

빅데이터 분석을 위한
데이터마이닝 방법론
SAS Enterprise Miner 활용사례를 중심으로

<<제4장>> 회귀분석

Chapter 4 Regression Analysis

강현철, 한상태, 최종후, 이성건, 김은석, 엄익현

Update: 2014. 4. 1.

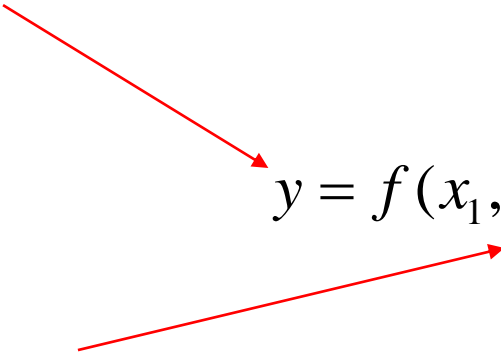
차례

- 4.1 선형 회귀분석(Linear Regression Analysis)
- 4.2 로지스틱 회귀분석(Logistic Regression Analysis)
- 4.3 회귀분석의 특징과 제약
- 4.4 분석사례 - 1: 선형 회귀분석
- 4.5 분석사례 - 2: 로지스틱 회귀분석
- 4.6 분석사례 - 3: 신용평점표 작성
- 4.7 연습문제

회귀분석(Regression Analysis)

- 반응변수(response variable)

- ✓ 목표변수(target variable)
- ✓ 종속변수(dependent variable)
- ✓ 설명(예측)되어지는 변수


$$y = f(x_1, x_2, \Lambda, x_p)$$

- 설명변수(explanatory variable)

- ✓ 입력변수(input variable)
- ✓ 독립변수(independent variable)
- ✓ 반응변수를 설명(예측)하는데 이용되는 변수

- ✓ 회귀분석이란 반응변수가 설명변수들에 의해 어떻게 설명(예측)되는지를 알아보기 위해 적절한 함수식으로 표현하여 분석하는 통계적 자료분석 방법

회귀분석의 종류

- 선형(linear) vs 비선형(nonlinear)

- ✓ 선형 회귀분석 : 반응변수와 설명변수의 관계를 선형함수로 표현
- ✓ 비선형 회귀분석 : 반응변수와 설명변수의 관계가 비선형

- 단순(simple) vs 다중(multiple)

- ✓ 단순 회귀분석 : 설명변수가 한 개
- ✓ 다중 회귀분석 : 설명변수가 두 개 이상

- 일변량(univariate) vs 다변량(multivariate)

- ✓ 일변량 회귀분석 : 반응변수가 한 개
- ✓ 다변량 회귀분석 : 반응변수가 두 개 이상

... 회귀분석의 종류

$$- y = \alpha + \beta x \quad \leftarrow \text{단순 선형회귀분석}$$

$$- y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \Lambda + \beta_p x_p \quad \leftarrow \text{다중 선형회귀분석}$$

$$- y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \gamma_{12} x_1 x_2 + \delta_1 x_1^2 + \delta_2 x_2^2 \quad \leftarrow \text{다항 회귀분석}$$

$$- y = \frac{m \exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \quad \leftarrow \text{비선형 회귀분석}$$

$$\begin{cases} y_1 = \alpha_1 + \beta_{11} x_1 + \beta_{12} x_2 + \Lambda + \beta_{1p} x_p \\ y_2 = \alpha_2 + \beta_{21} x_1 + \beta_{22} x_2 + \Lambda + \beta_{2p} x_p \\ y_3 = \alpha_3 + \beta_{31} x_1 + \beta_{32} x_2 + \Lambda + \beta_{3p} x_p \end{cases} \quad \leftarrow \text{다변량 회귀분석}$$

회귀(Regression)

Table 8.1. Galton's 1885 cross-tabulation of 928 adult children born of 205 midparents, by their height and their midparent's height.

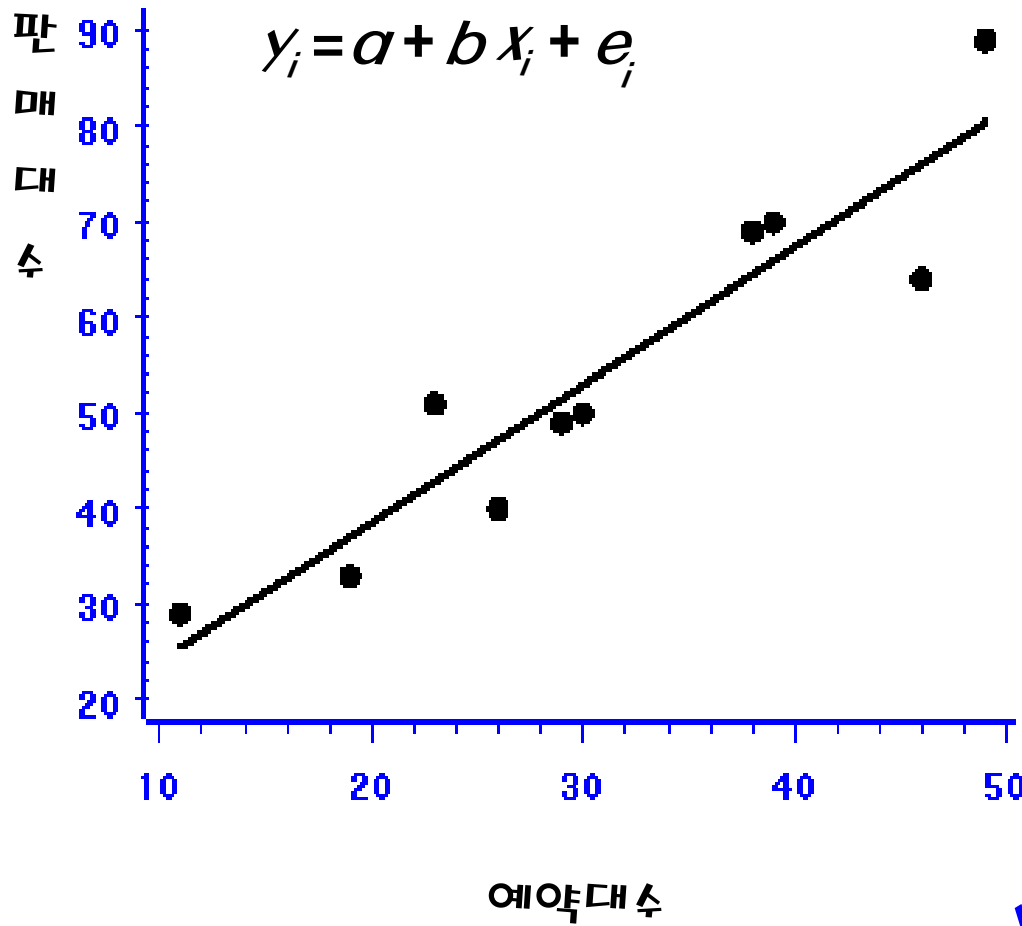
Height of the mid-parent in inches	Height of the adult child														Total no. of adult children	Total no. of mid-parents	Medians
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7			
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—	—	—	—

Source: Galton (1886a).

Note: All female heights were multiplied by 1.08 before tabulation. Galton added an explanatory footnote to the table: "In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents." Galton republished these data in 1889, where they are referred to as the R.F.F. Data (Record of Family Faculties); he then noted that the first row must be in error (four children cannot have five sets of parents), but he claimed that "the bottom line, which looks suspicious, is correct" (p. 208).

Francis Galton(1822~1911) : 아버지의 키와 아들의 키의 관계를 연구

4.1.1 단순 회귀모형(Simple Regression)



x	y	\hat{y}
11	29	25.5
19	33	37.1
23	51	42.8
26	40	47.1
29	49	51.5
30	50	52.9
38	69	64.4
39	70	65.9
46	64	76.0
49	89	80.3

$$y_i = a + b x_i = 9.74 + 1.44 x_i$$

회귀계수(모수)의 추정

단순 선형 회귀모형 $\rightarrow y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \Lambda, n$

추정된 회귀직선 $\rightarrow \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i = a + b x_i, \quad i = 1, \Lambda, n$

$$e_i = y_i - \hat{y}_i \quad \leftarrow \text{잔차 (residual)}$$

- 최소제곱추정 (Least Square Estimation)

$$\text{Min} \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - b x_i)^2$$

$$\Rightarrow b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad a = \bar{y} - b \bar{x}$$

회귀계수에 대한 해석과 검정

표 4.1 단순 회귀분석의 결과

Variable	DF	Parameter	Standard	T for H0:	
		Estimate	Error	Parameter=0	Prob > T
INTERCEP	1	9.736154	6.62064516	1.471	0.1796
X	1	1.440769	0.20044164	7.188	0.0001

- $H_0: \beta=0$

$$t = \frac{b - \beta}{s.e.(b)}$$

- ✓ 자유도 $n-1$ 인 t -분포를 따른다.
- ✓ $s.e.(b)$ 는 b 의 표준오차(standard error)이다.

4.1.2 다중 회귀모형(Multiple Regression)

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \Lambda + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \Lambda, n$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \mathbf{M} \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \Lambda & x_{1p} \\ 1 & x_{21} & \Lambda & x_{2p} \\ \mathbf{M} & \mathbf{M} & \Lambda & \mathbf{M} \\ 1 & x_{n1} & \Lambda & x_{np} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \mathbf{M} \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \mathbf{M} \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

<<사례>> 영업수익 평가지수

Correlation						
Variable	Label	x1	x2	x3	x4	y
x1	창의력	1.0000				
x2	단순추론능력	0.6010	1.0000			
x3	복합추론능력	0.1032	0.4208	1.0000		
x4	계량능력	0.3937	0.5746	0.5477	1.0000	
y	영업수익 평가지수	0.5310	0.7459	0.4982	0.9443	1.0000

어떤 회사에서는 신입사원에 대해 4과목(x_1 =창의력, x_2 =단순추론능력, x_3 =복합추론능력, x_4 =계량능력)의 적성검사를 실시하여 왔다. 이 회사에서는 이러한 적성검사 과목들이 사원의 업무능력을 평가하는데 타당하지를 알아보기 위하여 입사 후 일년 간의 실적을 평가하여 ‘업무능력지수(y)’를 산출하였다.

분산분석표 및 회귀계수 추정치

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4816.9644	1204.24110	263.55	<.0001
Error	45	205.6214	4.56936		
Corrected Total	49	5022.5858			

Root MSE	2.1376	R-Square	0.9591
Dependent Mean	106.6220	Adj R-Sq	0.9554
Coeff Var	2.0048		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	Intercept	1	73.15526	1.68258	43.48	<.0001	0
x1	창의력	1	0.14245	0.10157	1.40	0.1676	0.05498
x2	단순추론능력	1	0.84501	0.13186	6.41	<.0001	0.28250
x3	복합추론능력	1	-0.27220	0.16825	-1.62	0.1127	-0.06116
x4	계량능력	1	0.76269	0.03949	19.31	<.0001	0.79383

분산분석표(ANOVA Table)

- 제곱합의 분할

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

TSS	=	SSR	+	SSE
전체제곱합		회귀제곱합		오차제곱합

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}$$

- 분산분석표 (ANOVA table)

요인	제곱합	자유도	평균제곱	분산비
회귀	SSR	p	MSR=SSR/p	F=MSR/MSE (p-value)
오차	SSE	n-p-1	MSE=SSE/(n-p-1)	
전체	TSS	n-1		

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

회귀계수에 대한 검정

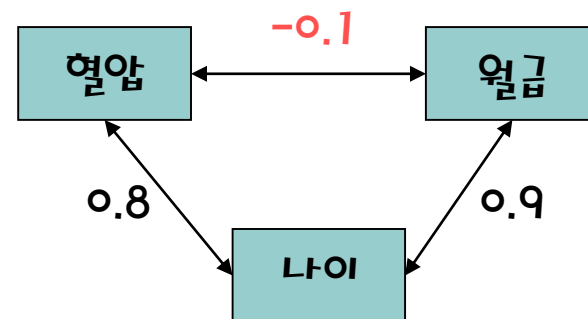
- 회귀계수에 대한 검정

$$H_0 : \beta_j = 0 \leftarrow t = b_j / \text{s.e.}(b_j) \sim t(n - p - 1)$$

- 표준화 회귀계수

$$y = \alpha^* + \beta_1^* z_1 + \beta_2^* z_2 + \Lambda + \beta_p^* z_p + \varepsilon^* \leftarrow z_j = (x_j - \bar{x}_j) / s_j$$

- 편상관계수(partial correlation coefficient)

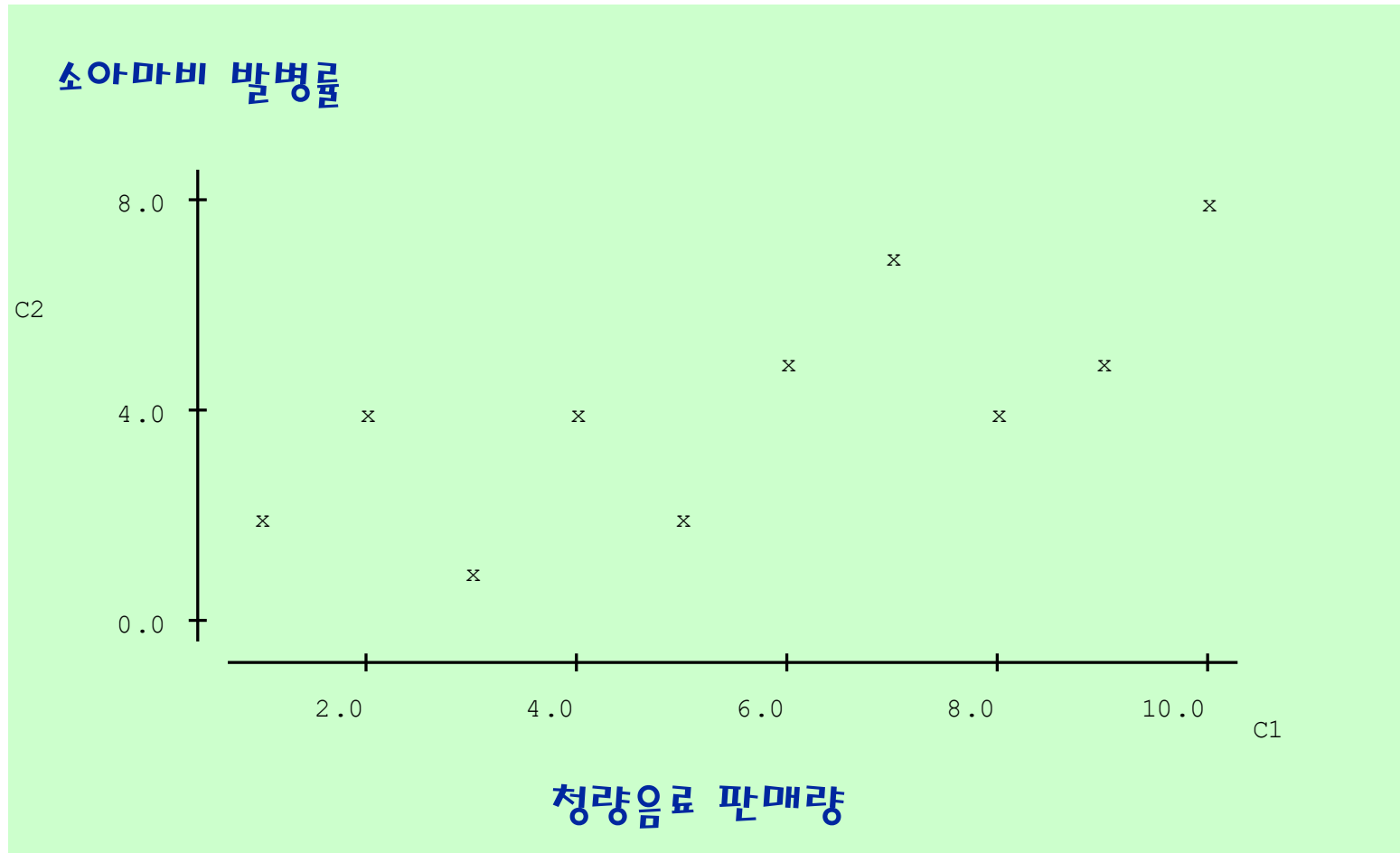


<<예>> 다중 회귀분석의 결과

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	6	722.54361	120.42393	22.433	0.0001
Error	24	128.83794	5.36825		
C Total	30	851.38154			
Root MSE		2.31695	R-square	0.8487	
Dep Mean		47.37581	Adj R-sq	0.8108	
C.V.		4.89057			

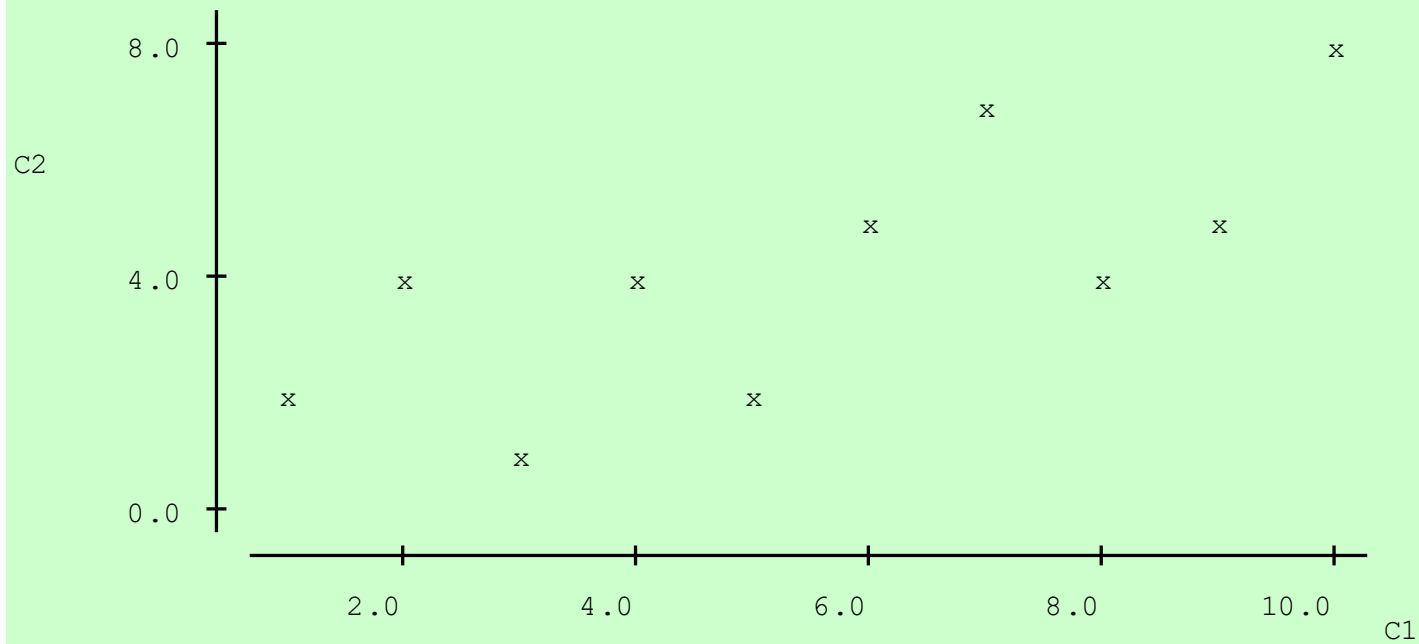
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	102.934479	12.40325810	8.299	0.0001
RUNTIME	1	-2.628653	0.38456220	-6.835	0.0001
AGE	1	-0.226974	0.09983747	-2.273	0.0322
WEIGHT	1	-0.074177	0.05459316	-1.359	0.1869
RUNPULSE	1	-0.369628	0.11985294	-3.084	0.0051
MAXPULSE	1	0.303217	0.13649519	2.221	0.0360
RSTPULSE	1	-0.021534	0.06605428	-0.326	0.7473

매개변수(Lurking Variables)

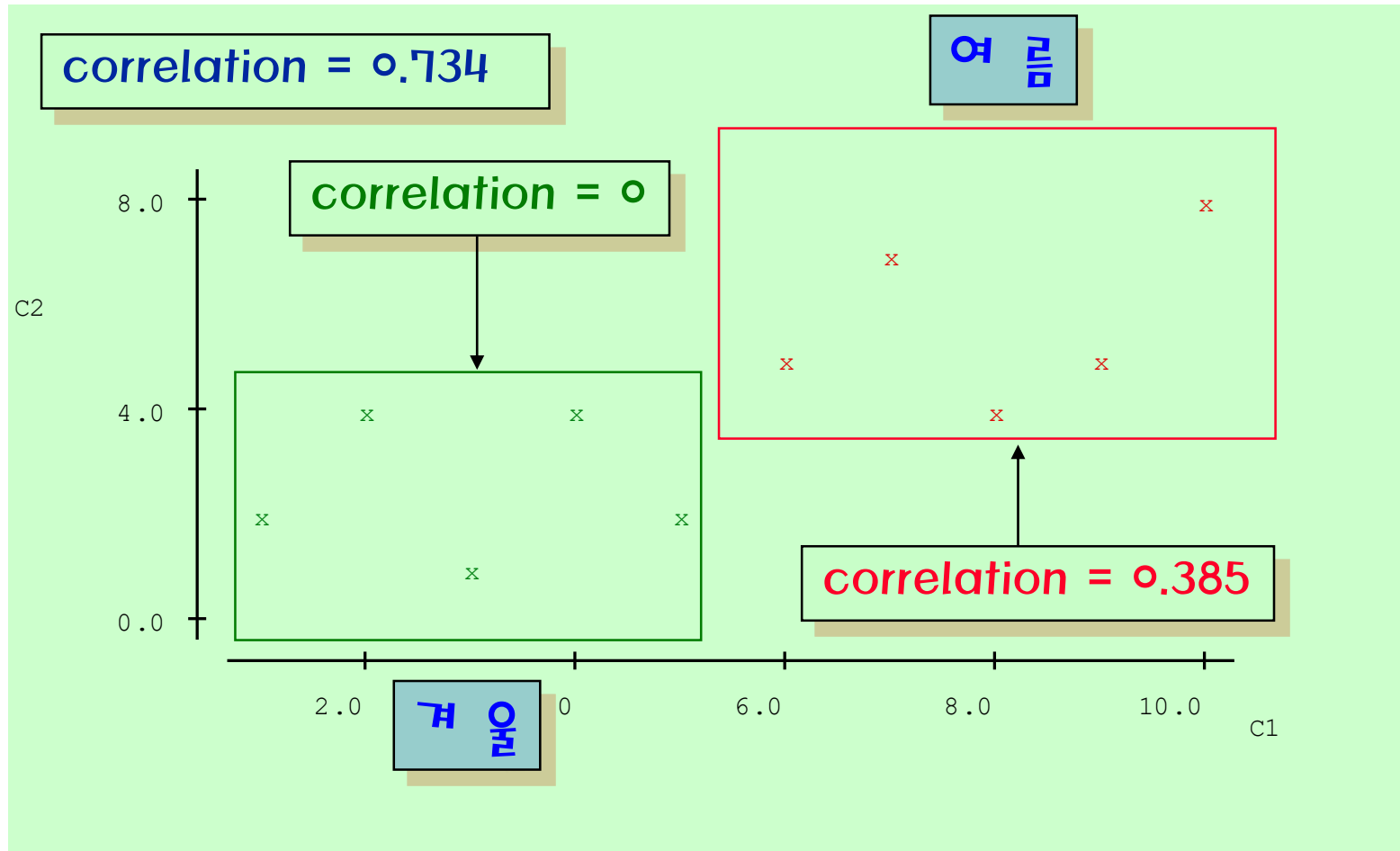


... 매개변수(Lurking Variables)

correlation = 0.734



... 매개변수(Lurking Variables)



입력변수의 선택

- 전진선택법 (Forward Selection)

- ✓ 입력변수를 각 변수의 기여도에 따라서 하나씩 추가하면서 선택하는 방법이다. 이 방법은 계산시간이 빠르다는 장점이 있지만, 한 번 선택된 변수는 절대로 제거되지 않는다는 단점이 있다.

- 후진소거법 (Backward Elimination)

- ✓ 모든 변수를 포함하는 완전모형으로부터 시작하여 불필요한 변수를 하나씩 제거해 나가는 방법이다. 이 방법은 중요한 변수가 모형에서 제외될 가능성이 적으므로 비교적 안전한 방법이라 할 수 있다. 그러나 한 번 제외된 변수는 다시 선택되지 못한다는 단점이 있다.

- 단계적 방법 (Stepwise Method)

- ✓ 전진선택법에 후진소거법을 결합한 것으로서, 매 단계마다 선택과 제거를 반복하면서 중요한 변수를 찾아내는 방법이다. 이 방법은 중요한 변수를 하나씩 추가로 선택하면서 이미 선택된 변수들이 제거될 수 있는 지를 매 단계마다 검토하는 방법이다. 그러나 이 방법에 의해서 찾아진 모형도 모든 가능한 회귀를 통해서 얻어진 모형들보다 못할 수 있다.

- 모든 가능한 회귀

- ✓ 가능한 모든 축소모형을 고려하여 가장 좋은 모형을 찾아내는 방법이다. 이 방법은 가장 안전한 방법이라고 할 수 있지만, 입력변수가 많은 경우에는 탐색시간이 매우 많이 걸리며 현실적으로 사용하기 어려운 경우가 종종 있다.

변수선택 요약

Stepwise Selection: Step 1

Variable x4 Entered: R-Square = 0.8917 and C(p) = 73.0476

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4478.61411	4478.61411	395.19	<.0001
Error	48	543.97169	11.33274		
Corrected Total	49	5022.58580			

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4		계량능력	1	0.8917	0.8917	73.0476	395.19	<.0001
2	x2		단순추론능력	2	0.0617	0.9534	7.1888	62.31	<.0001
3	x3		복합추론능력	3	0.0038	0.9573	4.9670	4.13	0.0478
4	x1		창의력	4	0.0018	0.9591	5.0000	1.97	0.1676
5		x1	창의력	3	0.0018	0.9573	4.9670	1.97	0.1676

분산분석표 및 회귀계수 추정치

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4807.97642	1602.65881	343.52	<.0001
Error	46	214.60938	4.66542		
Corrected Total	49	5022.58580			

Root MSE	2.15996	R-Square	0.9573
Dependent Mean	106.62200	Adj R-Sq	0.9545
Coeff Var	2.02581		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	Intercept	1	73.70797	1.65288	44.59	<.0001	0
x2	단순추론능력	1	0.94356	0.11274	8.37	<.0001	0.31545
x3	복합추론능력	1	-0.33374	0.16413	-2.03	0.0478	-0.07498
x4	계량능력	1	0.77258	0.03927	19.68	<.0001	0.80412


차례

- 4.1 선형 회귀분석(Linear Regression Analysis)
- 4.2 로지스틱 회귀분석(Logistic Regression Analysis)
- 4.3 회귀분석의 특징과 제약
- 4.4 분석사례 - 1: 선형 회귀분석
- 4.5 분석사례 - 2: 로지스틱 회귀분석
- 4.6 분석사례 - 3: 신용평점표 작성
- 4.7 연습문제

4.2.1 로지스틱 단순 회귀모형

- 목표변수가 이항형 또는 다항형으로 나타나는 경우가 있다. 예를 들어, 소비자가 어떤 상품을 구입할 것인지 아닌지(구입=1, 구입하지 않음=0)를 나타내는 변수는 이항형이고, 고객의 신용등급(A=매우 좋음, B=좋음, C=좋지 않음, D=매우 좋지 않음)을 나타내는 변수는 다항형이다.

$$y = 0.1 + 0.01x$$

x	y		\hat{y}
10	0		0.2
100	1		1.1
1000	1		10.1

- 로지스틱 회귀분석

$$\log \frac{P(y=1|x)}{1-P(y=1|x)} = \alpha + \beta x + \varepsilon$$

<<사례>> 독성실험 자료

번호	용량(g)	사망유무
1	0	무
2	0	무
3	0	무
4	0	무
5	1	아
6	1	무
7	1	무
8	1	무
9	2	무
10	2	아
11	2	아
12	2	아
13	3	아
14	3	아
15	3	아
16	3	아

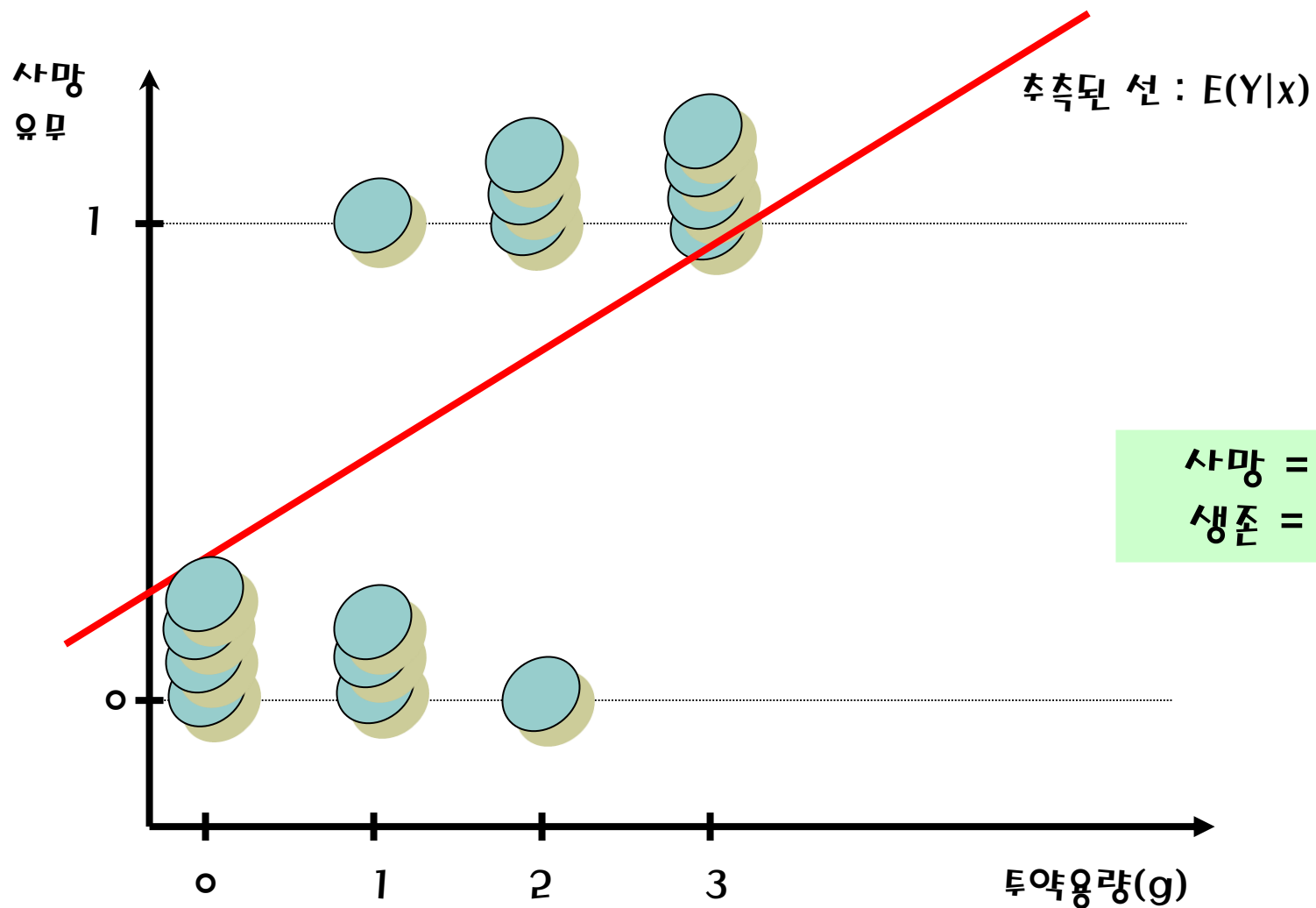
분석목적

약의 성분 → 사망유무

X

Y

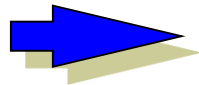
잘못된 분포가정



$P(Y|x=1)$

용량(x)	실험대상수	사망수(Y')	사망비율
0	4	0	0
1	4	1	1/4
2	4	2	2/4
3	4	4	1

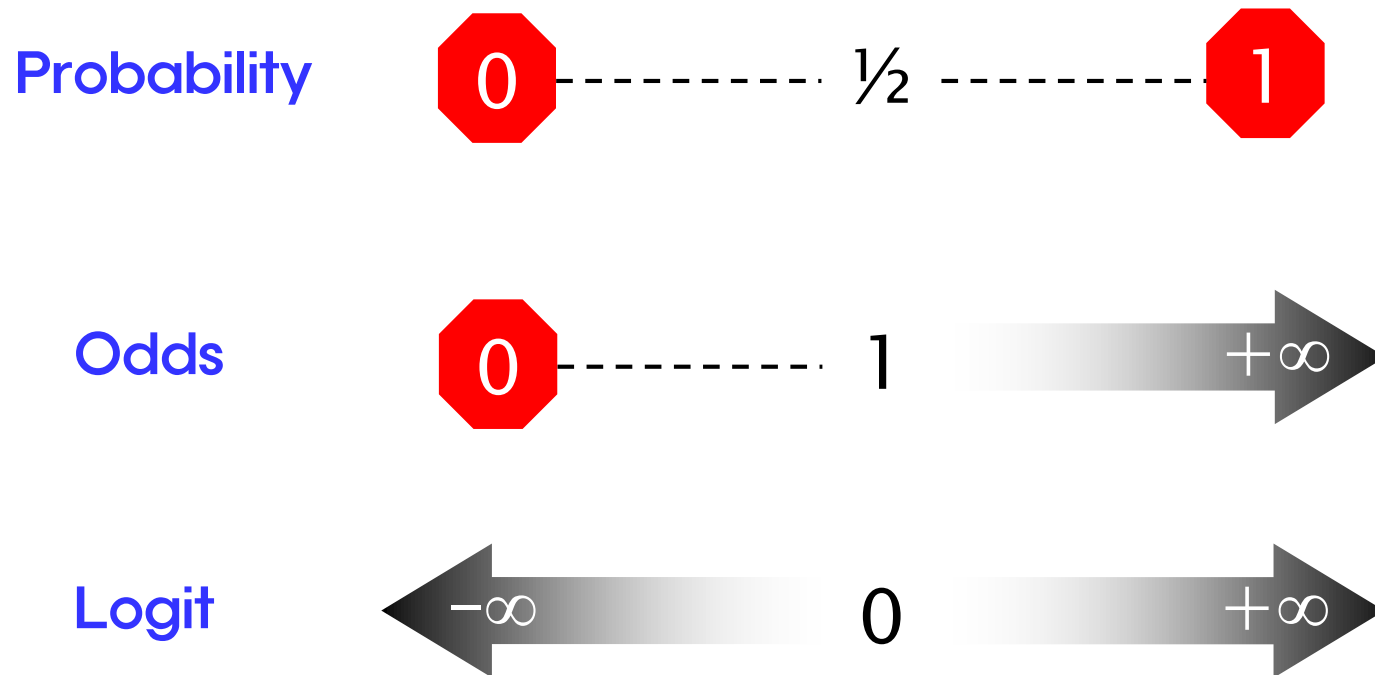
$$1/4 = P(Y/x=1)$$



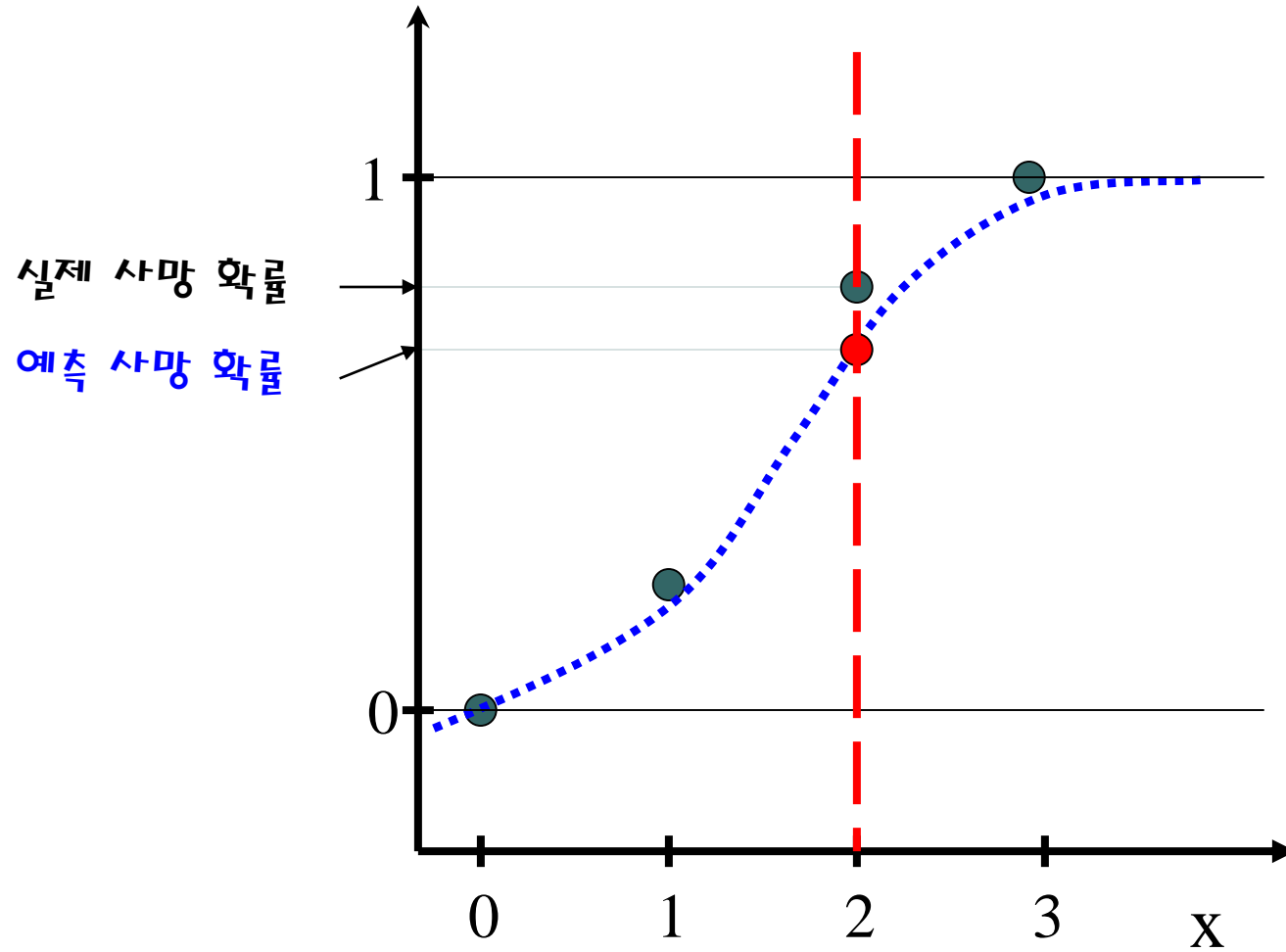
$P(Y/x=1)$ 를 x 에 의해 쉽게 설명한다면....

로짓모형

$$\text{logit}(P) = \log \text{ odds} = \ln\left(\frac{P}{1-P}\right)$$



$$\hat{P}(Y|x=1)$$



4.2.2 로지스틱 회귀분석

$$\log \frac{P(y = 1|x_1, \dots, x_p)}{1 - p(y = 1|x_1, \dots, x_p)} = \alpha + \beta_1 x_1 + \dots + \beta_p x_p$$

$$\hat{P}(y = 1|x_1, \dots, x_p) = \frac{\exp(a + b_1 x_1 + \dots + b_p x_p)}{1 + \exp(a + b_1 x_1 + \dots + b_p x_p)}$$

- **오즈비(Odds Ratio)**

$$Odds\ Ratio = \frac{\exp[\alpha + \beta_1 x_1 + \dots + \beta_i(x_i + 1) + \dots + \beta_p x_p]}{\exp[\alpha + \beta_1 x_1 + \dots + \beta_i(x_i) + \dots + \beta_p x_p]} = \exp(\beta_i)$$

- ✓ 오즈비가 1 보다 작다(계수가 음의 값을 갖는다)는 것은 입력변수 x 가 감소방향으로 영향을 미침을 의미하고, 반대로 오즈비가 1 보다 크다(계수가 양의 값을 갖는다)는 것은 증가방향으로 영향을 미침을 의미한다.
- ✓ 예를 들어, 월수입 x (단위 100만원)를 입력변수로 하고 어떤 상품에 대한 구입 여부(1 =구입, 0 =구입하지 않음) y 를 목표변수로 하여 분석하는 경우에 $b=3.73$ 이라고 해보자. 이는 x 가 1단위(백만원) 증가하면 구매하지 않을 확률에 대한 구매할 확률의 상대비가 $\exp(3.73)=42$ 배 증가한다는 것을 의미한다.

<<사례>> 신용평가 문제

대출금	대출금잔액	담보금	대출사유	직업	근무년수	신용거래수	신용상태	최초신용	P(나쁨)	P(좋음)
2300	102370	120953	Homelmp	Office	2	13	0	91	0.04	0.96
2400	34863	47471	Homelmp	Mgr	12	21	1	70	0.14	0.86
2400	98449	117195	Homelmp	Office	4	13	0	94	0.03	0.97
2900	103949	112505	Homelmp	Office	1	13	0	96	0.03	0.97
2900	104373	120702	Homelmp	Office	2	13	0	102	0.03	0.97
2900	7750	67996	Homelmp	Other	16	8	1	122	0.68	0.32
2900	61962	70915	DebtCon	Mgr	2	37	1	283	0.19	0.81
3000		14500	Homelmp	Other	3	2	1	9		
3000		14100	Homelmp	Other	1	19	1	104		
3200	74864	87266	Homelmp	ProfExe	7	12	0	251	0.08	0.92
3200	23159		Homelmp	Mgr	20	9	1	118		
3800		73189					0			
3300	130518	164317	DebtCon	Other	9	33	1	192	1.00	0.00
3600	52337	63989	Homelmp	Office	20	20	0	204	0.00	1.00
$\hat{P}(\text{신용상태} = \text{좋음}) = \frac{\exp(1.7 + 2.3X_1 - 0.45X_2 + \dots)}{1 + \exp(1.7 + 2.3X_1 - 0.45X_2 + \dots)}$									0.97	1.00
4000	54543	61777	Homelmp	Office	21	19	0	206	0.01	0.99
4000	26572	31960	Homelmp	Office	11	8	1	118	0.10	0.90
4100	57992	63797	DebtCon	ProfExe	7	31	0	166	0.22	0.78
4200	56544	59218	Homelmp	Office	19	20	0	211	0.00	1.00

차례

- 4.1 선형 회귀분석(Linear Regression Analysis)
- 4.2 로지스틱 회귀분석(Logistic Regression Analysis)
- 4.3 회귀분석의 특징과 제약
- 4.4 분석사례 - 1: 선형 회귀분석
- 4.5 분석사례 - 2: 로지스틱 회귀분석
- 4.6 분석사례 - 3: 신용평점표 작성
- 4.7 연습문제

회귀분석의 특징

- **장점**

- ✓ 친밀성(familiarity)
- ✓ 실제성(feasibility)
- ✓ 해석상의 편리(interpretability)

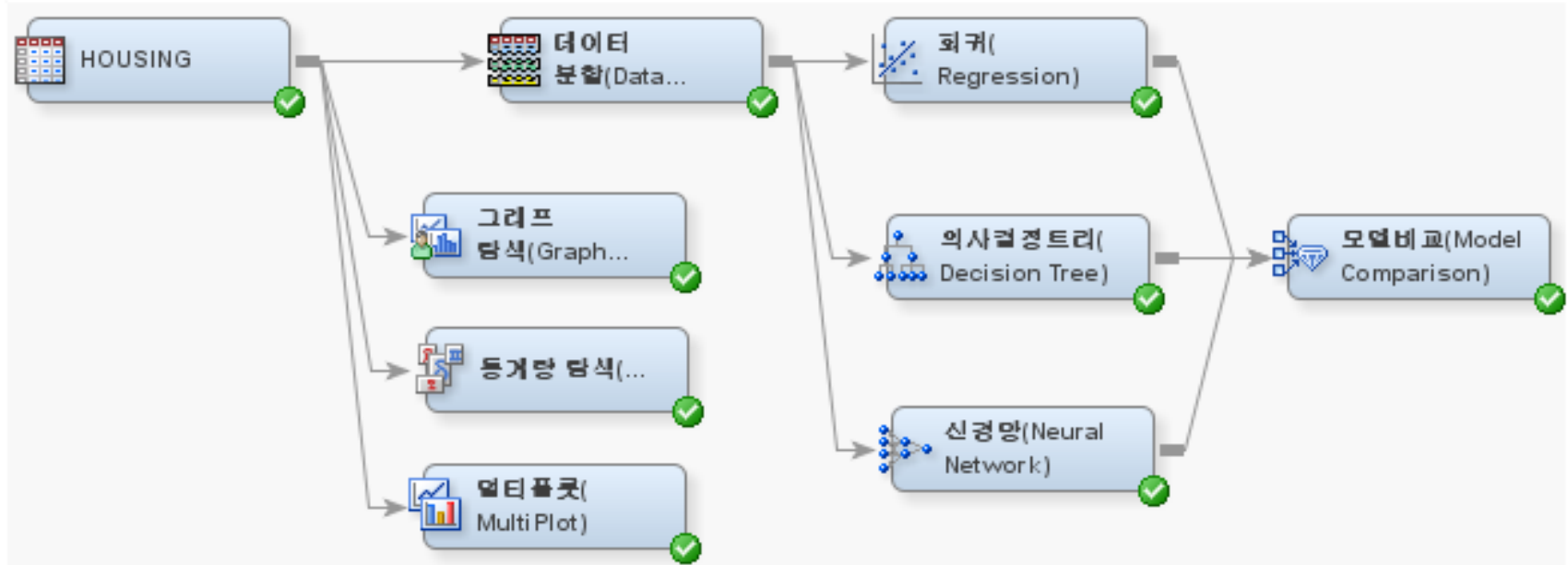
- **단점과 대안**

- ✓ 부적절하거나 불필요한 입력변수 : 변수선택방법 사용
- ✓ 선형성 : 다항회귀모형, 의사결정나무분석, 신경망분석 등 사용
- ✓ 교호작용의 결여 : 다항회귀모형, 의사결정나무분석 등 사용
- ✓ 명목형 변수 : 가변수(dummy variable) 사용
- ✓ 결측값 : 대체(imputation)

차례

- 4.1 선형 회귀분석(Linear Regression Analysis)
- 4.2 로지스틱 회귀분석(Logistic Regression Analysis)
- 4.3 회귀분석의 특징과 제약
- 4.4 분석사례 - 1: 선형 회귀분석
- 4.5 분석사례 - 2: 로지스틱 회귀분석
- 4.6 분석사례 - 3: 신용평점표 작성
- 4.7 연습문제

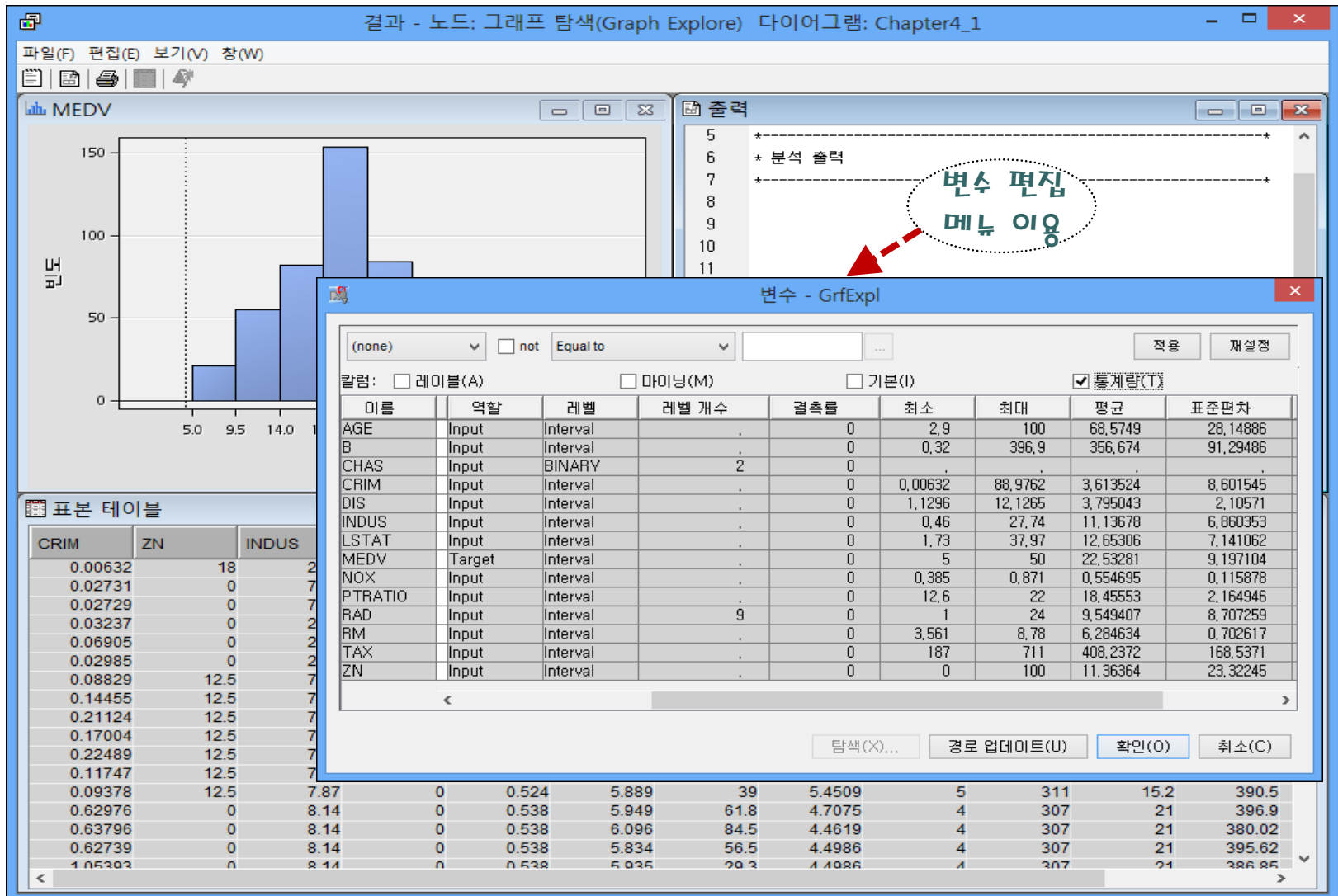
4.4.1 분석흐름도 작성



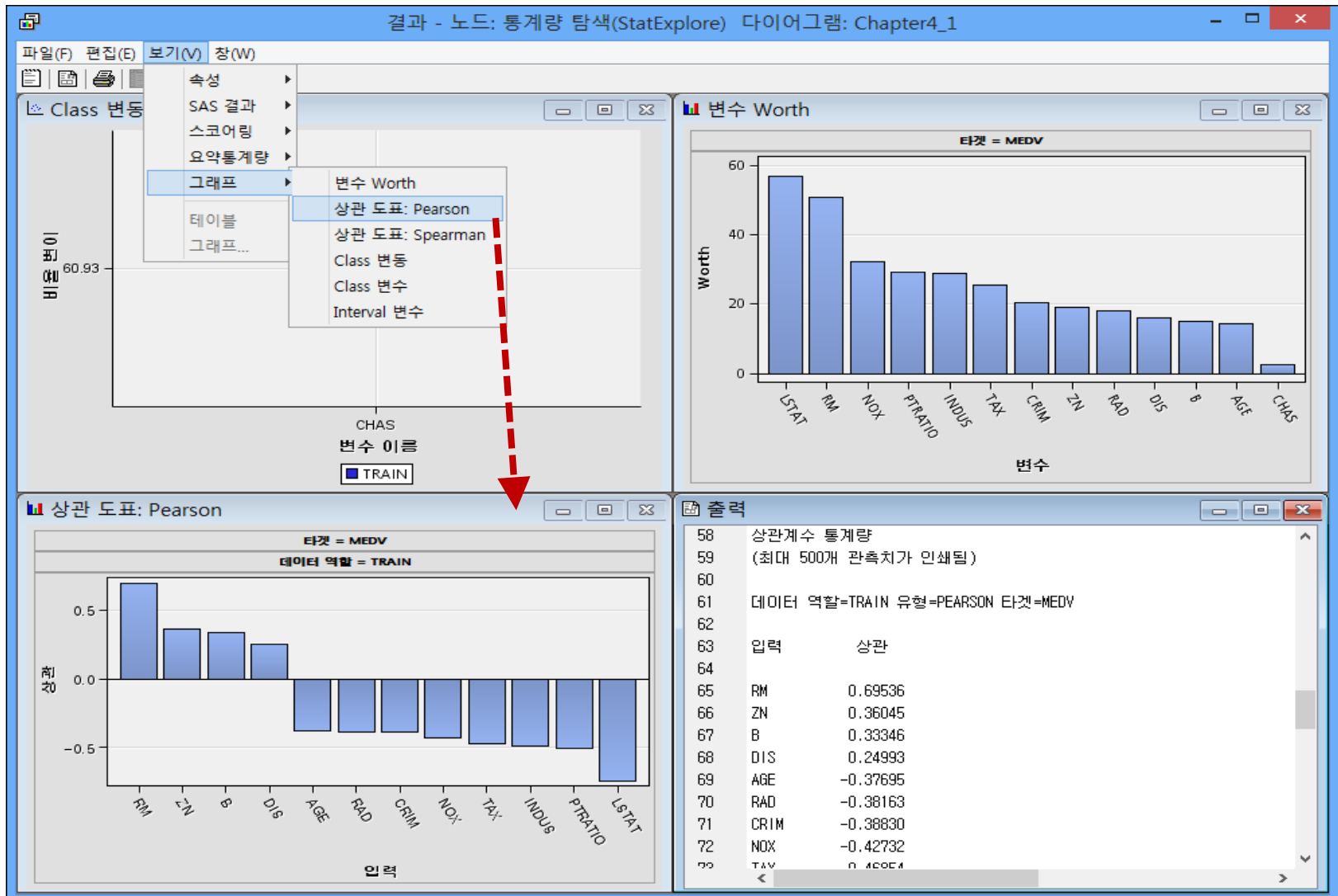
- 데이터 소스: HOUSING

- ✓ 변수 MEDV의 역할 칼럼을 Target으로 지정한다.
- ✓ 변수 CHAS의 레벨 칼럼을 Binary로 지정하고, 나머지 변수들의 레벨 칼럼은 Interval로 지정한다.

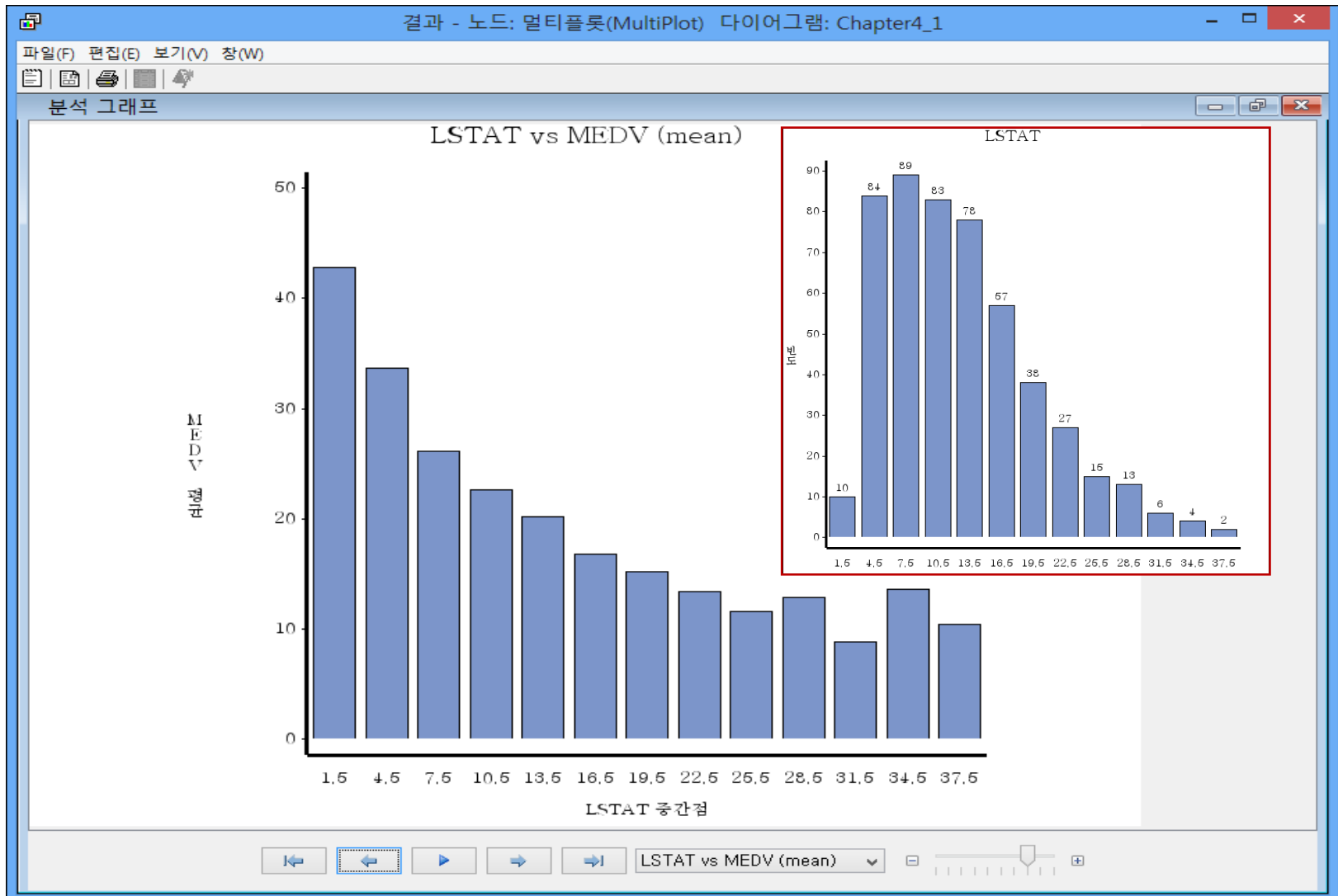
4.4.2 변수들의 분포에 대한 탐색



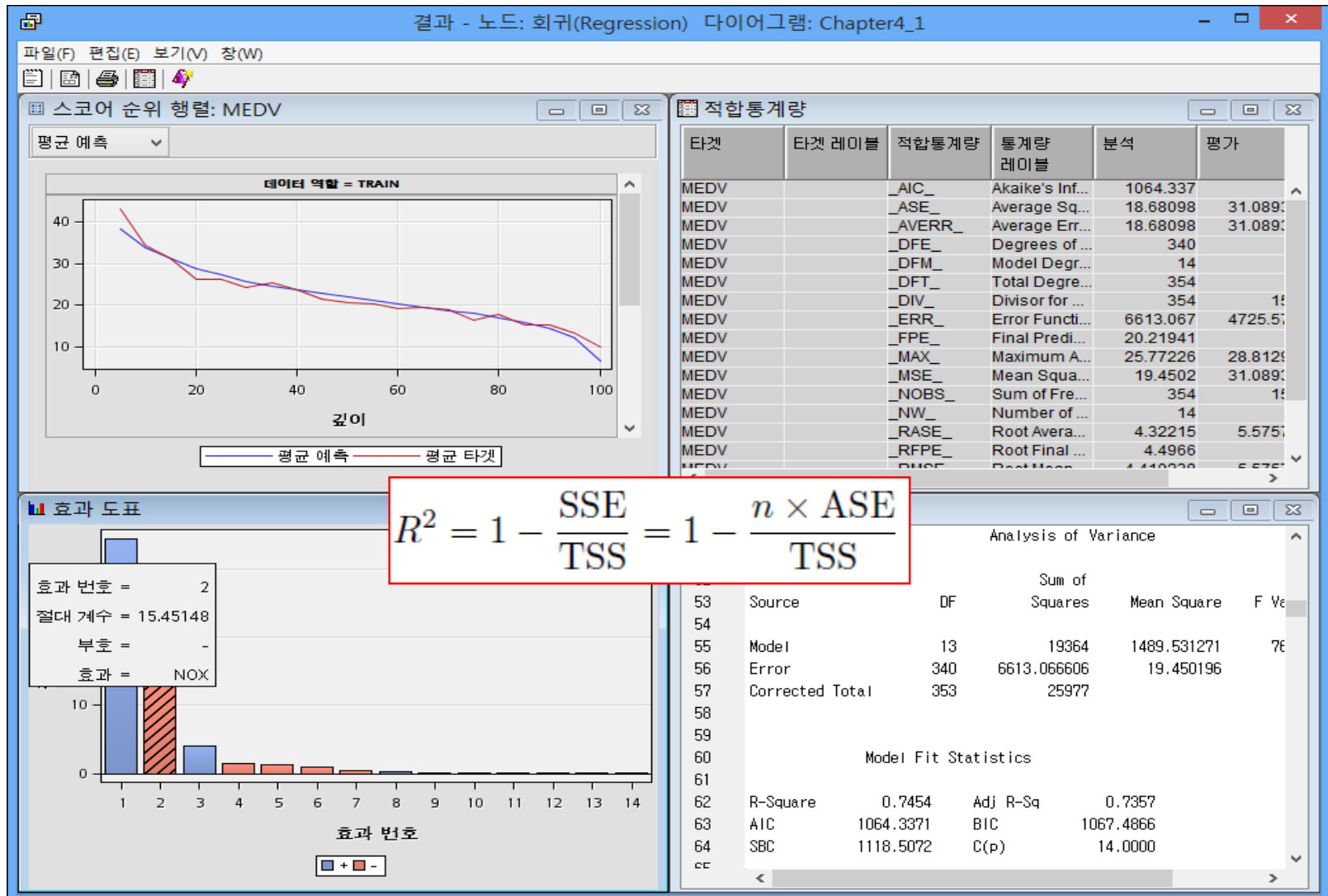
통계량 탐색(StatExplore) 노드 - 결과



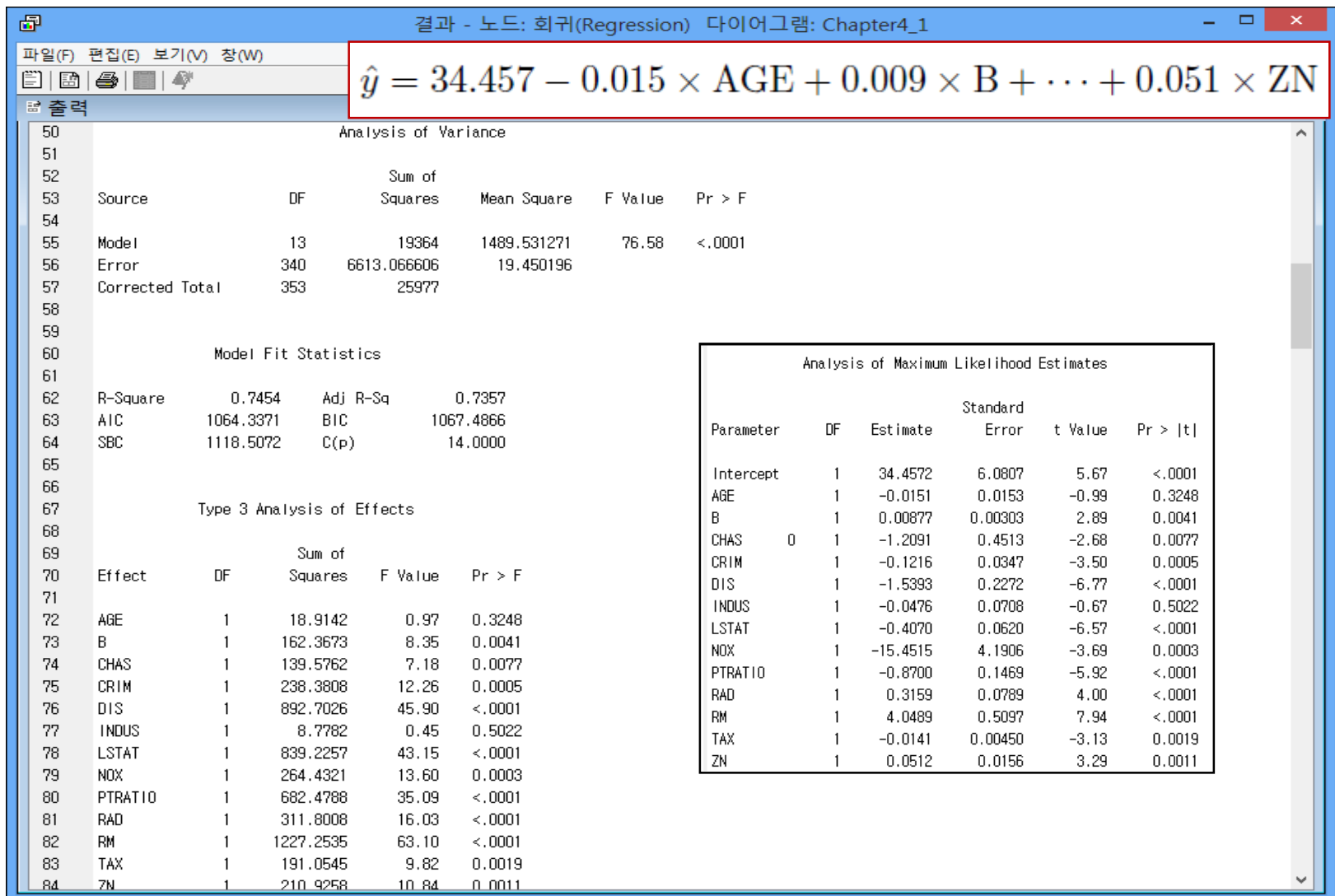
멀티 플롯(Multi Plot) 노드 - 결과



4.4.3 회귀(Regression) 노드의 실행과 결과 보기



회귀(Regression) 노드 - 결과: 출력 윈도우



차례

- 4.1 선형 회귀분석(Linear Regression Analysis)
- 4.2 로지스틱 회귀분석(Logistic Regression Analysis)
- 4.3 회귀분석의 특징과 제약
- 4.4 분석사례 - 1: 선형 회귀분석
- 4.5 분석사례 - 2: 로지스틱 회귀분석
- 4.6 분석사례 - 3: 신용평점표 작성
- 4.7 연습문제

분석사례 - 2를 위한 다이어그램

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

표본추출 탐색 수정 모델 평가 유틸리티 응용 프로그램 시계열

Chapter4_2

BUYTEST → 예측값 처리(Impute) → 데이터 분할(Data...) → 회귀(Regression) → 회귀-변수선택 → 모델비교(Model Comparison)

회귀-다항

이름 바꾸기

노드 이름:

회귀-다항

확인(O) 취소(C)

변수 편집...

업데이트

실행

모델 패키지 생성...

결과...

경로를 SAS 프로그램으로 내보내기

자르기

복사(C)

삭제

이름 바꾸기

모두 선택

노드 선택

노드 연결

노드 연결 해제

다이어그램 | 로그

다이어그램 Chapter4_2 열림

회귀 노드의 속성 패널

속성

일반

노드 ID: Reg2

가져온 데이터

내보낸 데이터

노트

분석

변수

방정식(Equation)

주효과(Main Effects): 예

2요인 교호작용(Two-Fa): 예

다항식 항(Polynomial T): 예

다항식 차수(Polynomial2): 예

사용자 항(User Terms): 아니요

항 편집기(Term Editor)

Class 타겟(Class Targ): 예

회귀 유형(Regression T로지스틱 회귀)

연결함수(Link Function): 로짓(Logit)

모델 옵션(Model Option)

절편 생략(Suppress Int): 아니요

입력 코딩(Input Coding): GLM

일반

일반 속성

4.5.1 변수선택 방법의 적용

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

DM Project

데이터 소스

다이어그램

Chapter2

Chapter3

Chapter4_2

표본추출 탐색 수정 모델 평가 유틸리티 응용 프로그램 시계열

BUYTEST

입력 처리(Impute)

데이터 분할(Data...)

선택 옵션(Selection Options)

속성	값
순차(Sequential Order)	아니요
변수 추가 기준 유의수준(Entry Significance)	0.2
변수 제거 기준 유의수준(Stay Significance)	0.1
시작 변수 개수(Start Variable Number)	0
종단할 변수 개수(Stop Variable Number)	0
후보 효과 추가(Force Candidate Effects)	0
계층효과(Hierarchy Effects)	CLASS
효과규칙 이동(Moving Effect Rule)	None
최대 단계 수(Maximum Number of Steps)	100

최대 단계 수(Maximum Number of Steps)

단계별 선택법에 대한 최대 단계 수입니다.

확인(O) 취소(C)

다이어그램

실행 완료

hckang(으)로서의 hckang hckang-pc에 연결

변수선택 과정의 요약

결과 - 노트: 회귀-변수선택 다이어그램: Chapter4_2																	
파일(F) 편집(E) 보기(V) 창(W)																	
출력																	
Summary of Stepwise Selection																	
	Effect		Number		Score	Wald											
	Step	Entered	Removed	DF	In	Chi-Square	Chi-Square	Pr > ChiSq									
	1	BUY18		1	1	36.3008		<.0001									
	2	BUY12		1	2	29.9847		<.0001									
	3	IMP_AGE		1	3	17.7058		<.0001									
	4	IMP_MARRIED		1	4	22.3246		<.0001									
	5	LOC		7	5	26.8528		0.0004									
	6	IMP_OWNSHOME		1	6	8.4599		0.0036									
	7	COA6		1	7	4.1983		0.0405									
	8	RETURN24		1	8	4.1920		0.0406									
	9	IMP_FICO		1	9	2.4281		0.1192									
	10		IMP_FICO	1	8		2.4261	0.1193									
The selected model is the model trained in the last step (Step 10). It consists of the																	
Intercept BUY12 BUY18 COA6 IMP_AGE IMP_MARRIED IMP_OWNSHOME LOC RETURN24																	
Likelihood Ratio Test for Global Null Hypothesis: BETA=0																	
-2 Log Likelihood		Likelihood															
Intercept	Intercept &	Ratio															
Only	Covariates	Chi-Square	DF	Pr > ChiSq													
2165.913	2022.295	143.6177	14	<.0001													
Type 3 Analysis of Effects																	
		Wald															
Effect	DF	Chi-Square	Pr > ChiSq														
BUY12	1	25.6923	<.0001														
BUY18	1	59.2229	<.0001														
COA6	1	4.4192	0.0355														
IMP_AGE	1	28.5753	<.0001														
IMP_MARRIED	1	20.4123	<.0001														
IMP_OWNSHOME	1	7.7806	0.0053														
LOC	5	9.3640	0.0954														
RETURN24	1	4.0723	0.0436														
Wald																	
Effect	DF	Chi-Square	Pr > ChiSq														

회귀계수 추정치

결과 - 노트: 회귀-변수선택 다이어그램: Chapter4_2								
파일(F) 편집(E) 보기(V) 창(W)								
출력								
Analysis of Maximum Likelihood Estimates								
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)	
Intercept	1	-0.9929	0.5398	3.38	0.0659		0.371	
BUY12	1	-0.9514	0.1877	25.69	<.0001	-0.2306	0.386	
BUY18	1	0.9927	0.1290	59.22	<.0001	0.3104	2.699	
COA6 0	1	-0.5959	0.2835	4.42	0.0355		0.551	
COA6 1	0	0	
IMP_AGE	1	-0.0381	0.00712	28.58	<.0001	-0.2080	0.963	
IMP_MARRIED 0	1	-0.6422	0.1421	20.41	<.0001		0.526	
IMP_MARRIED 1	0	0	
IMP_OWNSHOME 0	1	0.4028	0.1444	7.78	0.0053		1.496	
IMP_OWNSHOME 1	0	0	
LOC A	1	-0.3194	0.3298	0.94	0.3329		0.727	
LOC B	1	-0.1337	0.2248	0.35	0.5520		0.875	
LOC C	1	0.5395	0.2588	4.35	0.0371		1.715	
LOC D	1	0.4737	0.2709	3.06	0.0803		1.606	
LOC E	1	-0.2954	0.2249	1.73	0.1889		0.744	
LOC F	1	-0.2929	0.2276	1.66	0.1981		0.746	
LOC G	1	-0.3272	0.2743	1.42	0.2330		0.721	
LOC H	0	0	
RETURN24 0	1	0.6178	0.3062	4.07	0.0436		1.855	
RETURN24 1	0	0	
Odds Ratio Estimate								
Effect								
BUY12		0.386						
BUY18		2.699						
COA6 0 vs 1		0.551						

$$\begin{cases} f = -0.9929 - 0.9514 \times \text{BUY12} + \dots + 0.6178 \times \text{RETURN24}(0) \\ \hat{P}(y = 1) = \exp(f) / [1 + \exp(f)] \end{cases}$$

범주형 변수에 대한 코딩: 가변수(Dummy Variable)

결과 - 노드: 회귀-변수선택 다이어그램: Chapter4_2

파일(F) 편집(E) 보기(V) 창(W)

출력

Class Level Information

Class	Value	Design Variables								
CLIMATE	10	1	0	0						
	20	0	1	0						
	30	0	0	1						
COA6	0	1	0							
	1	0	1							
DISCBUY	0	1	0							
	1	0	1							
IMP_MARRIED	0	1	0							
	1	0	1							
IMP_ORGSRC	C	1	0	0	0	0	0	0	0	0
	D	0	1	0	0	0	0	0	0	0
	I	0	0	1	0	0	0	0	0	0
	O	0	0	0	1	0	0	0	0	0
	P	0	0	0	0	1	0	0	0	0
	R	0	0	0	0	0	1	0	0	0
	U	0	0	0	0	0	0	1	0	0
IMP_OWNSHOME	0	1	0							
	1	0	1							
IMP_SEX	F	1	0							
	M	0	1							
LOC	A	1	0	0	0	0	0	0	0	0
	B	0	1	0	0	0	0	0	0	0
	C	0	0	1	0	0	0	0	0	0
	D	0	0	0	1	0	0	0	0	0

CLIMATE	Deviation 방식			GLM 방식		
	D(10)	D(20)	D(30)	D(10)	D(20)	D(30)
10	1	0	0	1	0	0
20	0	1	0	0	1	0
30	-1	-1	0	0	0	0

4.5.2 교호작용과 이차항의 추가

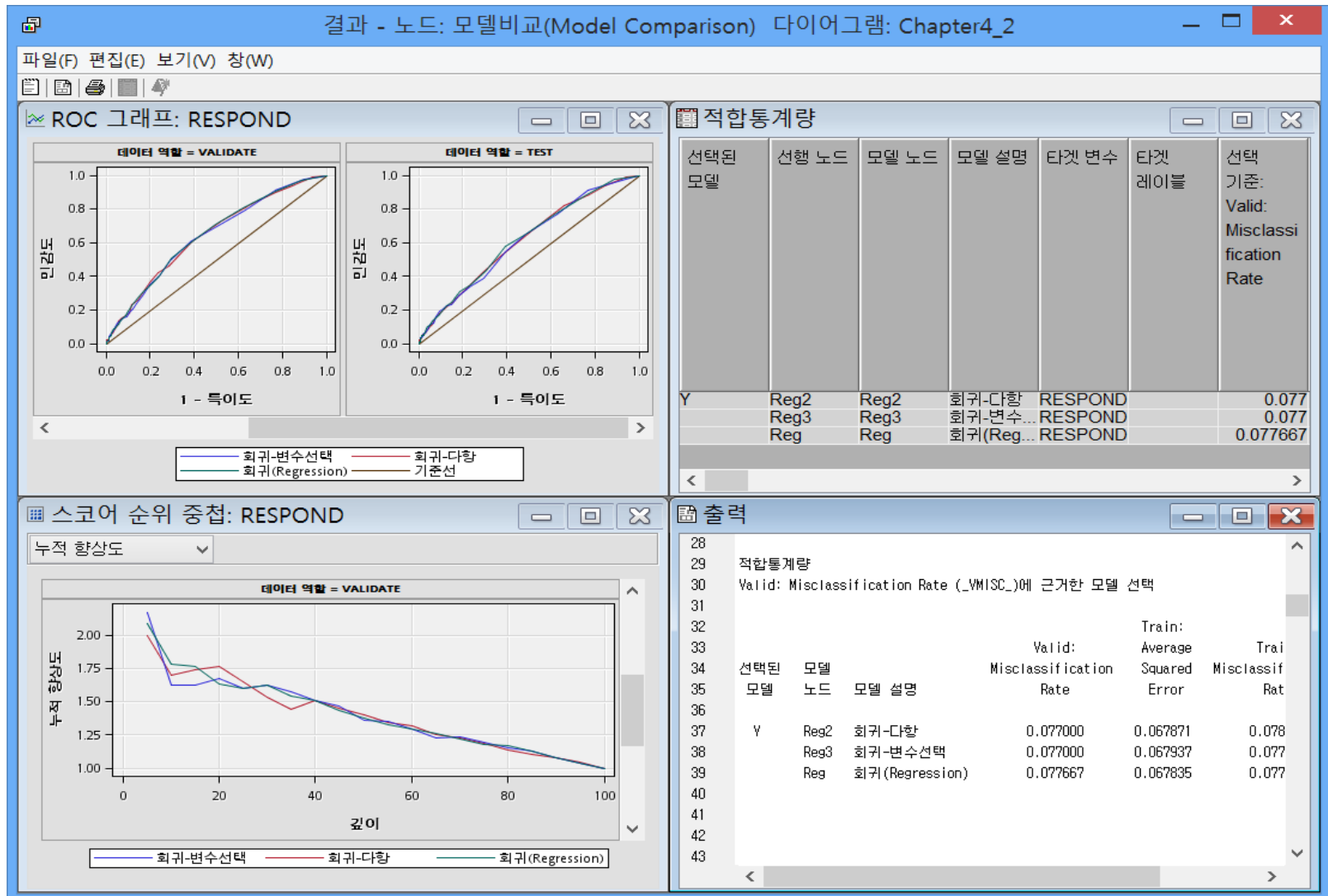
결과 - 노드: 회귀-다항 다이어그램: Chapter4_2

파일(F) 편집(E) 보기(V) 창(W)

출력

Summary of Stepwise Selection									
		Effect	Number		Score	Wald			
Step	Entered	Removed	DF	In	Chi-Square	Chi-Square	Pr >	ChiSq	
1	BUY18		1	1	36.3008		<.0001		
2	BUY12*IMP_AGE		1	2	34.8129		<.0001		
3	<=COA6*LOC		15	5	58.5790		<.0001		
4		COA6*LOC	7	4		4.7332	0.6925		
5	<=IMP_MARRIED*IMP_OWNHOME		3	7	19.9588		0.0002		
6	IMP_AGE*IMP_FICO		1	8	24.4259		<.0001		
7	RETURN24		1	9	4.5427		0.0331		
8	<=CLIMATE*COA6		2	11	5.2120		0.0738		
9		CLIMATE*COA6	2	10		4.5516	0.1027		
10	CLIMATE*COA6		2	11	5.2120		0.0738		
11		CLIMATE*COA6	2	10		4.5516	0.1027		
The selected model is the model trained in the last step (Step 11). It consists of the following effects:									
Intercept BUY18 CLIMATE COA6 IMP_MARRIED IMP_OWNHOME LOC RETURN24 IMP_MARRIED*IMP_OWNHOME									
Likelihood Ratio Test for Global Null Hypothesis: BETA=0									
-2 Log Likelihood		Likelihood							
Intercept	Intercept &	Ratio							
Only	Covariates	Chi-Square	DF	Pr > ChiSq					
2165.913	2012.413	153.5003	15	<.0001					
Type 3 Analysis of Effects									
				Type 3 Analysis of Effects					
Effect	DF	Wald	Chi-Square	Pr > ChiSq					
BUY18	1	59.3494	<.0001						
CLIMATE	0	.	.						
COA6	1	4.6536	0.0310						
IMP_MARRIED	1	25.2704	<.0001						
IMP_OWNHOME	1	11.9344	0.0006						
LOC	5	9.3147	0.0972						
RETURN24	1	4.4041	0.0359						
IMP_MARRIED*IMP_OWNHOME	1	6.2159	0.0127						
BUY12*IMP_AGE	1	25.1964	<.0001						
IMP_AGE*IMP_FICO	1	23.7983	<.0001						

4.5.3 모형 평가



4.5.4 예측확률 계산

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

표본추출 탐색 수정 모델 평가 유틸리티 응용 프로그램 시계열

Chapter4_2

BUYTEST → 결측값 처리(Impute) → 데이터 분할(Data...) → BUYROLL → 스코어(Score)

그래프 탐색(Graph...) 통계량 탐색(...) 멀티플롯(Multi Plot)

회귀(Regression) 회귀-변수선택 회귀-다항

모델비교(Model Comparison)

속성

분석 변수

평가 리포트

범주 수(Number of E20)

ROC 그래프(ROC Curve)

재계산(Recompute) 아니요

모델 선택

데이터 선택 기본

통계량 선택(Select) 기본

그리드 선택 통계량(기본)

선택 테이블(Select) 분석용(Training)

깊이 선택(Selection) 10

스코어

선택 편집기(Select)

리포트

선택한 모델

타겟 RESPOND

모델 노드 Reg2

모델 설명 회귀-다항

선택 기준 Valid: Misclassification

일반

일반 속성

실행 완료

선택 편집기(Selection Editor)-WORK.EMOUTFIT

선택된 모델	선택된 노드	모델 노드	모델 설명	타겟 변수	타겟 레이블	선택 기준: Valid: Misclassification Rate
아니요	Reg2	Reg2	회귀-다항	RESPOND		0.077
예	Reg3	Reg3	회귀-변수선택	RESPOND		0.077
아니요	Reg	Reg	회귀(Regression)	RESPOND		0.0776666667

확인(O) 취소(C)

스코어(Score) 노드 - 속성 패널

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

표준추출 탐색 수정 모델 평가 유틸리티 응용 프로그램 시계열

Chapter4_2

BUYTEST → 결측값 처리(Impute) → 데이터 분할(Data...) → BUYROLL → 스코어(Score)

그래프 탐색(Graph...) 통계량 탐색(...) 멀티플롯(Multi Plot) 회귀(Regression) 회귀-변수선택 회귀-다항 모델비교(Model Comparison)

스코어 노드의 속성 패널

속성

일반

노드 ID Score

가져온 데이터

내보낸 데이터

분석

변수

스코어 데이터 유형

고정된 출력 이름 사예

변수 숨기기(Hide Variable)

숨길 대상 선택(Hide Selection)

데이터 스코어(Score)

평가용(Validation) 아니요

검증용(Test) 아니요

스코어 코드 생성

최적화된 코드(Optimize)

C 스코어 코드(C Score)

Java 스코어 코드(Java)

Java 패키지 이름(Java Package Name)

사용자 패키지 이름(User Package Name)

일반 속성

실행 완료

내보낸 데이터 - 스코어(Score)

포트	테이블	역할	데이터 존재함
TRAIN	EMWS6.Score_TRAIN	Train	예
VALIDATE	EMWS6.Score_VALIDATE...	Validate	예
TEST	EMWS6.Score_TEST	Test	예
SCORE	EMWS6.Score_SCORE	Score	예

찾아보기(B)... 탐색(X)... 속성(P)... 확인(O)

스코어(Score) 노드 - 탐색

탐색 - EMWS6.Score_SCORE

파일(F) 보기(V) 작업(A) 창(W)

표본 속성

속성	값
행	알 수 없음
칼럼	34
라이브러리	EMWS6
멤버	SCORE_SCORE
유형	VIEW
표본추출 방법	Top
데이터 크기	기본
가져온 행	2000
난수초기값	12345

적용(L) 그래프(O)...

표본 통계량

관측치 ...	변수 이름	레이블	유형	결측률	최소
1	WARN	Warnings			
2	CLIMATE		CLASS	0	
3	EM_CLA...	Predictio...	CLASS	0	
4	ID		CLASS	0	
5	IMP_OR...	Imputed ...	CLASS	0	
6	IMP_SEX	Imputed ...	CLASS	0	
7	L_RESP...	Into: RES...	CLASS	0	
8	LOC		CLASS	0	
9	ORGSRC		CLASS	5.3	
10	SEX		CLASS	2.95	
11	AGE		VAR	2.95	1
12	BUY12		VAR	0	
13	BUY18		VAR	0	
14	BUY6		VAR	0	

EMWS6.Score_SCORE

Warnin...	Into: R...	Unnor...	Predict...	Predict...	b_RES...	세그먼트	Probability for level 1 of RESPOND	Probab...	Predict
0		0	0.097184	0.902816	5	5	0.097184	0.9028160	
0		0	0.026117	0.973883	19	19	0.026117	0.9738830	
0		0	0.049712	0.950288	14	14	0.049712	0.9502880	
0		0	0.059958	0.940042	11	11	0.059958	0.9400420	
0		0	0.055893	0.944107	12	12	0.055893	0.9441070	
0		0	0.060857	0.939143	11	11	0.060857	0.9391430	
0		0	0.071099	0.928901	8	8	0.071099	0.9289010	
0		0	0.189198	0.810802	1	1	0.189198	0.8108020	
0		0	0.056068	0.943932	12	12	0.056068	0.9439320	
0		0	0.06242	0.93758	10	10	0.06242	0.937580	
0		0	0.04158	0.95842	16	16	0.04158	0.958420	
0		0	0.038711	0.961289	17	17	0.038711	0.9612890	
0		0	0.071967	0.928033	8	8	0.071967	0.9280330	
0		0	0.025049	0.974951	19	19	0.025049	0.9749510	

차례

- 4.1 선형 회귀분석(Linear Regression Analysis)
- 4.2 로지스틱 회귀분석(Logistic Regression Analysis)
- 4.3 회귀분석의 특징과 제약
- 4.4 분석사례 - 1: 선형 회귀분석
- 4.5 분석사례 - 2: 로지스틱 회귀분석
- 4.6 분석사례 - 3: 신용평점표 작성
- 4.7 연습문제

분석사례-3을 위한 분석흐름도

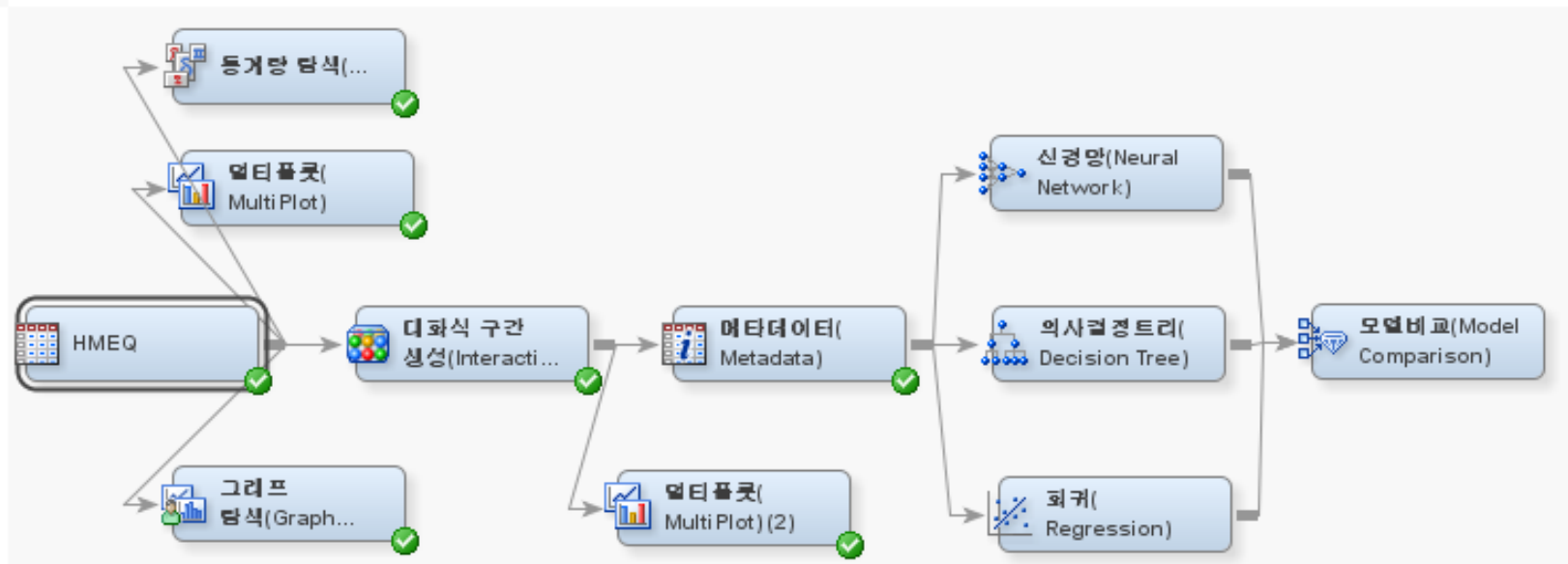
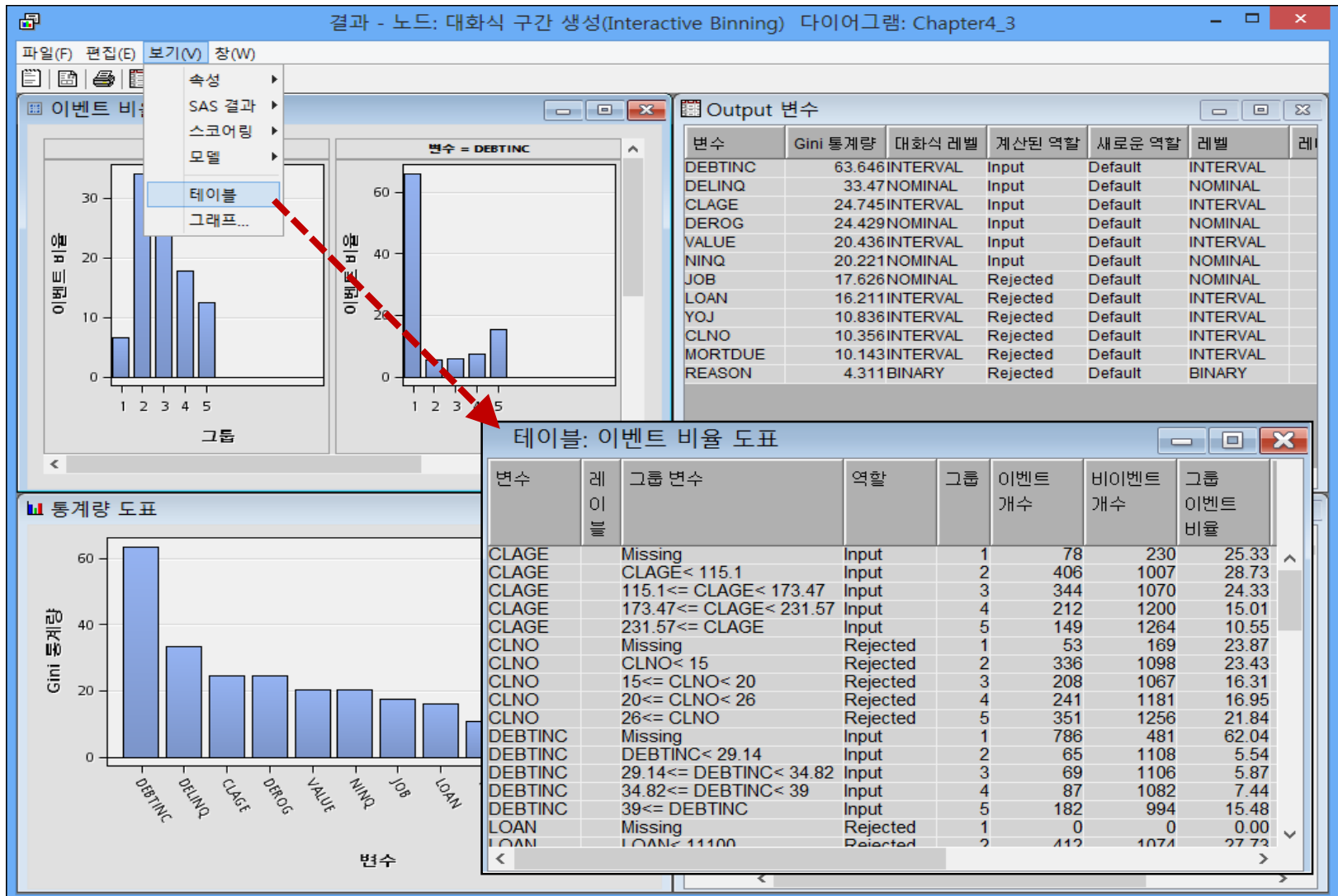


표 4.5 대화식 범주화의 예

그룹	DEBTINC	DELINQ	CLAGE	VALUE	DEROG
1	Missing	Missing, 0	Missing, 150 미만	Missing	Missing
2	44 미만	1	150~240	47500 미만	0
3	44 이상	2 이상	240 이상	47500~200000	1 이상
4				200000 이상	

4.6.1 대화식 구간화: Interactive Binning 노트



대화식 구간 생성(Interactive Binning) 노드 - 속성 패널

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

DM Project

- 데이터 소스
- 다이어그램
 - Chapter2
 - Chapter3_1
 - Chapter3_2
 - Chapter4_1
 - Chapter4_3

대화식 구간 생성 노드의 속성 패널

속성

문식

변수

대화식 범주화(Interactive Binning)

결과값 범주화 여부(Tre에)

기존 그룹화 사용(Use Old)

Interval 변수 옵션

레벨 규칙 적용(Apply)

방법(Method) Quantile

그룹 개수 4

Class 변수 옵션

회귀 레벨 그룹화(Grouping)

임계치 비율(Cutoff Value) 0.5

가져오기/내보내기 옵션

그룹화 데이터 가져오기(Import)

데이터셋 가져오기

스코어

그룹 레벨(Group Level) Nominal

변수 선택 방법(Variable Selection Method) Gini 통계량

Gini 임계치(Gini Cutoff) 0.0

리포트

그룹화 데이터 생성(Create Metadata)

새로운 방법(Create Method) 사용

일반

일반 속성

실행 완료

Chapter4_3

대화식 범주화(Interactive Binning)

변수 선택

선택한 변수 DEBTINC

변수 그룹화

이벤트 개수

임계치

그룹

계산된 WOE 새로운 WOE

값	그룹	임계치	이벤트 수	비이벤트 개수	합계	이벤트 비율
MISSING	1		786.0	481.0	1267.0	0.62
DEBTINC ...	2	29.14	65.0	1108.0	1173.0	0.055
29.14 <= D...	3	34.82	69.0	1106.0	1175.0	0.059
34.82 <= D...	4	39.0	87.0	1082.0	1169.0	0.074
DEBTINC ...	5		182.0	994.0	1176.0	0.155

변수 통계량

원본 Gini 63,646

새로운 Gini 63,646

상세 레벨

☒ 상세

☐ 구간화

모든 변경 사항 재설정

닫기

대화식 범주화: DEBTINC(구간형 변수)의 경우

대화식 범주화(Interactive Binning)

변수 선택
선택한 변수: DEBTINC

이전 ← 다음 →

변수 그룹화

이벤트 개수

계수

그룹

계산된 WOE 새로운 WOE

범주 분할

기존 임계치: N/A
새로운 임계치 입력:
44

확인 취소

값	그룹	임계치	이벤트 수	비이벤트 개수	합계	이벤트 비율
MISSING	1		786,0	481,0	1267,0	0,62
DEBTINC ...	2	29,14	65,0	1108,0	1173,0	0,055
29,14 <= D...	3	34,82	69,0	1106,0	1175,0	0,059
34,82 <= D...	4	39,0	87,0	1082,0	1169,0	0,074
39 <= DEB...	5	44,0	94,0	979,0	1073,0	0,088
44 <= DEB...	5		88,0	15,0	103,0	0,854

변수 통계량

원본 Gini 63,646
새로운 Gini 63,646

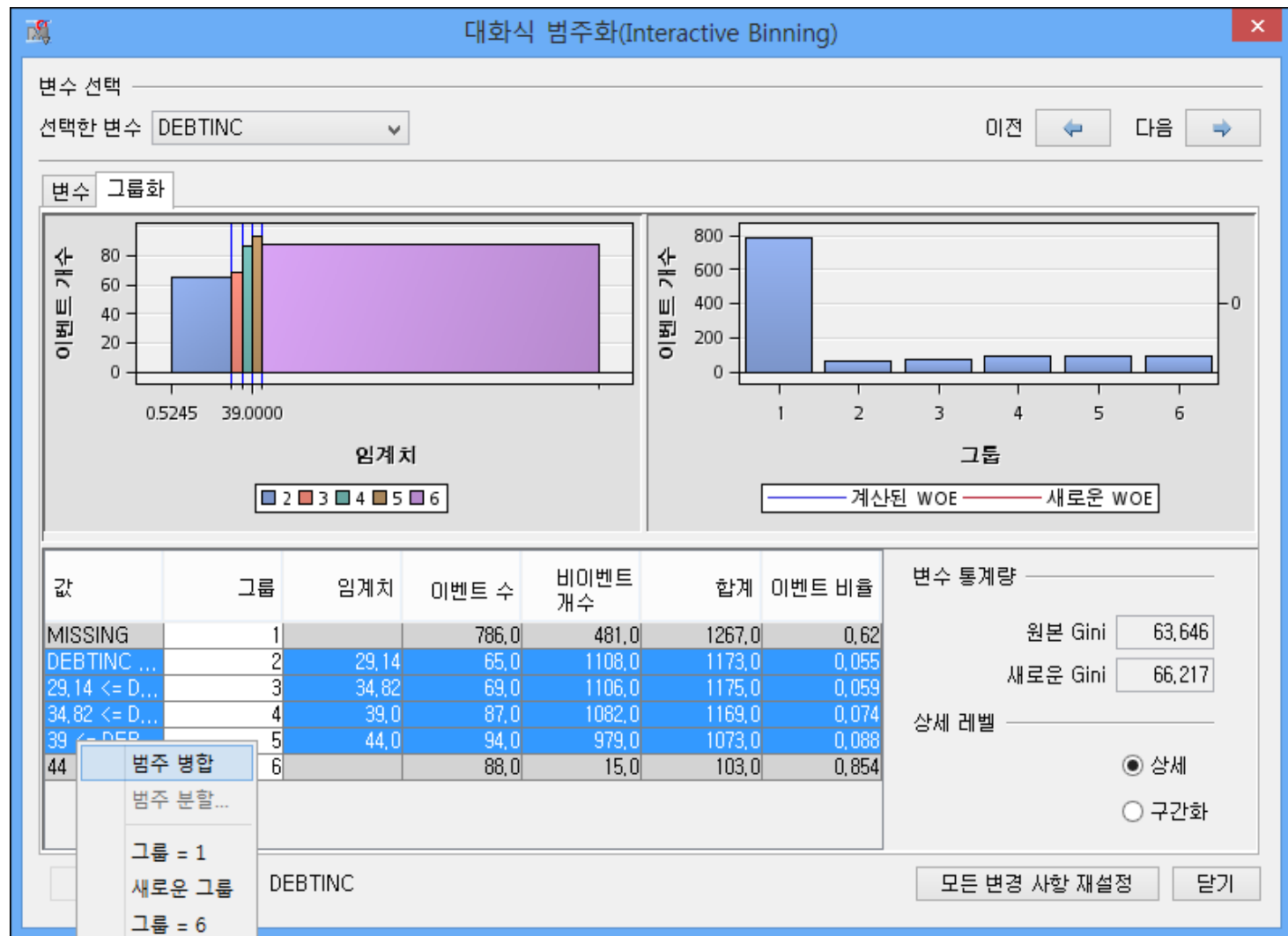
상세 레벨

☒ 상세
☐ 구간화

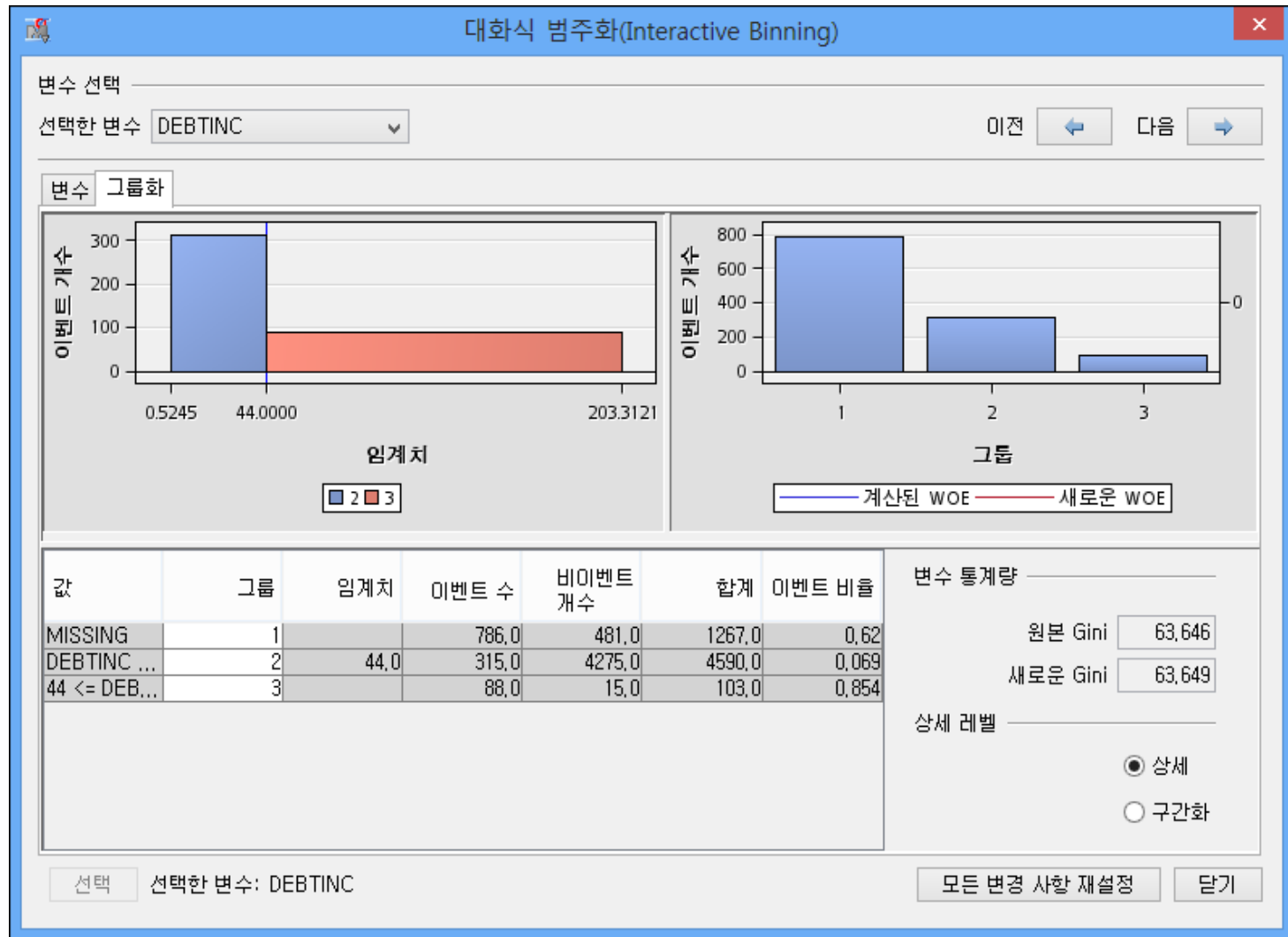
모든 변경 사항 재설정 닫기

범주 병합
범주 분할...
그룹 = 4
새로운 그룹
그룹 = 6

대화식 범주화 - 범주 병합



변수 DEBTINC에 대한 범주화



대화식 범주화: DELINQ(범주형 변수)의 경우

대화식 범주화(Interactive Binning)

변수 선택

선택한 변수

DELINQ

이전

다음

변수

그룹화

이벤트 개수

이벤트 개수

값	그룹	임계치	이벤트 수	비이벤트 개수	합계	이벤트 비율
MISSING	1		72,0	508,0	580,0	0,124
0	1		583,0	3596,0	4179,0	0,14
1	2		222,0	432,0	654,0	0,339
2	3		112,0	138,0	250,0	0,448
3	3		71,0	58,0	129,0	0,55
4	3					0,59
5	3					0,816
6	3					1,0
7	3					1,0
8	3					1,0
10	3					1,0
11	3					1,0
12	3					1,0
13	3					1,0
15	3					1,0

변수 통계량

원본 Gini

33,470

새로운 Gini

32,388

상세 레벨

☒ 상세

☐ 구간화

그룹 선택

?

그룹 선택

3

1

2

3

할당...

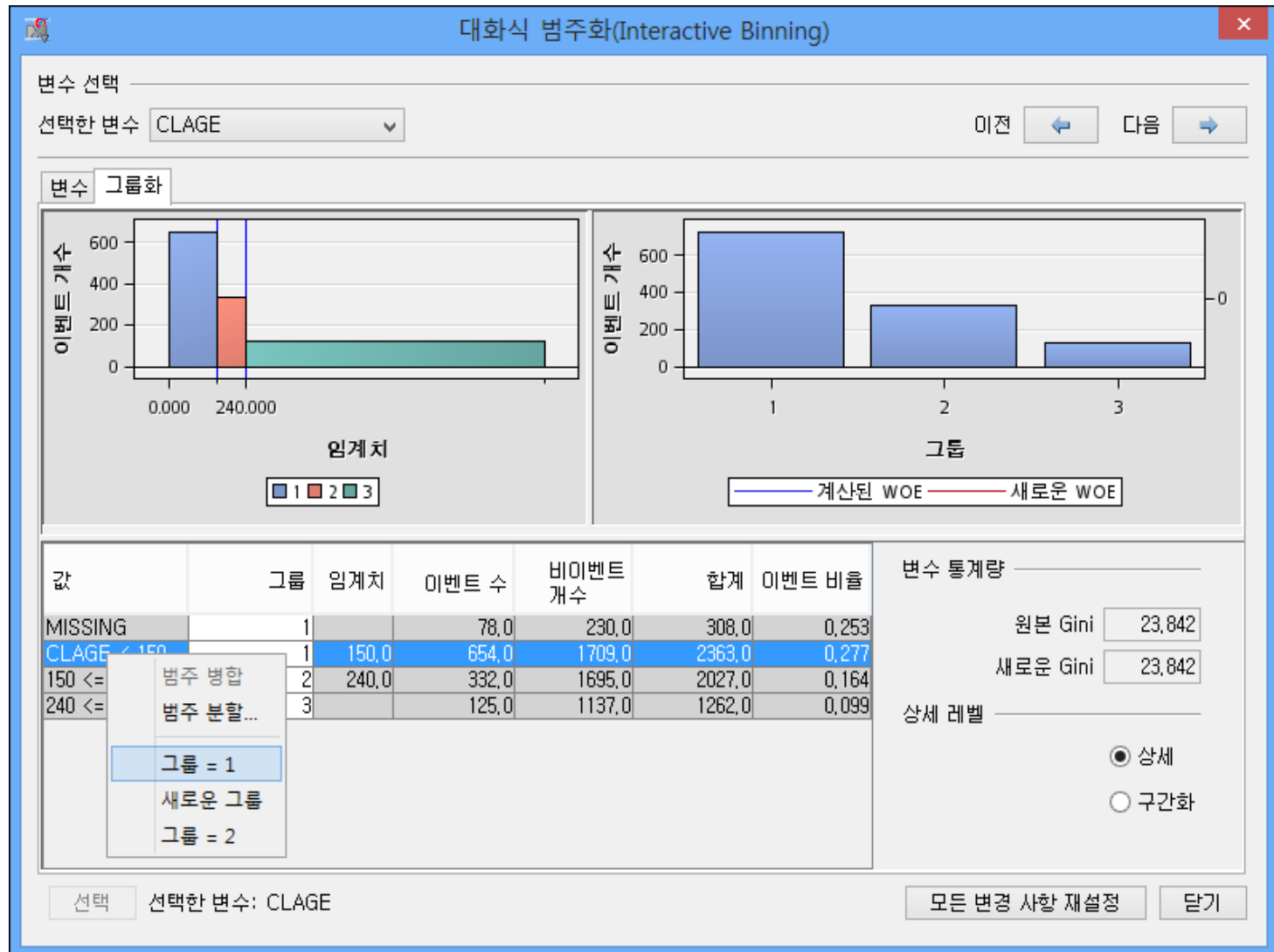
선택

선택한 변수: DELINQ

모든 변경 사항 재설정

닫기

대화식 범주화: 변수 CLAGE의 경우



4.6.2 변수들의 설정 변경: Metadata 노트

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

표본추출 탐색 수정 모델 평가 유틸리티 응용 프로그램 시계열

DM Project

- 데이터 소스
- 다이어그램
 - Chapter2
 - Chapter3.1
 - Chapter3.2
 - Chapter4.1

메타데이터 노트의 속성 패널

속성

일반

노드 ID Meta

가져온 데이터

내보낸 데이터

노드

분석

가져올 항목 선택

요약 아니요

고급 관리자 아니요

기각된 변수(Reject)

기각된 변수 숨기 아니요

조합 규칙(Comb)없음

변수

분석(Train)

트랜잭션(Transaction)

평가(Validate)

검증(Test)

스코어(Score)

상태

생성 시간 14, 2, 14 오후 3:2

실행 ID 16b7fbc8-ad5d-4

최근 오픈

일반

일반 속성

실행 완료

Chapter4.3

변수 - Meta

(none) ☐ not Equal to

적용 재설정

칼럼: ☐ 레이블(A) ☐ 마이닝(M) ☐ 기본(I) ☐ 통계량(T)

이름	참치기	숨기기	역할	새로운 역할	레벨	새로운 레벨	새로운 순서
BAD		기본	Target	Target	BINARY	Default	오름차순
CLAGE		기본	Rejected	Rejected	Interval	Default	기본
CLNO		기본	Rejected	Rejected	Interval	Default	내림차순
DEBTINC		기본	Rejected	Rejected	Interval	Default	서식이 설정된 내림차
DELINQ		기본	Rejected	Rejected	Nominal	Default	서식이 설정된 오름차
DEROG		기본	Rejected	Rejected	Nominal	Default	오름차순
GRP_CLAGE		기본	Input	Input	Nominal	Default	기본
GRP_CLNO		기본	Input	Rejected	Nominal	Default	기본
GRP_DEBTINC		기본	Input	Input	Nominal	Default	기본
GRP_DELINQ		기본	Input	Input	Nominal	Default	기본
GRP_DEROG		기본	Input	Input	Nominal	Default	기본
GRP_JOB		기본	Input	Rejected	Nominal	Default	기본
GRP_LOAN		기본	Input	Rejected	Nominal	Default	기본
GRP_MORTDUE		기본	Input	Rejected	Nominal	Default	기본
GRP_NINQ		기본	Input	Rejected	Nominal	Default	기본
GRP_REASON		기본	Input	Rejected	Nominal	Default	기본
GRP_VALUE		기본	Input	Input	Nominal	Default	기본
GRP_YOJ		기본	Input	Rejected	Nominal	Default	기본
JOB		기본	Rejected	Rejected	Nominal	Default	기본
LOAN		기본	Rejected	Rejected	Interval	Default	기본
MORTDUE		기본	Rejected	Rejected	Interval	Default	기본
NINQ		기본	Rejected	Rejected	Nominal	Default	기본
REASON		기본	Rejected	Rejected	BINARY	Default	기본
VALUE		기본	Rejected	Rejected	Interval	Default	기본
YOJ		기본	Rejected	Rejected	Interval	Default	기본

다이어그램

탐색(X)... 경로 업데이트(U) 확인(O) 취소(C)

4.6.3 로지스틱 회귀분석을 이용한 계수 추정

Type 3 Analysis of Effects

Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
GRP_CLAGE	2	134.3582	<.0001
GRP_DEBTINC	2	1078.8534	<.0001
GRP_DELINQ	2	237.7478	<.0001
GRP_DEROG	2	117.8789	<.0001
GRP_VALUE	3	122.0245	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard		Wald		Exp (Est)
			Error	Chi-Square	Pr>ChiSq		
Intercept	1	-4.0086	0.3991	100.89	<.0001	55.070	
GRP_CLAGE	1 1	-1.3861	0.1389	99.61	<.0001	3.999	
GRP_CLAGE	2 1	-0.4884	0.1470	11.04	0.0009	1.630	
GRP_CLAGE	3 0	0	
GRP_DEBTINC	1 1	1.5979	0.3060	27.27	<.0001	0.202	
GRP_DEBTINC	2 1	4.5202	0.3067	217.16	<.0001	0.011	
GRP_DEBTINC	3 0	0	
GRP_DELINQ	1 1	1.9949	0.1379	209.27	<.0001	0.136	
GRP_DELINQ	2 1	0.9747	0.1688	33.36	<.0001	0.377	
GRP_DELINQ	3 0	0	
GRP_DEROG	1 1	2.0150	0.1920	110.16	<.0001	0.133	
GRP_DEROG	2 1	0.9343	0.1213	59.31	<.0001	0.393	
GRP_DEROG	3 0	0	
GRP_VALUE	1 1	-3.5041	0.4898	51.19	<.0001	33.253	
GRP_VALUE	2 1	-0.2164	0.2247	0.93	0.3354	1.242	
GRP_VALUE	3 1	0.6390	0.1921	11.07	0.0009	0.528	
GRP_VALUE	4 0	0	

4.6.4 평점표 작성

- 사후확률 추정

$$\begin{cases} f = -4.0086 + 1.5979 \times \text{GRP_DEBTINC}(-1) + \dots + 0 \times \text{GRP_DEROG}(2) \\ \hat{P}(y = 0) = \exp(f) / [1 + \exp(f)] \end{cases}$$

- 회귀계수 추정치의 보정

$$\text{보정된 추정치} = (\text{회귀계수 추정치}) - (\text{가장 작은 회귀계수 추정치})$$

- POD를 이용한 변환

$$\text{평점} = \text{보정된 추정치} \times (\text{POD} / \log(2))$$

평점표 작성의 예

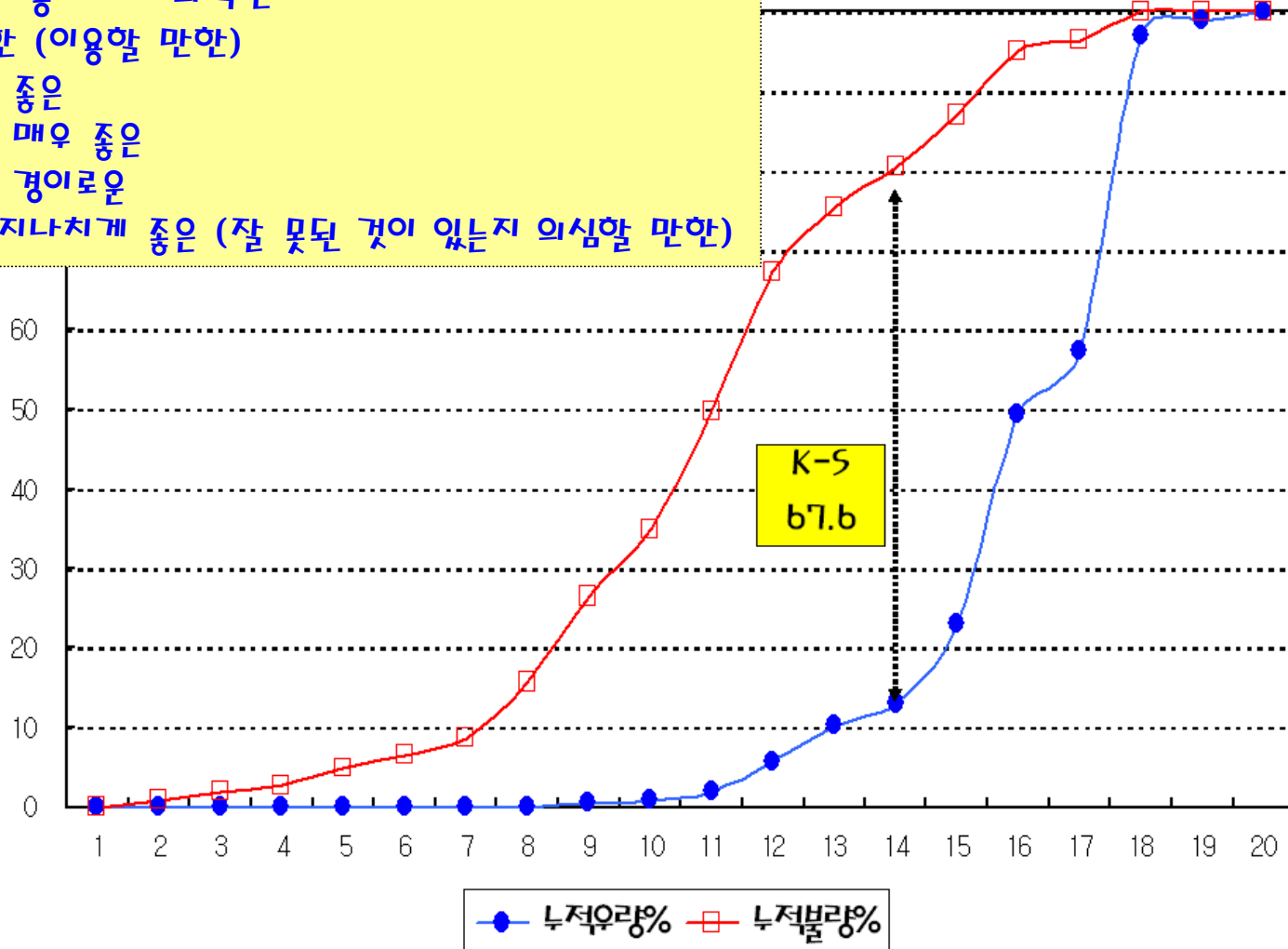
변수	범주	회귀 계수 추정치	보정된 추정치	평점	범위
DEBTINC 수입 대비 부채비율	Missing	1.5979	1.5979	115	326 (32.1%)
	44 미만	4.5202	4.5202	326	
	44 이상	0.0000	0.0000	0	
CLAGE 가장 오래된 거래의 개월 수	Missing, 150 미만	-1.3861	0.0000	0	100 (9.9%)
	150~240	-0.4884	0.8977	65	
	240 이상	0.0000	1.3861	100	
DELINQ 부실거래의 수	Missing, 0	1.9949	1.9949	144	144 (14.2%)
	1	0.9747	0.9747	70	
	2 이상	0.0000	0.0000	0	
VALUE 현재의 자산가치	Missing	-3.5041	0.0000	0	299 (29.5%)
	47500 미만	-0.2164	3.2877	237	
	47500~200000	0.6390	4.1431	299	
	200000 이상	0.0000	3.5041	253	
DEROG 주요 부실 거래의 수	Missing	2.0150	2.0150	145	145 (14.3%)
	0	0.9343	0.9343	67	
	1 이상	0.0000	0.0000	0	
				합계	1014

4.6.4 평점표의 타당성 평가

평점그룹	빈도			평점그룹		전체		누적		K-S
	우량	불량	전체	우량%	불량%	우량%	불량%	우량%	불량%	
0-50	0	0	0	-	-	0.0	0.0	0.0	0.0	0.0
50-100	0	1	1	0.0	100.0	0.0	0.1	0.0	0.1	0.1
100-150	0	8	8	0.0	100.0	0.0	0.7	0.0	0.8	0.8
150-200	0	14	14	0.0	100.0	0.0	1.2	0.0	2.0	2.0
200-250	0	8	8	0.0	100.0	0.0	0.7	0.0	2.7	2.7
250-300	0	26	26	0.0	100.0	0.0	2.2	0.0	4.9	4.9
300-350	0	18	18	0.0	100.0	0.0	1.5	0.0	6.4	6.4
350-400	0	26	26	0.0	100.0	0.0	2.2	0.0	8.6	8.6
400-450	4	86	90	4.4	95.6	0.1	7.2	0.1	15.8	15.7
450-500	19	126	145	13.1	86.9	0.4	10.6	0.5	26.4	25.9
500-550	14	101	115	12.2	87.8	0.3	8.5	0.8	34.9	34.1
550-600	55	176	231	23.8	76.2	1.2	14.8	2.0	49.7	47.7
600-650	173	208	381	45.4	54.6	3.6	17.5	5.6	67.2	61.6
650-700	221	97	318	69.5	30.5	4.6	8.2	10.2	75.4	65.2
700-750	128	61	189	67.7	32.3	2.7	5.1	12.9	80.5	67.6
750-800	487	79	566	86.0	14.0	10.2	6.6	23.1	87.1	64.0
800-850	1259	95	1354	93.0	7.0	26.4	8.0	49.5	95.1	45.6
850-900	366	16	382	95.8	4.2	7.7	1.3	57.2	96.4	39.2
900-950	1898	43	1941	97.8	2.2	39.8	3.6	97.0	100.0	3.0
950-1000	84	0	84	100.0	0.0	1.8	0.0	98.8	100.0	1.2
1000-	63	0	63	100.0	0.0	1.3	0.0	100.0	100.0	0.0
전체	4771	1189	5960	80.1	19.9	100.0	100.0	최대값(K-S)		67.6

K-S(Kolmogorov-Smirnov) 통계량

- 20 이하: 이용가치가 희박한
- 40: 적당한 (이용할 만한)
- 40 ~ 50: 좋은
- 50 ~ 60: 매우 좋은
- 60 ~ 75: 경이로운
- 75 이상: 지나치게 좋은 (잘 못된 것이 있는지 의심할 만한)



민감도와 특이도

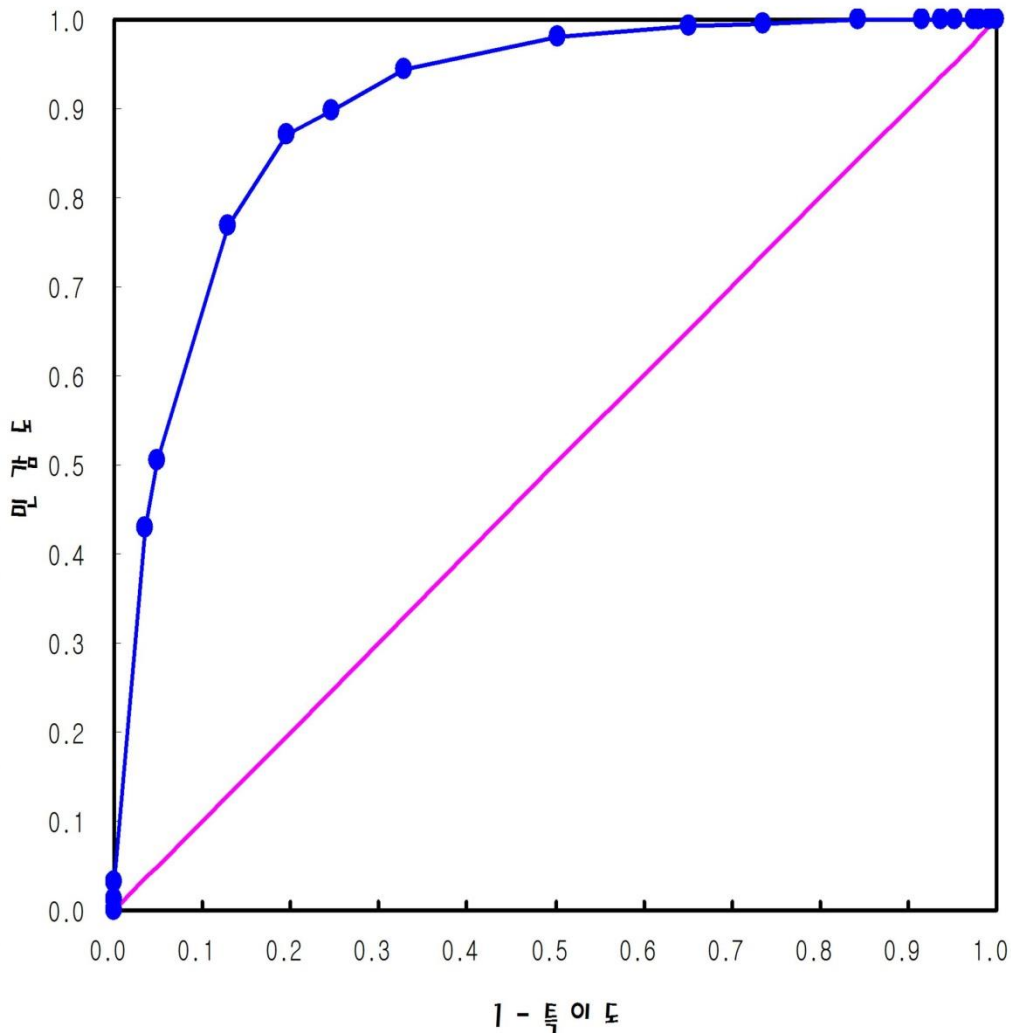
- n : 전체 개체수
 - n_1 : 우량으로 관측된 개체수
 - n_0 : 불량으로 관측된 개체수
 - n_{11} : 우량으로 관측된 n_1 개의 개체 중 우량으로 정분류된 개체수
 - n_{00} : 불량으로 관측된 n_0 개의 개체 중 불량으로 정분류된 개체수
-
- 정확도(Accuracy; 정분류율) = $(n_{11} + n_{00})/n$
 - 오분류율(Misclassification Rate) = $[(n_1 - n_{11}) + (n_0 - n_{00})]/n$
 - 민감도(Sensitivity) = n_{11}/n_1
 - 특이도(Specificity) = n_{00}/n_0

... 민감도와 특이도

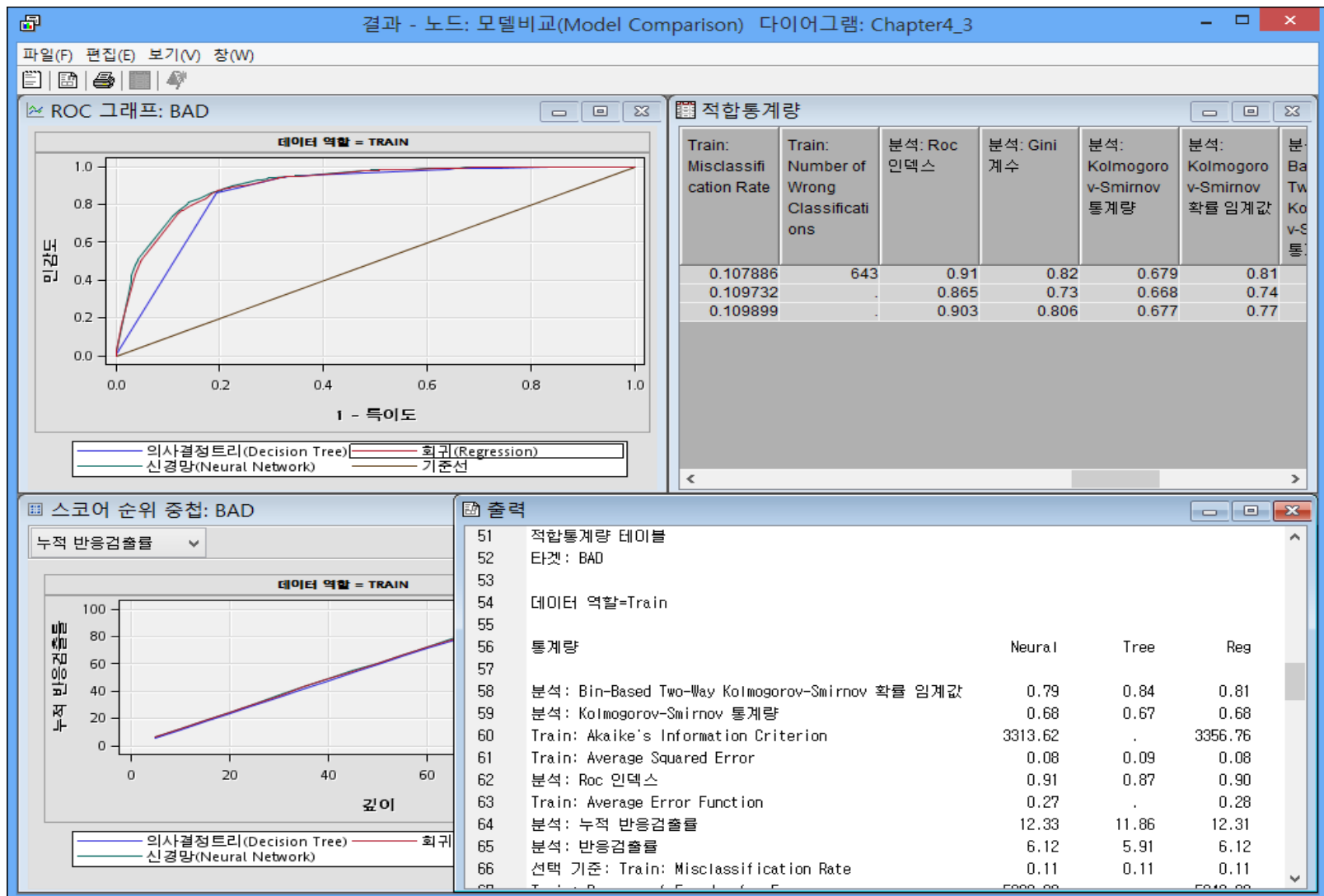
절단값	빈도		누적		정분류		예측 정확도		
	우량	불량	우량	불량	우량	불량	정확도	민감도	특이도
0	-	-	-	-	4771	0	0.801	1.000	0.000
50	0	1	0	1	4771	1	0.801	1.000	0.001
100	0	8	0	9	4771	9	0.801	1.000	0.008
150	0	14	0	23	4771	23	0.801	1.000	0.019
200	0	8	0	31	4771	31	0.801	1.000	0.026
250	0	26	0	57	4771	57	0.801	1.000	0.048
300	0	18	0	75	4771	75	0.801	1.000	0.063
350	0	26	0	101	4771	101	0.801	1.000	0.085
450	4	86	4	187	4767	187	0.800	0.999	0.157
500	19	126	23	313	4748	313	0.797	0.995	0.263
550	14	101	37	414	4734	414	0.794	0.992	0.348
600	55	176	92	590	4679	590	0.785	0.981	0.496
650	173	208	265	798	4506	798	0.756	0.944	0.671
700	221	97	486	895	4285	895	0.719	0.898	0.753
750	128	61	614	956	4157	956	0.697	0.871	0.804
800	487	79	1101	1035	3670	1035	0.616	0.769	0.870
850	1259	95	2360	1130	2411	1130	0.405	0.505	0.950
900	366	16	2726	1146	2045	1146	0.343	0.429	0.964
950	1898	43	4624	1189	147	1189	0.025	0.031	1.000
1000	84	0	4708	1189	63	1189	0.011	0.013	1.000
Max	63	0	4771	1189	0	1189	0.000	0.000	1.000

ROC(Receiver Operation Characteristic) 곡선

민감도	1-특이도	C
0.000	0.000	0.000
0.013	0.000	0.000
0.031	0.000	0.000
0.429	0.036	0.008
0.505	0.050	0.006
0.769	0.130	0.051
0.871	0.196	0.055
0.898	0.247	0.045
0.944	0.329	0.075
0.981	0.504	0.168
0.992	0.652	0.146
0.995	0.737	0.084
0.999	0.843	0.106
1.000	0.915	0.072
1.000	0.937	0.022
1.000	0.952	0.015
1.000	0.974	0.022
1.000	0.981	0.007
1.000	0.992	0.012
1.000	0.999	0.007
1.000	1.000	0.001
합계		0.902



모델 비교(Model Comparison) 노드 - 결과



결과 - 노드: 모델비교(Model Comparison) 다이어그램: Chapter4_3

파일(F) 편집(E) 보기(V) 창(W)

ROC 그래프

속성

- SAS 결과
- 스코어링
- 모델
- 평가
- 테이블
- 그래프...

데이터 역할 = TRAIN

1 - 특이도

민감도

의사결정트리(Decision Tree)

신경망(Neural Network)

회귀(Regression)

기준선

테이블: ROC 그래프: BAD

True Positive 개수	민감도	True Negative 개수	False Positive 개수	1 - 특이도	특이도	False Negative 개수	그룹 내 첫번째
4762	0.998114	281	908	0.763667	0.236333	9	0.1012
4763	0.998323	267	922	0.775442	0.224558	8	0.0956
4765	0.998742	254	935	0.786375	0.213625	6	0.0872
4765	0.998742	251	938	0.788898	0.211102	6	0.0783
4766	0.998952	238	951	0.799832	0.200168	5	0.0600
4767	0.999162	213	976	0.820858	0.179142	4	0.0594
4770	0.99979	177	1012	0.851135	0.148865	1	0.0428
4771	1	159	1030	0.866274	0.133726	0	0.0338
4771	1	155	1034	0.869638	0.130362	0	0.0208
4771	1	126	1063	0.894029	0.105971	0	0.0102
4771	1	111	1078	0.906644	0.093356	0	.00009
1	1	1	1	1	0	0	0
0	1	1	1	0	0	0	0
1	1	1	1	1	0	0	0

스코어 순위 중첩: BAD

누적 반응검출률

데이터 역할 = TRAIN

깊이

의사결정트리(Decision Tree)

신경망(Neural Network)

회귀(Regression)

기준선

테이블: 스코어 순위 중첩: BAD

누적반응률	반응검출률	누적 반응검출률	기준 반응률	기준 누적반응률	기준 반응검출률
95.20861	5.946797	5.946797	80.05034	80.05034	5
94.91638	5.910291	11.85709	80.05034	80.05034	5
94.81898	5.910291	17.76738	80.05034	80.05034	5
94.77027	5.910291	23.67767	80.05034	80.05034	5
94.74105	5.910291	29.58796	80.05034	80.05034	5
94.72157	5.910291	35.49825	80.05034	80.05034	5
94.70765	5.910291	41.40854	80.05034	80.05034	5
94.69721	5.910291	47.31883	80.05034	80.05034	5
94.6891	5.910291	53.22913	80.05034	80.05034	5
94.6826	5.910291	59.13942	80.05034	80.05034	5
94.67729	5.910291	65.04971	80.05034	80.05034	5
94.67286	5.910291	70.96	80.05034	80.05034	5
94.66911	5.910291	76.87029	80.05034	80.05034	5
94.6659	5.910291	82.78058	80.05034	80.05034	5
94.67955	5.910291	88.69087	80.05034	80.05034	5

차례

- 4.1 선형 회귀분석(Linear Regression Analysis)
- 4.2 로지스틱 회귀분석(Logistic Regression Analysis)
- 4.3 회귀분석의 특징과 제약
- 4.4 분석사례 - 1: 선형 회귀분석
- 4.5 분석사례 - 2: 로지스틱 회귀분석
- 4.6 분석사례 - 3: 신용평점표 작성
- 4.7 연습문제

회귀(Regression) 노드 - 속성 패널과 항 편집기

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

표본추출 탐색 수정 모델 평가 유틸리티 응용 프로그램 시계열

Chapter4_2

회귀 노드의 속성 패널

속성

일반

노드 ID Reg2

가져온 데이터

내보낸 데이터

노트

분석

변수

방정식(Equation)

주효과(Main Effects) 예

2요인 교호작용(Two-Factor) 예

다항식 항(Polynomial Terms) 예

다항식 차수(Polynomial Order) 2

사용자 항(User Terms) 예

항 편집기(Term Editor)

Class 타겟(Class Target)

회귀 유형(Regression) 로지스틱 회귀

연결함수(Link Function) 로짓(Logit)

모델 옵션(Model Option)

절편 생략(Suppress Intercept) 아니요

입력 코딩(Input Coding) Deviation

항 편집기(Term Editor)

항 편집기로 교호작용 항과 모델 항의 순서를 지정하여 모델을 사용자 정의할 수 있습니다.

다이어그램 Chapter4_2 열림

hckang(으)로서의 hckang hckang-pc에 연결