경영데이터분석기초

- SPSS, Excel을 활용한 통계분석 -

유 진 호

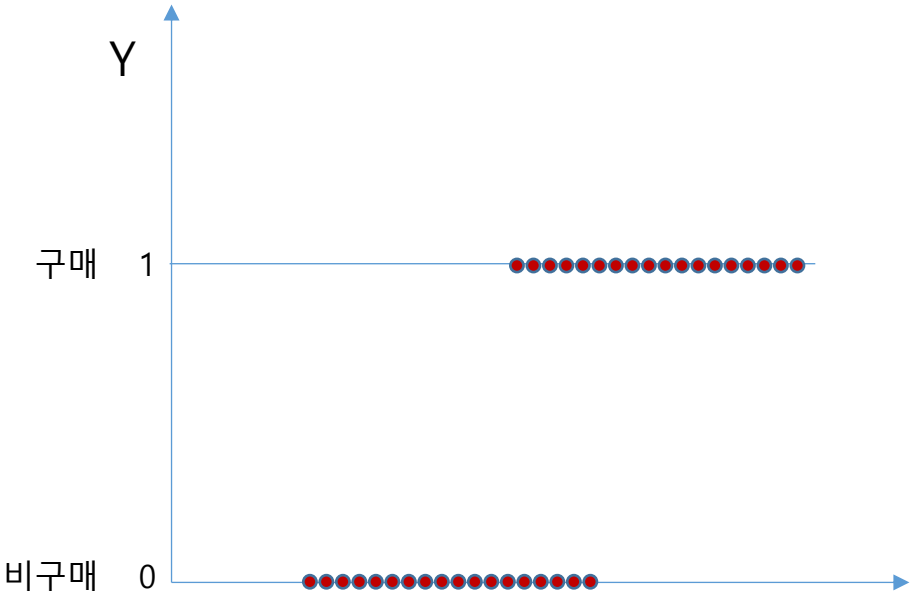
jhyoo@smu.ac.kr

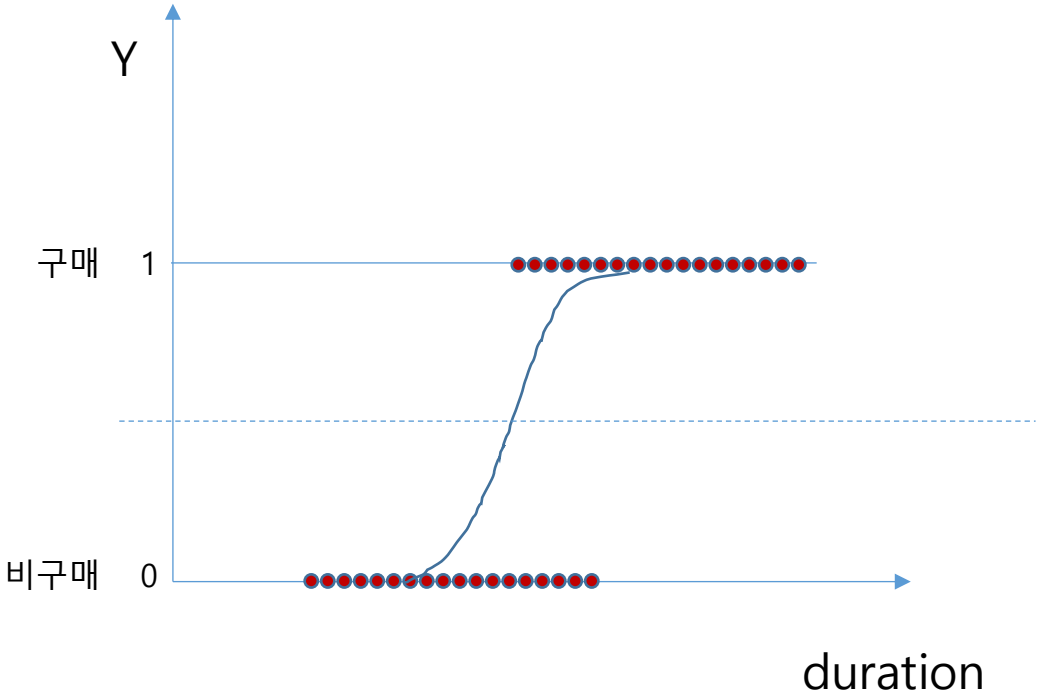
척도와 분석간의 관계		독립변수 X	
		범주형 자료	연속형 자료
종속변수 Y	범주형 자료	교차분석 (x^2검정)	로지스틱 회귀분석 판별분석
	연속형 자료	ANOVA(분산분석)	회귀분석

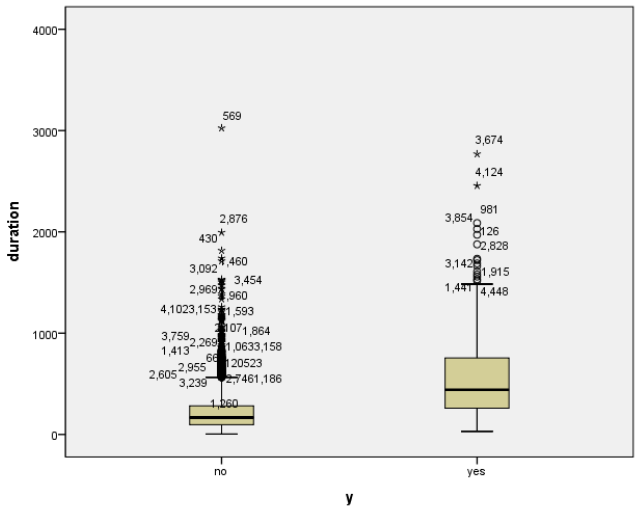
- 로지스틱 회귀(Logistic Regression)는 독립변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용하는 통계 기법이다.
- 로지스틱 회귀의 목적은 일반적인 회귀 분석의 목표와 동일하게 종속 변수와 독립 변수간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용하는 것이다.
- 이는 독립 변수의 선형 결합으로 종속 변수를 설명한다는 관점에서는 선형 회귀 분석과 유사하다.
- 하지만 로지스틱 회귀는 선형 회귀분석과는 다르게 종속 변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류(Classification) 기법으로도 볼 수 있다.

로지스틱 회귀모형은 로짓(logit) 회귀모형이라고도 불리며, 일반적으로 종속변수가 두 개의 범주를 가지는 명목형 변수인 경우 사용된다. 종속변수가 1과 0만의 값을 갖는 가변수(dummy variable)인 경우에 y의 기대값을 나타내는 반응함수의 모양이 x가 증가함에 따라 y의 값이 1로 서서히 수렴하는 S형 곡선이 되도록 추정하는데, 이와 같은 함수를 로지스틱 함수(logistic function)라 부른다.

로지스틱 회귀분석(logistic regression)에서는 단지 두 개의 값만을 가지는 종속변수(구매=1, 비구매=0)와 독립변수들 간의 인과관계를 로지스틱 함수를 이용하여 추정한다. 예를 들어, Duration(캠페인 상담 진행시간)을 보고 구매(1)할 것인지 비구매(0)할 것인지 통계적으로 예측하고자 할 때, 사용한다.







각 duration 마다 **상품 구입확률을 계산해서 S자 곡선 추정**

- **종속변수가 binary(이진 반응 변수)로 분류되는 경우 => 로지스틱 회귀 분석(Logistic Regression)**
  - 예) duration 에 따라, 구매(1)할 것인지 비구매(0)할 것인 통계적으로 분석하고자 할 때,

- **로지스틱 회귀분석에 중요한 요소 3가지**
  - ◈"Odds"라고 불리는 파라미터, ◈"로짓 변환", ◈ 출력값을 만들어내는 "시그모이드 함수"

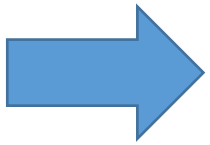
- Odds(오즈) = p/(1 − p)
  - 사건이 일어나지 않을 확률 대비 사건이 일어날 확률, Odds의 범위는 [0, ∞]

- 로짓변환: 로그변환하면 그 범위는 [-∞ , +∞]

$$\text{logit}(p) = \log \frac{p}{1-p}$$

,

$$\ell = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$



- 우리가 알고 싶은 값: **p (구매 확률)**
  - 어떤 사건이 발생할 확률(p)을 계산하는 것
  - 확률을 구하면 사건이 발생할지, 아니면 발생하지 않을 지 예측할 수 있음

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

p의 변화 모습은: S곡선

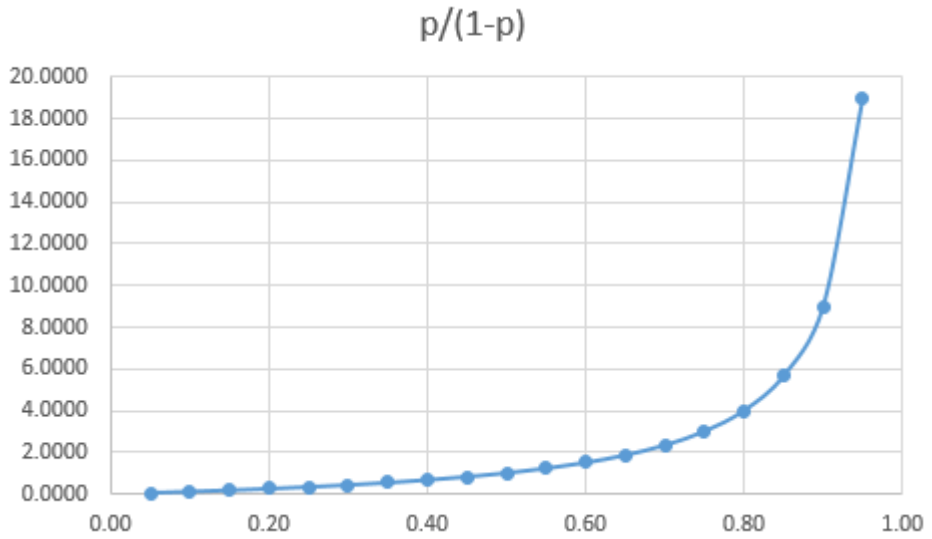
Sigmoid(시그모이드) 함수

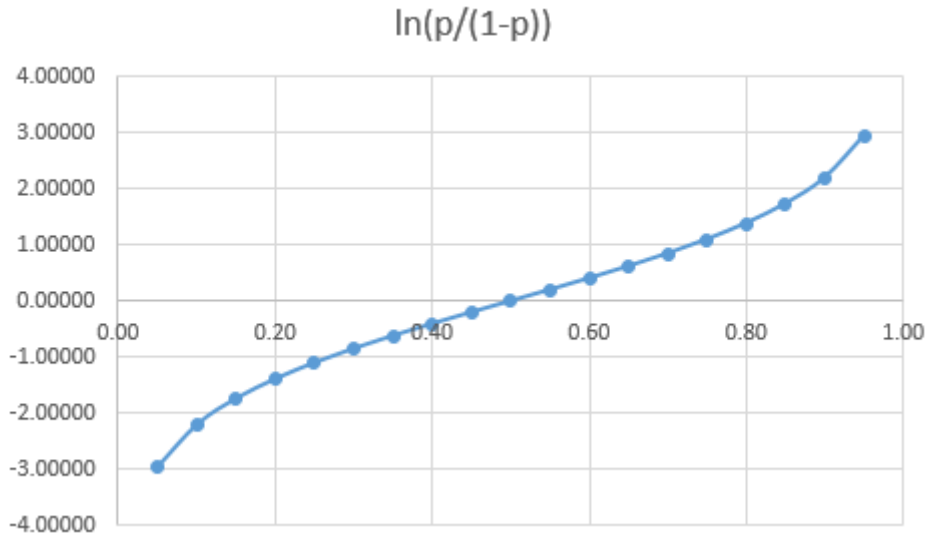
=> X 가 주어졌을 때 성공확률(p)을 예측하는 로지스틱 회귀분석은, 학습데이터를 잘 설명하는 시그모이드 함수의 B0과 B1를 찾는 문제임
.성공확률(p)을 예측한다는 것은 S 곡선을 추정한다는 것임, 마찬가지로 S 곡선을 추정한다는 p를 예측한다는 것임
.성공확률(p)를 알면, L을 안다. X를 알고 L을 알기 때문에 B0과 B1를 찾을 수 있다.

Odds

로짓 변환

p: 구매확률	1-p	p/(1-p)	ln(p/(1-p))
0.05	0.95	0.0526	-2.94444
0.10	0.90	0.1111	-2.19722
0.15	0.85	0.1765	-1.73460
0.20	0.80	0.2500	-1.38629
0.25	0.75	0.3333	-1.09861
0.30	0.70	0.4286	-0.84730
0.35	0.65	0.5385	-0.61904
0.40	0.60	0.6667	-0.40547
0.45	0.55	0.8182	-0.20067
0.50	0.50	1.0000	0.00000
0.55	0.45	1.2222	0.20067
0.60	0.40	1.5000	0.40547
0.65	0.35	1.8571	0.61904
0.70	0.30	2.3333	0.84730
0.75	0.25	3.0000	1.09861
0.80	0.20	4.0000	1.38629
0.85	0.15	5.6667	1.73460
0.90	0.10	9.0000	2.19722
0.95	0.05	19.0000	2.94444
1.00	0.00	#DIV/0!	#DIV/0!





$$\text{logit}(p) = \log \frac{p}{1-p}$$

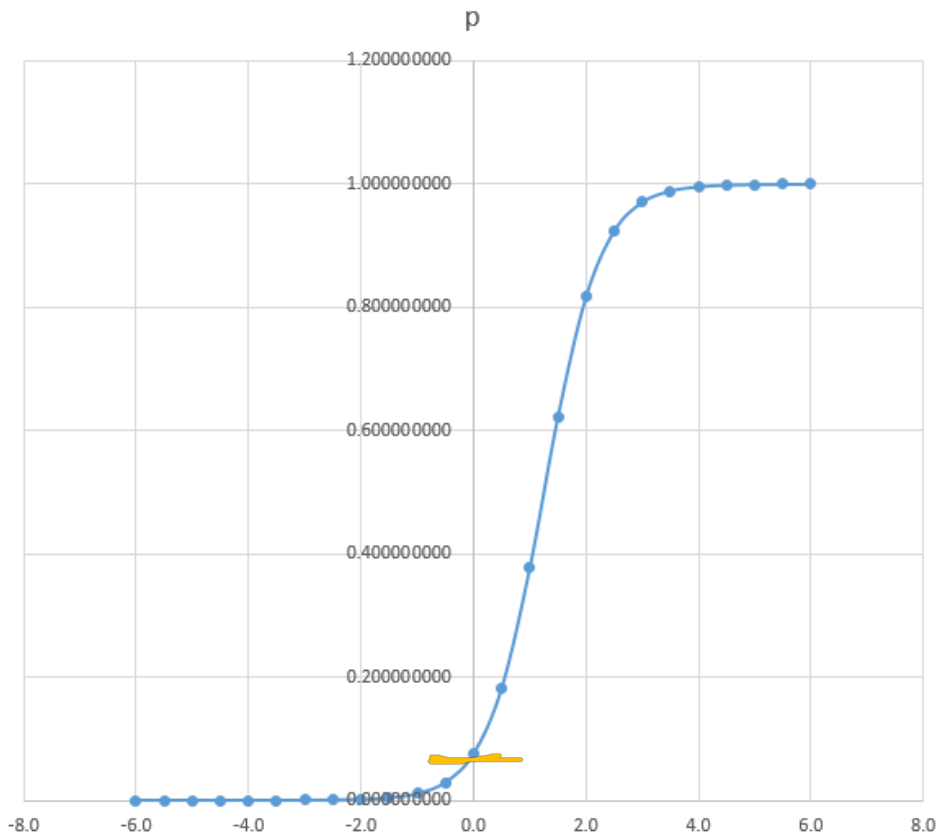
,

$$\ell = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

x	B0	B1	-(B0+B1*X)	p
-6.0	-2.5	2	14.5	0.000000504
-5.5			13.5	0.000001371
-5.0			12.5	0.000003727
-4.5			11.5	0.000010130
-4.0			10.5	0.000027536
-3.5			9.5	0.000074846
-3.0			8.5	0.000203427
-2.5			7.5	0.000552779
-2.0			6.5	0.001501182
-1.5			5.5	0.004070138
-1.0			4.5	0.010986943
-0.5			3.5	0.029312231
0.0			2.5	0.075858180
0.5			1.5	0.182425524
1.0			0.5	0.377540669
1.5			-0.5	0.622459331
2.0			-1.5	0.817574476
2.5			-2.5	0.924141820
3.0			-3.5	0.970687769
3.5			-4.5	0.989013057
4			-5.5	0.995929862
4.5			-6.5	0.998498818
5			-7.5	0.999447221
5.5			-8.5	0.999796573
6			-9.5	0.999925154

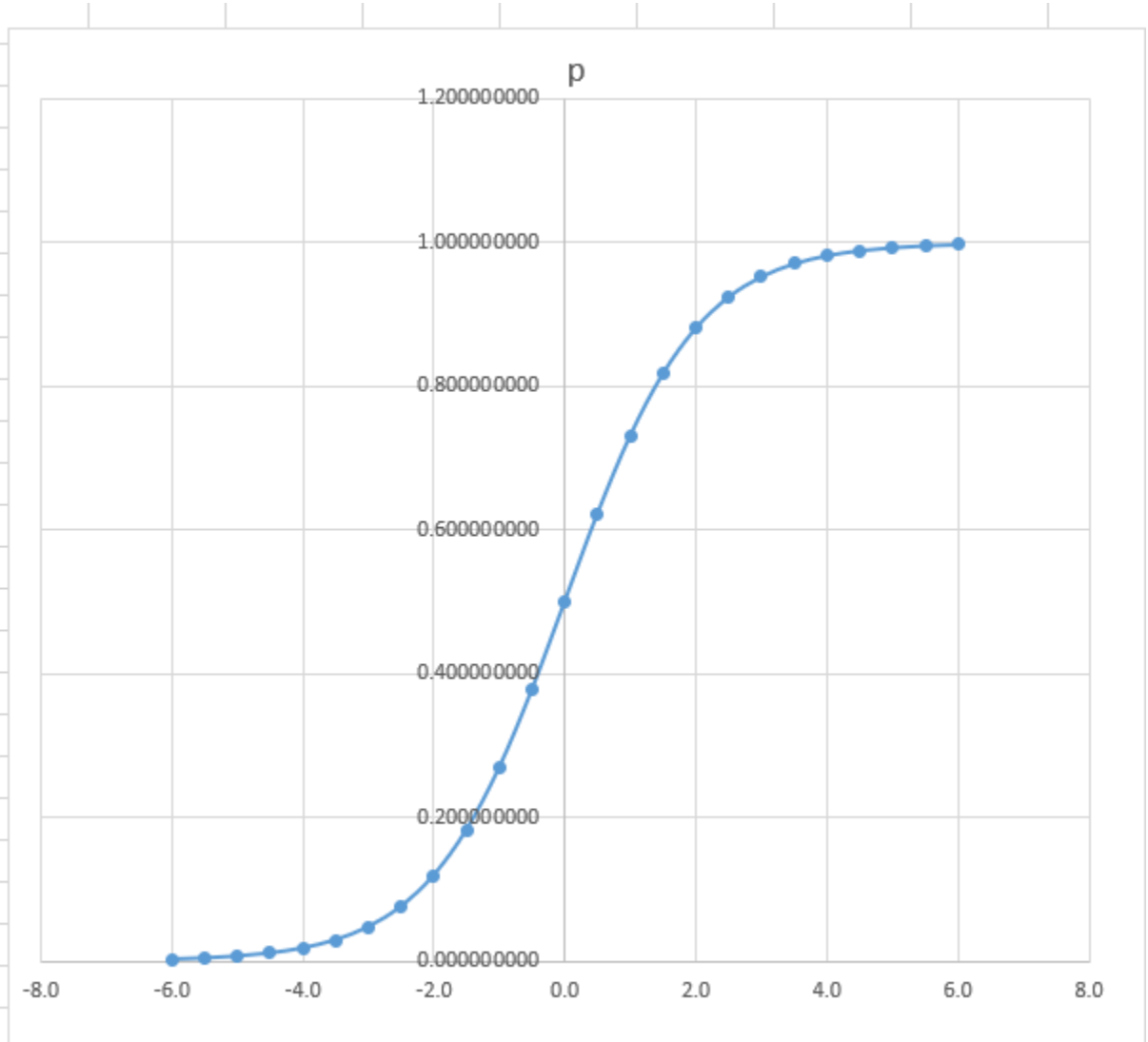
$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- 우리가 알고 싶은 값: p .
  - 어떤 사건이 발생할 확률을 계산하는 것
  - 확률을 구하면 사건이 발생할지, 아니면 발생하지 않을 지 예측할 수 있음

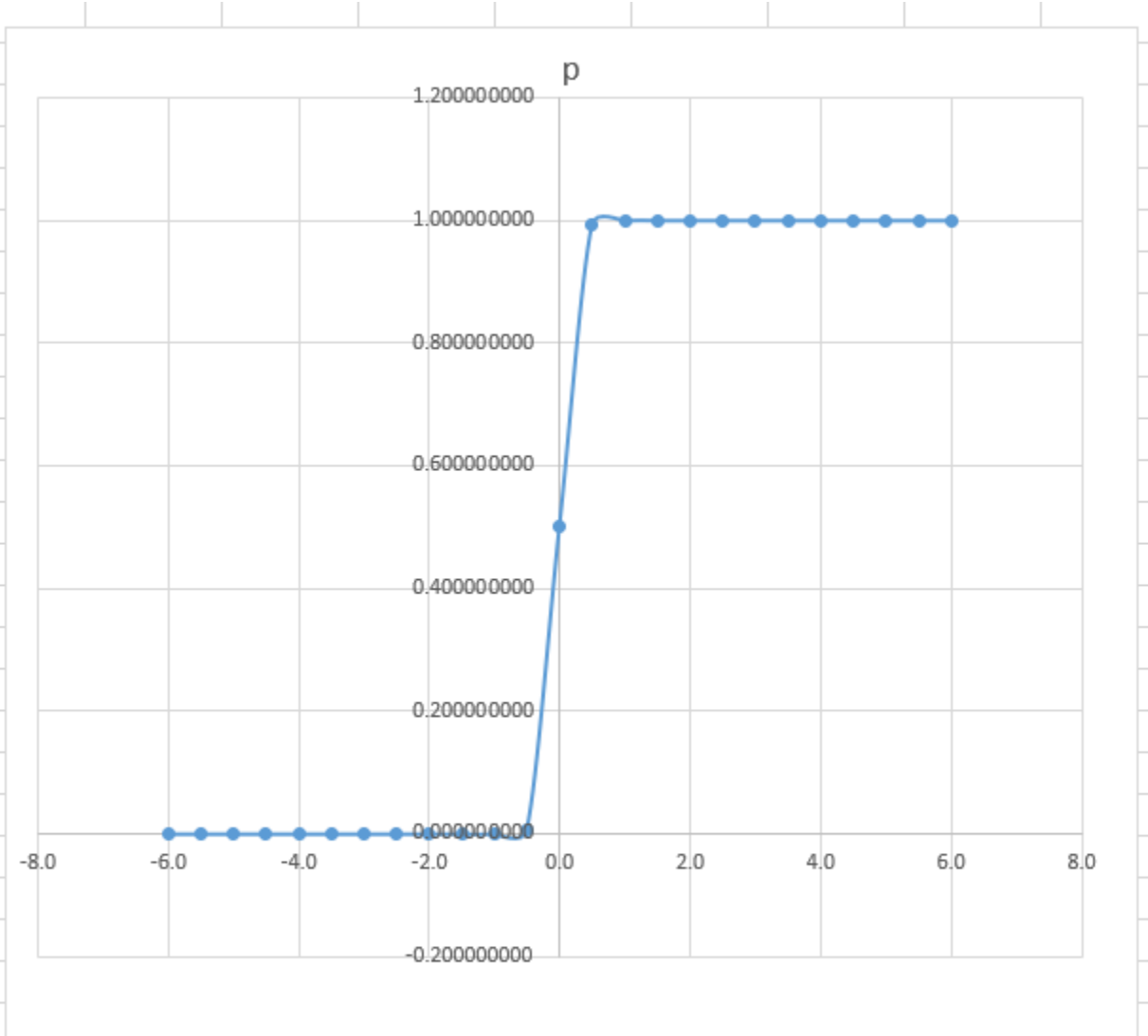


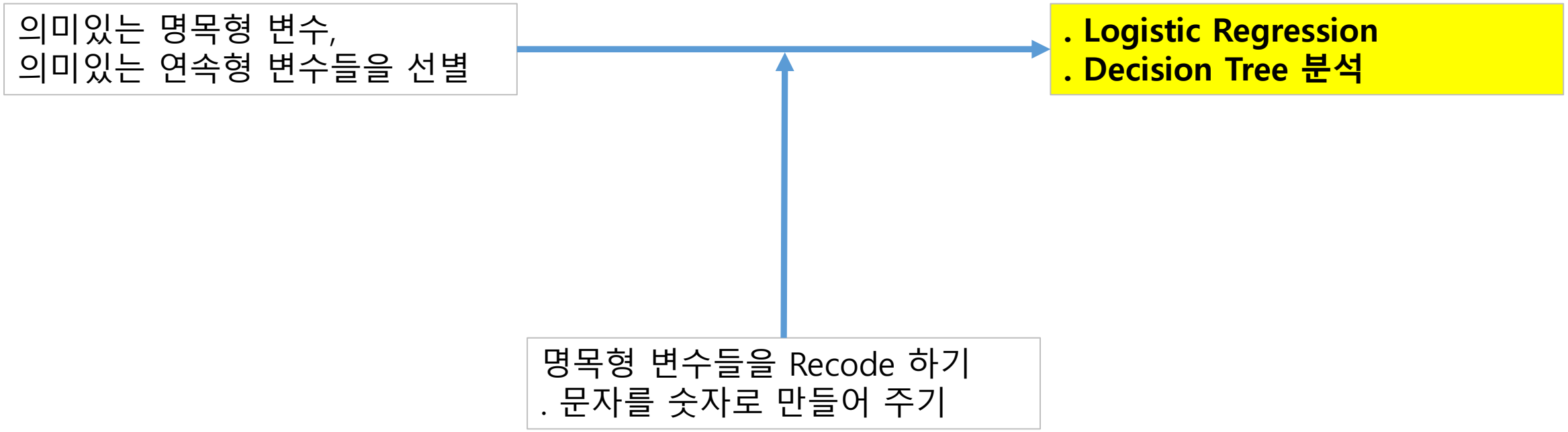
Sigmoid(시그모이드) 함수 그래프

x	B0	B1	-(B0+B1*X)	p
-6.0	0	1	6	0.002472623
-5.5			5.5	0.004070138
-5.0			5	0.006692851
-4.5			4.5	0.010986943
-4.0			4	0.017986210
-3.5			3.5	0.029312231
-3.0			3	0.047425873
-2.5			2.5	0.075858180
-2.0			2	0.119202922
-1.5			1.5	0.182425524
-1.0			1	0.268941421
-0.5			0.5	0.377540669
0.0			0	0.500000000
0.5			-0.5	0.622459331
1.0			-1	0.731058579
1.5			-1.5	0.817574476
2.0			-2	0.880797078
2.5			-2.5	0.924141820
3.0			-3	0.952574127
3.5			-3.5	0.970687769
4			-4	0.982013790
4.5			-4.5	0.989013057
5			-5	0.993307149
5.5			-5.5	0.995929862
6			-6	0.997527377



x	B0	B1	-(B0+B1*X)	p
-6.0	0	10	60	0.000000000
-5.5			55	0.000000000
-5.0			50	0.000000000
-4.5			45	0.000000000
-4.0			40	0.000000000
-3.5			35	0.000000000
-3.0			30	0.000000000
-2.5			25	0.000000000
-2.0			20	0.000000002
-1.5			15	0.000000306
-1.0			10	0.000045398
-0.5			5	0.006692851
0.0			0	0.500000000
0.5			-5	0.993307149
1.0			-10	0.999954602
1.5			-15	0.999999694
2.0			-20	0.999999998
2.5			-25	1.000000000
3.0			-30	1.000000000
3.5			-35	1.000000000
4			-40	1.000000000
4.5			-45	1.000000000
5			-50	1.000000000
5.5			-55	1.000000000
6			-60	1.000000000



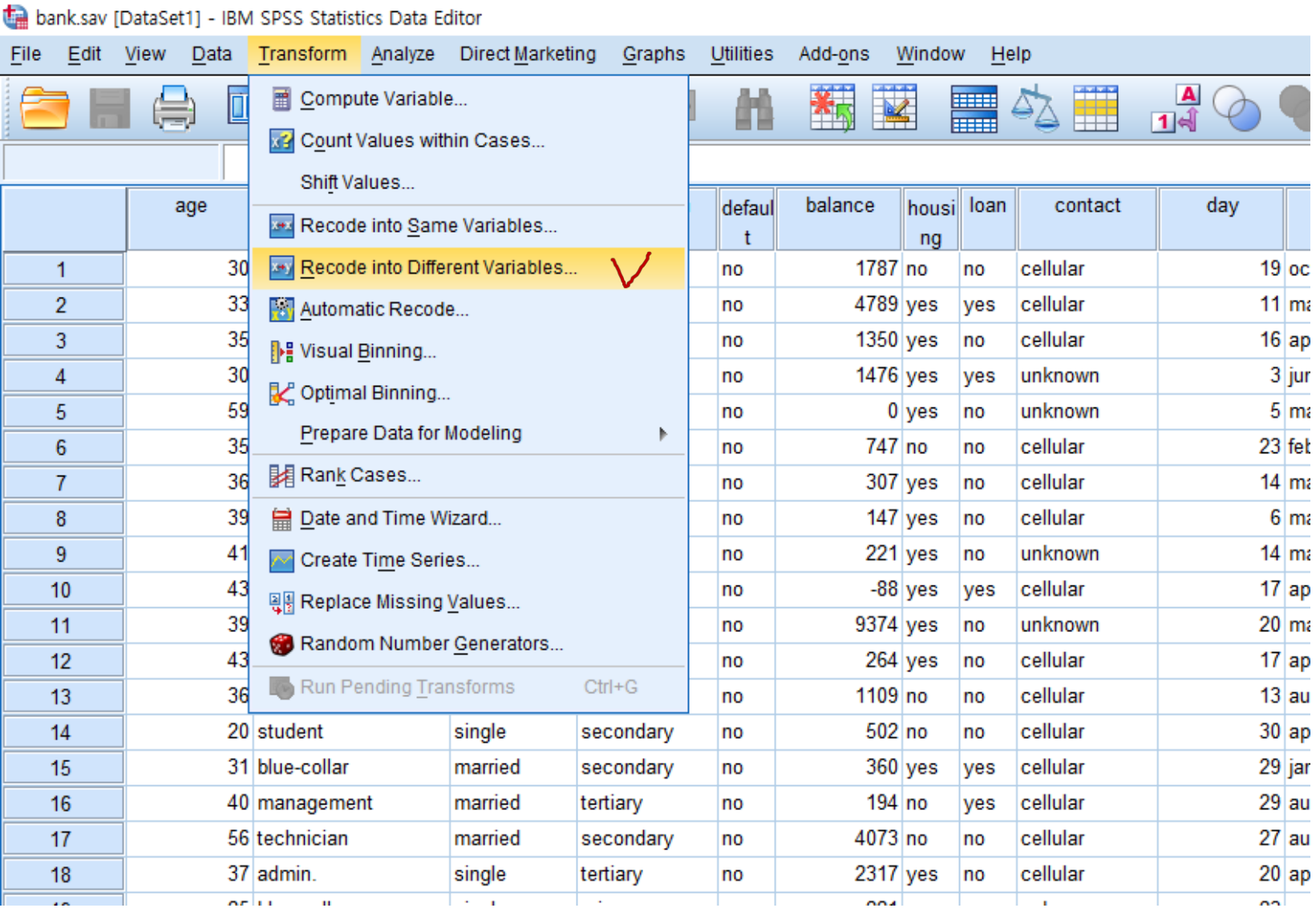

의미있는 명목형 변수,
의미있는 연속형 변수들을 선별
. Logistic Regression
. Decision Tree 분석
명목형 변수들을 Recode 하기
. 문자를 숫자로 만들어 주기

bank.sav [데이터집합1] - IBM SPSS Statistics Data Editor

파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 창(W) 도움말(H)

변수 계산(C)...
케이스 내의 값 빈도(O)...
값 이동(F)...
같은 변수로 코딩변경(S)...
다른 변수로 코딩변경(R)...
자동 코딩변경(A)...
비주얼 빈 만들기(B)...
최적의 빈 만들기(I)...
모형화를 위한 데이터 준비(P)
순위변수 생성(K)...
날짜 및 시간 마법사(D)...
시계열변수 생성(M)...
결측값 대체(V)...
난수 생성기(G)...
변환 중지(T) Ctrl+G

	age	job				balance	housing	loan	contact	day	month	duration	campaig
1	30	unemployed				1787	no	no	cellular	19	oct	79	
2	33	services				4789	yes	yes	cellular	11	may	220	
3	35	management				1350	yes	no	cellular	16	apr	185	
4	30	management				1476	yes	yes	unknown	3	jun	199	
5	59	blue-collar				0	yes	no	unknown	5	may	226	
6	35	management				747	no	no	cellular	23	feb	141	
7	36	self-employed				307	yes	no	cellular	14	may	341	
8	39	technician				147	yes	no	cellular	6	may	151	
9	41	entrepreneur				221	yes	no	unknown	14	may	57	
10	43	services				-88	yes	yes	cellular	17	apr	313	
11	39	services				9374	yes	no	unknown	20	may	273	
12	43	admin.				264	yes	no	cellular	17	apr	113	
13	36	technician				1109	no	no	cellular	13	aug	328	
14	20	student				502	no	no	cellular	30	apr	261	
15	31	blue-collar				360	yes	yes	cellular	29	jan	89	
16	40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	
17	56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	
18	37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	
19	25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	
20	31	services	married	secondary	no	132	no	no	cellular	7	jul	148	
21	38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	
22	42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	
23	44	services	single	secondary	no	106	no	no	unknown	12	jun	109	
24	44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	




```
RECODE y ('yes'=1) ('no'=0) INTO YG.
EXECUTE.
```

```
RECODE marital ('single'=1) ('married'=2) ('divorced'=3) INTO maritalG.
EXECUTE.
```

```
RECODE education ('primary'=1) ('secondary'=2) ('tertiary'=3) ('unknown'=SYSMIS) INTO eduG.
EXECUTE.
```

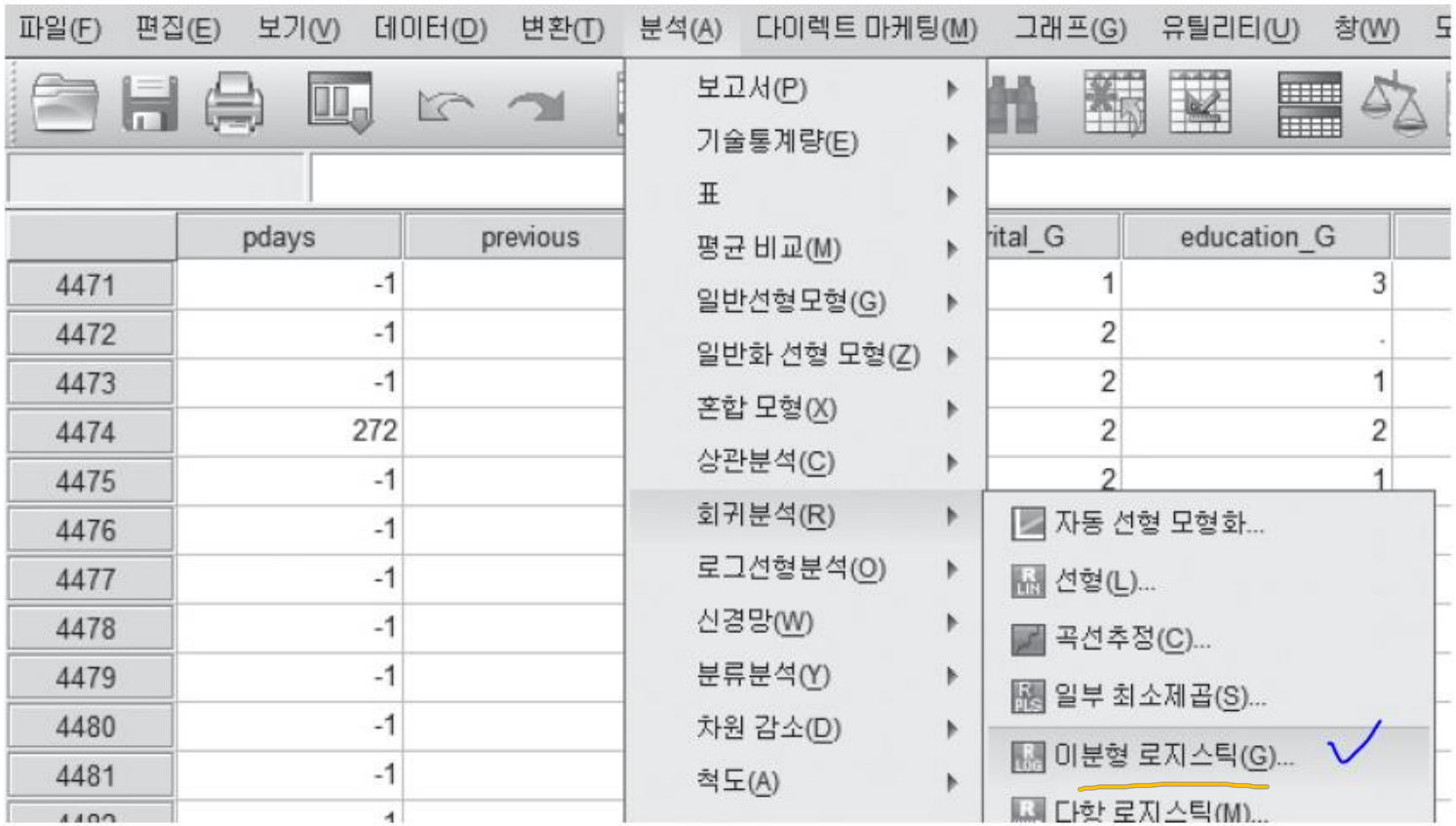
```
RECODE y ('yes'=1) ('no'=0) INTO yG.
EXECUTE.
RECODE marital ('single'=1) ('married'=2) ('divorced'=3) INTO maritalG.
EXECUTE.

RECODE education ('primary'=1) ('secondary'=2) ('tertiary'=3) ('unknown'=SYSMIS) INTO eduG.
EXECUTE.
RECODE housing ('yes'=1) ('no'=0) INTO housingG.
EXECUTE.
RECODE loan ('yes'=1) ('no'=0) INTO loanG.
EXECUTE.

RECODE contact ('cellular'=1) ('telephone'=2) ('unknown'=SYSMIS) INTO contactG.
EXECUTE.
RECODE poutcome ('failure'=0) ('success'=1) ('unknown'=SYSMIS) ('other'=SYSMIS) INTO
poutcomeG.
EXECUTE.

FREQUENCIES VARIABLES=marital maritalG education eduG housing housingG loan loanG
contact contactG poutcome poutcomeG y yG
  /ORDER=ANALYSIS.
```

① [분석]–[회귀분석]–[이분형 로지스틱]



명목형 변수,
연속형 변수들을 가지고

Logistic Regression 분석
.분류표 제공(예측력)

6개로 실습

age
pdays
duration
marital -> maritalG
Housing -> HousingG
Loan -> LoanG

⑥ 결과

분류표[a]

감시됨			예측		
			y_G		분류정확 %
			0	1	
1단계	y_G	0	3937	63	98.4
		1	442	79	15.2
	전체 퍼센트				88.8
2단계	y_G	0	3941	59	98.5
		1	430	91	17.5
	전체 퍼센트				89.2
3단계	y_G	0	3936	64	98.4
		1	427	94	18.0
	전체 퍼센트				89.1
4단계	y_G	0	3934	66	98.4
		1	423	98	18.8
	전체 퍼센트				89.2
5단계	y_G	0	3935	65	98.4
		1	421	100	19.2
	전체 퍼센트				89.3

a. 절단값은 .500입니다.

y를 예측하는데 5단계가 실행되었는데, 예측력이 가장 높은 것은 5단계이다. 이 때는 duration, pdays, marital, housing, loan 등 5개 독립변수가 사용된 것으로 나타났고, 예측력은 89.3%이다.

2단계에서 duration, housing 2개의 변수만 사용하고도 예측력이 89.2%가 나왔다. 즉, duration, housing 2개 변수로도 y를 예측하는데 충분히 활용될 수 있다는 의미이다.

특히, duration은 1~5단계에서 모두 가장 먼저 선정된 변수로 나타나, age, duration, pdays, housing, loan, marital 중 y를 예측하는데 가장 큰 영향을 주는 변수로 판정되었다. age는 1~5단계에서 모두 선정되지 않아 y를 예측하는데 영향력이 없는 변수로 판정되었다.

방정식에 포함된 변수

감시됨		B	S.E.	Wals	자유도	유의확률	Exp(B)
1단계[a]	duration	.004	.000	429.085	1	.000	1.004
	상수항	−3.256	.085	1481.990	1	.000	.039
2단계[b]	duration	.004	.000	438.564	1	.000	1.004
	housing_G(1)	.866	.106	66.762	1	.000	2.378
3단계[c]	상수항	−3.743	.112	1126.754	1	.000	.024
	duration	.004	.000	439.146	1	.000	1.004
	pdays	.004	.000	70.991	1	.000	1.004
	housing_G(1)	.993	.109	82.830	1	.000	2.700
	상수항	−4.027	.122	1088.329	1	.000	.018
4단계[d]	duration	.004	.000	439.573	1	.000	1.004
	pdays	.004	.000	67.817	1	.000	1.004
	housing_G(1)	.985	.109	80.927	1	.000	2.678
	loan_G(1)	.888	.185	22.967	1	.000	2.430
	상수항	−4.824	.214	507.712	1	.000	.008
5단계[e]	duration	.004	.000	434.924	1	.000	1.004
	pdays	.004	.000	66.508	1	.000	1.004
	marital_G			8.747	2	.013	
	marital_G(1)	−.091	.171	.283	1	.595	.913
	marital_G(2)	−.371	.159	5.470	1	.019	.690
	housing_G(1)	.976	.110	79.218	1	.000	2.655
	loan_G(1)	.882	.186	22.538	1	.000	2.415
	상수항	−4.565	.250	332.742	1	.000	.010

a. 변수가 1: 단계에 진입했습니다 duration. duration.
b. 변수가 2: 단계에 진입했습니다 housing_G. housing_G.
c. 변수가 2: 단계에 진입했습니다 pdays. pdays.
d. 변수가 2: 단계에 진입했습니다 loan_G. loan_G.
e. 변수가 2: 단계에 진입했습니다 marital_G. marital_G.

- duration이 1초 증가할 때마다 로짓이 0.004 증가한다는 의미
- duration이 1초 증가할 때마다 **구매가능성(Exp(B))**이 1.004배 증가한다는 의미

$$\ell = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

- **Exp(B)=구매가능성=오즈비**
  =구매 안 할 확률(1-p) 대비 구매할 확률(p)

- 회귀계수 B가 양수(+)면, Exp(B)는 1보다 커지기 때문에, 구매(Y=1) 가능성이 높아진다는 의미
- 회귀계수 B가 음수(-)면, Exp(B)는 1보다 작아지기 때문에 구매(Y=1) 가능성이 낮아진다는 의미

- 오즈비가 1보다 크면 증가의 의미
  (구매 안 할 확률 보다 구매할 확률이 커진다는 의미)
- 오즈비가 1보다 작으면 감소의 의미
  (구매 안 할 확률이 구매할 확률 보다 커진다는 의미)

Exp(B) = 오즈비 = p / (1-p) = 구매가능성

항이 제거된 경우의 모형[a]

변수		로그-우도 모형	-2 로그 우도에서 변경	자유도	변화량의 유의확률
1 단계	duration	-1622.984	544.216	1	.000
2 단계	duration	-1600.835	569.081	1	.000
	housingG	-1351.737	70.885	1	.000
3 단계	duration	-1573.884	579.426	1	.000
	pdays	-1316.958	65.573	1	.000
	housingG	-1328.909	89.474	1	.000
4 단계	duration	-1563.979	586.837	1	.000
	pdays	-1301.906	62.691	1	.000
	housingG	-1314.198	87.275	1	.000
	loanG	-1284.302	27.484	1	.000
5 단계	duration	-1556.241	580.020	1	.000
	pdays	-1297.069	61.677	1	.000
	housingG	-1308.888	85.316	1	.000
	loanG	-1279.682	26.903	1	.000
	maritalG	-1270.573	8.685	2	.013

a. 조건부 모수 추정값 기준

-2LL(Log-Likelihood)값 활용

. 범위(0<= -2LL < 무한대)

. 값이 클수록 classification에 영향을 준다고 할 수 있음

. 0에 가까울 수록 영향력이 없는 변수

숙제 (1개의 엑셀파일에 만들기)

- Recode 해서 분포 확인하기 => 엑셀(*.xls)로 내보내기
- 엑셀에서 오즈비, 시그모이드 함수 그리기 & 해석하기
- SPSS에서 로지스틱 회귀분석 실시하고 엑셀(*.xls)로 내보내기 후 해석하기