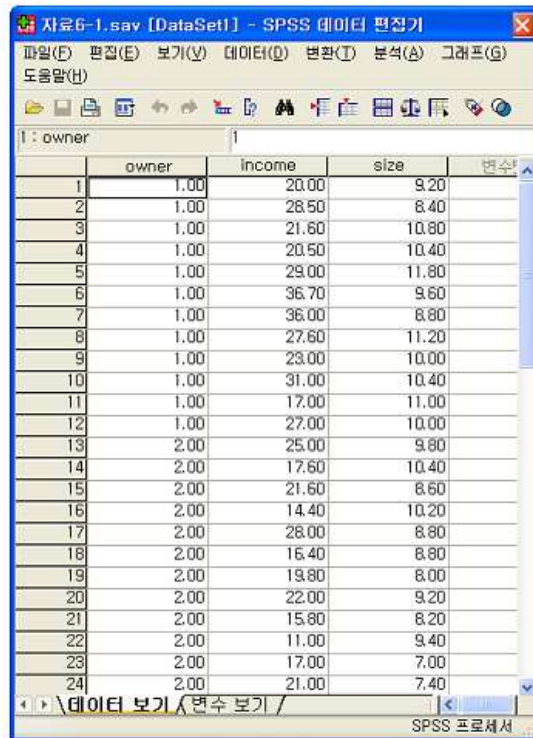
참고 : SPSS 및 SAS 분석 예

로지스틱 회귀분석의 예 - SPSS

1) SPSS 분석의 예

SPSS 로지스틱 회귀분석 절차

① 자료입력



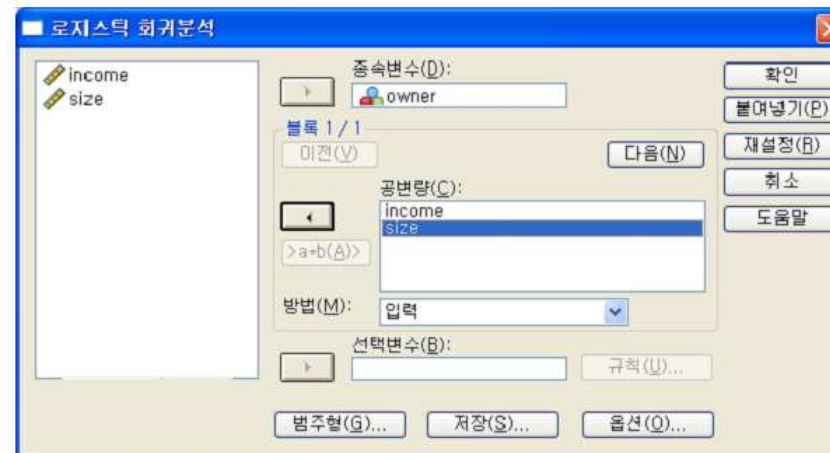
자료6-1.sav [DataSet1] - SPSS 데이터 편집기

	owner	income	size	변수
1	1.00	20.00	9.20	
2	1.00	28.50	8.40	
3	1.00	21.60	10.80	
4	1.00	20.50	10.40	
5	1.00	29.00	11.80	
6	1.00	36.70	9.60	
7	1.00	36.00	8.80	
8	1.00	27.60	11.20	
9	1.00	23.00	10.00	
10	1.00	31.00	10.40	
11	1.00	17.00	11.00	
12	1.00	27.00	10.00	
13	2.00	25.00	9.80	
14	2.00	17.60	10.40	
15	2.00	21.60	8.60	
16	2.00	14.40	10.20	
17	2.00	28.00	8.80	
18	2.00	16.40	8.80	
19	2.00	19.80	8.00	
20	2.00	22.00	9.20	
21	2.00	15.80	8.20	
22	2.00	11.00	9.40	
23	2.00	17.00	7.00	
24	2.00	21.00	7.40	

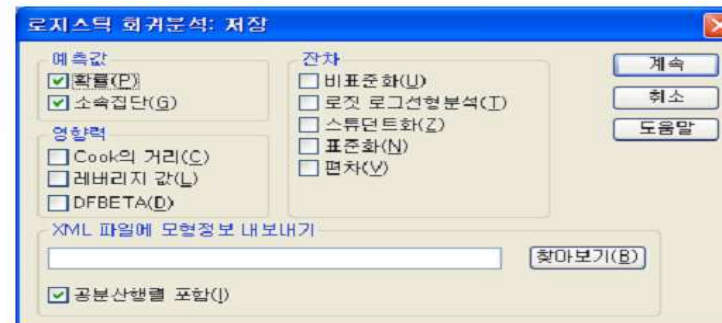
SPSS 프로세서

② 분석 절차

분석-회귀분석-이분형 로지스틱 절차



<로지스틱 회귀분석 대화상자>



<저장 대화상자>

로지스틱 회귀분석의 예 - SPSS

③ 출력결과

<분류표와 로지스틱 회귀계수(SPSS 출력결과)>

분류표^a

관측		예측값		
		OWNER	분류정확 %	
		1	2	
1 단계	OWNER 1	10	2	83.3
	2	2	10	83.3
전체 %				83.3

a. 절단값은 .500입니다.

방정식에 포함된 변수

		B	S.E.	Wald	자유도	유의확률	Exp(B)
1 단계	INCOME	-.333	.163	4.169	1	.041	.717
계 ^a	SIZE	-1.927	.925	4.337	1	.037	.146
	상수	25.934	11.485	5.099	1	.024	1.832E+11

a. 변수가 1: 단계에 진입했습니다 INCOME, SIZE.

: 로지스틱 회귀모형은

$$\ln\left(\frac{p}{1-p}\right) = 25.934 - 0.333 \times \text{income} - 1.927 \times \text{size}$$

단, 여기서 $p = \Pr(\text{owner} = 2)$ 임. 분류표에서 볼 때, 그룹 1, 2 모두 12 케이스에서 각각 2개씩 잘못 분류되었다는 것을 알 수 있음.

로지스틱 회귀분석의 예 - SPSS

<예측 확률과 예측그룹>

	owner	income	size	PRE_1	PGR_1	변수
1	1.00	20.00	9.20	.82537	2.00	
2	1.00	28.50	8.40	.56668	2.00	
3	1.00	21.60	10.80	.11274	1.00	
4	1.00	20.50	10.40	.28370	1.00	
5	1.00	29.00	11.80	.00158	1.00	
6	1.00	36.70	9.60	.00839	1.00	
7	1.00	36.00	8.80	.04756	1.00	
8	1.00	27.60	11.20	.00793	1.00	
9	1.00	23.00	10.00	.27159	1.00	
10	1.00	31.00	10.40	.01191	1.00	
11	1.00	17.00	11.00	.28522	1.00	
12	1.00	27.00	10.00	.08973	1.00	
13	2.00	25.00	9.80	.21990	1.00	
14	2.00	17.60	10.40	.50957	2.00	
15	2.00	21.60	8.60	.89822	2.00	
16	2.00	14.40	10.20	.81578	2.00	
17	2.00	28.00	8.80	.41668	1.00	
18	2.00	16.40	8.80	.97129	2.00	
19	2.00	19.80	8.00	.98079	2.00	
20	2.00	22.00	9.20	.70848	2.00	
21	2.00	15.80	8.20	.99244	2.00	
22	2.00	11.00	9.40	.98465	2.00	
23	2.00	17.00	7.00	.99888	2.00	
24	2.00	21.00	7.40	.99090	2.00	

: 추정된 로지스틱 회귀모형을 이용하여
각 케이스를 예측한 확률(PRE_1)과
예측그룹(PGR_1)을 나타낸 것.

예측그룹은 $p=0.5$ 를 기준으로 하여 배정.

여기서 예측확률은 $p = \Pr(\text{owner} = 2)$ 의 추정값임.

로지스틱 회귀분석의 예 - SAS

2) SAS 분석의 예

❶ 로지스틱 회귀분석을 위한 SAS 절차 : PROC LOGISTIC

참고로 SAS의 경우는 그룹값이 작은 것이 기준임. 즉, $p = \text{Pr}(\text{owner} = 1)$

(예) 로지스틱 회귀분석을 위한 SAS 프로그램

```
DATA mower;
  INPUT owner income size @@;
DATALINES;
  1 20.0 9.2 2 25.0 9.8
  ...
  1 27.0 10.0 2 21.0 7.4
RUN;
PROC LOGISTIC;
  model owner = income size
    /ctable;
  output out=result p=phat; RUN;
PROC SORT DATA=result;
  BY owner; RUN;
PROC PRINT DATA=result;
  var owner income size phat; RUN;
```

<출력결과>

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-25.9382	11.4873	5.0985	0.0239
income	1	0.3326	0.1629	4.1685	0.0412
size	1	1.9276	0.9256	4.3369	0.0373

추정된 로지스틱 회귀모형 계수. 추정된 회귀식은

$$\ln\left(\frac{p}{1-p}\right) = -25.938 + 0.333 \times \text{income} + 1.928 \times \text{size}$$

※ SAS에서는 $p = \text{Pr}(\text{owner} = 1)$ 을 기준으로 하므로, SPSS 결과와 부호가 바뀔에 유념

로지스틱 회귀분석의 예 - SAS

< 분류 테이블(SAS 출력결과 일부) >

Classification Table									
Correct		Incorrect		Percentages					
Prob	Non-	Non-	Non-	Non-	Sensi-	Speci-	False	False	
Level	Event	Event	Event	Event	tivity	ficity	POS	NEG	
0.100	11	6	6	1	70.8	91.7	50.0	35.3	14.3
0.200	10	7	5	2	70.8	83.3	58.3	33.3	22.2
0.300	10	8	4	2	75.0	83.3	66.7	28.6	20.0
0.400	10	9	3	2	79.2	83.3	75.0	23.1	18.2
0.500	10	9	3	2	79.2	83.3	75.0	23.1	18.2
0.600	9	9	3	3	75.0	75.0	75.0	25.0	25.0
0.700	7	10	2	5	70.8	58.3	83.3	22.2	33.3
0.800	7	11	1	5	75.0	58.3	91.7	12.5	31.3
0.900	5	12	0	7	70.8	41.7	100.0	0.0	36.8

: 분류테이블. 만약 분류 규칙을 "Prob Level=0.5"로 한다면(표에서 줄친 부분), 분류 결과는 옆과 같이 정리됨.

<분류표(p=0.5 기준인 경우)>

		분류그룹		합
		Owner=1	Owner=2	
실제그룹	Owner=1	10	2	12 (83.3%)
	Owner=2	3	9	12 (75.0%)
	합	13	11	24 (79.2%)

<예측확률(SAS 출력결과) >

Obs	owner	income	size	phat
1	1	20.0	9.2	0.17463
2	1	28.5	8.4	0.43332
...
23	2	17.0	7.0	0.00112
24	2	21.0	7.4	0.00910

: $p = \Pr(\text{owner} = 1)$ 의 추정값을 나타낸 결과임

다음시간에는

13강 나무모형(1)

 수고했습니다