로지스틱 회귀분석과 다중판별분석 수행 보고서

Y2023011 윤요섭

Ⅰ. 자료 및 분석 방법

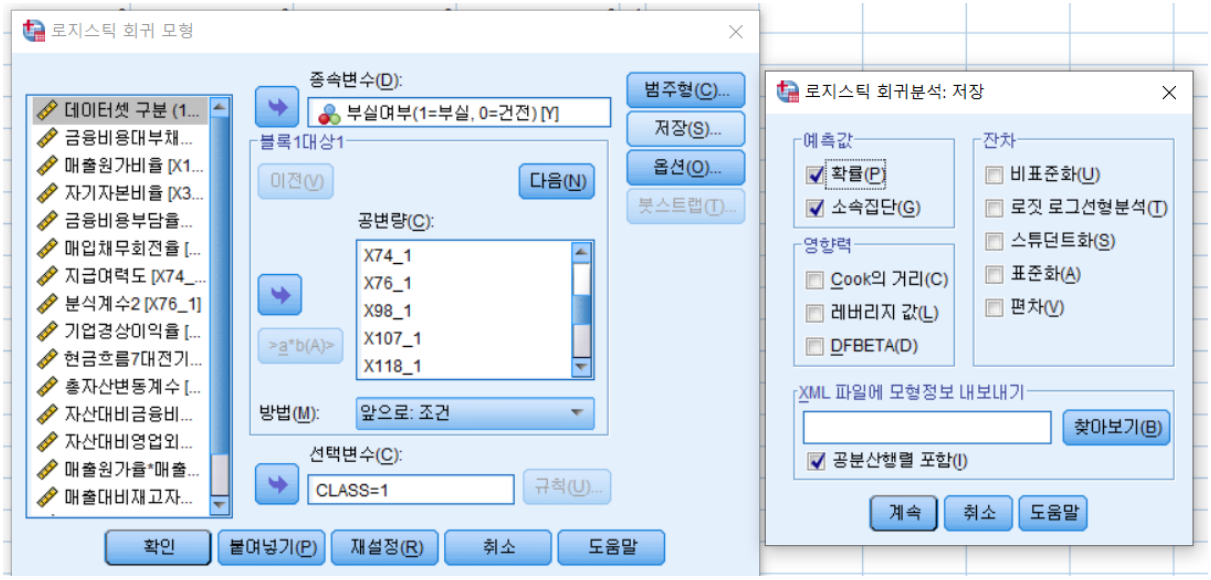
본 보고서는 bankruptcy.xlsx 데이터를 이용하여 IBM SPSS Statistics를 통해 실습과제를 수행하였다. bankruptcy.xlsx 데이터 변수와 이에 대한 설명은 다음과 같다.

변수명	변수설명
CLASS	데이터셋 구분 (1=학습, 0=검증)
X13_1	금융비용대부채비율
X18_1	매출원가비율
X36_1	자기자본비율
X42_1	금융비용부담율증가분
X43_1	매입채무회전율
X74_1	지급여력도
X76_1	분식계수2
X98_1	기업경상이익율
X107_1	현금흐름7대전기총부채
X118_1	총자산변동계수
X129_1	자산대비금융비용증가율
X130_1	자산대비영업외비용증가율
X133_1	매출원가율*매출원가평균증가비
X149_1	매출대비재고자산증가율
X157_1	총자본회전율*매출액증가비
Y	부실여부(1=부실, 0=건전)

bankruptcy.xlsx 데이터를 통해 로지스틱 회귀분석과 다중판별분석을 실시하였으며, 이 때, CLASS 변수가 1인 경우 학습용 데이터로 0인 경우 검증용 데이터로 활용하였다. 종속변수로는 Y(부실여부)로 설정하였으며, CLASS변수와 Y변수 외의 변수를 독립변수로 설정한 뒤, 학습을 진행하였다.

Ⅱ. 분석 과정과 결과 및 해석

Ⅱ-가. 이분형 로지스틱 모형을 통해 분석



다음과 같이, 종속변수로 부실여부, 선택변수로 데이터셋 구분을 한 뒤, 이 외의 변수들을 독립변수로 입력한 뒤, 예측값을 확률과 소속집단으로 선택한 뒤, 분석을 진행하였다.

분석 결과로 케이스 처리 요약은 아래와 같다.

케이스 처리 요약

가중되지 않은 케이스[a]		N	퍼센트
선택 케이스	분석에 포함	2134	79.9
	결측 케이스	0	.0
	합계	2134	79.9
비선택 케이스		536	20.1
합계		2670	100.0

a. 가중값을 사용하는 경우에는 전체 케이스 수의 분류표를 참조하십시오.

CLASS=1로 설정한 데이터 셋의 학습 데이터는 79.9%였으며, CLASS=0으로 설정한 데이터 셋의 검증 데이터의 경우, 20.1%로 구성되었다. 이분형 로지스틱 모형을 통한 분석의 경우, 총 15단계로 이루어져 있으며, 이는 15개의 독립변수가 선택되었다는 것을 의미한다.

로지스틱 회귀모형을 통해 분석 후 모형 요약과 해석은 아래와 같다.

모형 요약

단계	-2 Log 우도	Cox와 Snell의 R-제곱	Nagelkerke R-제곱
1	2517.823[a]	.187	.249
2	2061.560[b]	.343	.457
3	1908.062[b]	.389	.518
4	1839.570[b]	.408	.544
5	1779.111[b]	.425	.566
6	1747.342[b]	.433	.577
7	1723.621[b]	.439	.586
8	1701.445[b]	.445	.593
9	1679.679[b]	.451	.601
10	1665.525[b]	.454	.606
11	1654.145[b]	.457	.610
12	1647.228[b]	.459	.612
13	1639.502[b]	.461	.615
14	1635.049[b]	.462	.616
15	1631.051[b]	.463	.617

a. 모수 추정값이 .001보다 작게 변경되어 계산반복수 4에서 추정을 종료하였습니다.

b. 모수 추정값이 .001보다 작게 변경되어 계산반복수 6에서 추정을 종료하였습니다.

-2 Log 우도의 경우, 작을수록 가장 완벽한 모형에 적합하다고 볼 수 있으며, Cox와 Snell의 R-제곱, Nagelkerke R-제곱의 경우, 높을수록 가장 완벽한 모형에 적합하다고 볼 수 있다.
이를 통해, 변수가 15개 일 경우, -2 Log 우도가 가장 낮고, Cox와 Snell의 R-제곱, Nagelkerke R-제곱이 높으므로 가장 완벽한 모형에 적합하다고 볼 수 있다.

이분형 로지스틱 모형 분석 후, 분류표는 다음 장과 같다.

분류표[a]

감시됨			예측					
			선택 케이스[b]			비선택 케이스[c]		
			부실여부(1=부실, 0=건전)			부실여부(1=부실, 0=건전)		
			0	1	분류정확 %	0	1	분류정확 %
1 단계	부실여부(1=부실, 0=건전)	0	776	291	72.7	210	58	78.4
		1	278	789	73.9	69	199	74.3
	전체 퍼센트				73.3			76.3
2 단계	부실여부(1=부실, 0=건전)	0	802	265	75.2	205	63	76.5
		1	168	899	84.3	34	234	87.3
	전체 퍼센트				79.7			81.9
3 단계	부실여부(1=부실, 0=건전)	0	854	213	80.0	215	53	80.2
		1	162	905	84.8	37	231	86.2
	전체 퍼센트				82.4			83.2
4 단계	부실여부(1=부실, 0=건전)	0	868	199	81.3	219	49	81.7
		1	177	890	83.4	37	231	86.2
	전체 퍼센트				82.4			84.0
5 단계	부실여부(1=부실, 0=건전)	0	883	184	82.8	224	44	83.6
		1	172	895	83.9	37	231	86.2
	전체 퍼센트				83.3			84.9
6 단계	부실여부(1=부실, 0=건전)	0	888	179	83.2	227	41	84.7
		1	165	902	84.5	39	229	85.4
	전체 퍼센트				83.9			85.1
7 단계	부실여부(1=부실, 0=건전)	0	883	184	82.8	224	44	83.6
		1	162	905	84.8	41	227	84.7
	전체 퍼센트				83.8			84.1
8 단계	부실여부(1=부실, 0=건전)	0	891	176	83.5	221	47	82.5
		1	163	904	84.7	41	227	84.7
	전체 퍼센트				84.1			83.6
9 단계	부실여부(1=부실, 0=건전)	0	894	173	83.8	222	46	82.8
		1	158	909	85.2	38	230	85.8
	전체 퍼센트				84.5			84.3
10 단계	부실여부(1=부실, 0=건전)	0	894	173	83.8	222	46	82.8
		1	160	907	85.0	42	226	84.3
	전체 퍼센트				84.4			83.6
11 단계	부실여부(1=부실, 0=건전)	0	901	166	84.4	222	46	82.8
		1	163	904	84.7	42	226	84.3
	전체 퍼센트				84.6			83.6
12 단계	부실여부(1=부실, 0=건전)	0	905	162	84.8	226	42	84.3
		1	162	905	84.8	44	224	83.6
	전체 퍼센트				84.8			84.0
13 단계	부실여부(1=부실, 0=건전)	0	905	162	84.8	223	45	83.2
		1	162	905	84.8	45	223	83.2
	전체 퍼센트				84.8			83.2
14 단계	부실여부(1=부실, 0=건전)	0	905	162	84.8	225	43	84.0
		1	162	905	84.8	43	225	84.0
	전체 퍼센트				84.8			84.0
15 단계	부실여부(1=부실, 0=건전)	0	903	164	84.6	227	41	84.7
		1	162	905	84.8	45	223	83.2
	전체 퍼센트				84.7			84.0

a. 절단값은 .500입니다.

b. 선택 케이스 데이터셋 구분 (1=학습, 0=검증) EQ 1

c. 비선택 케이스 데이터셋 구분 (1=학습, 0=검증) NE 1

15단계로 진행된 로지스틱 회귀분석 모형에서 단계별 학습 데이터의 정확도의 경우, 12단계부터 14단계까지 84.8%로 가장 높았다. 단계별 검증 데이터의 정확도의 경우, 6단계에서 85.1%로 가장 높았다.

검증 데이터의 정확도로만 본다면 6단계가 85.1%로 가장 높으나, 분석의 목적이 부실 여부를 맞추는 것이므로, 실제로 부실일 경우에 부실을 잘 예측했는지 확인할 필요가 있다.

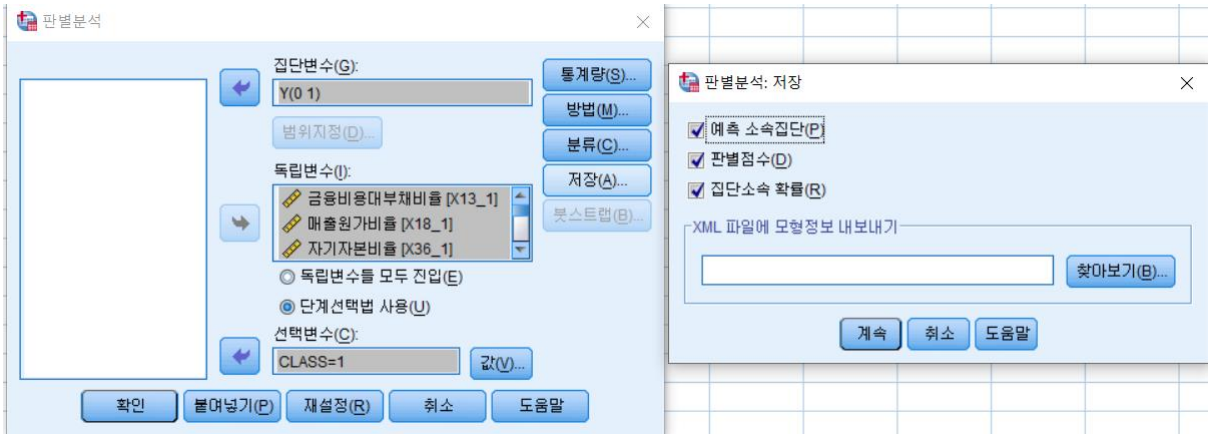
이를 확인하기 위해, Recall(재현도) 값의 결과인 실제로 부실일 때, 모델이 부실이라고 예측한 비율을 각 모델별로 비교하였다. 비교 대상의 단계별 모형으로는 -2 Log 우도가 가장 낮고, Cox와 Snell의 R-제곱, Nagelkerke R-제곱이 높았던 15단계의 모형과 학습 데이터의 정확도가 가장 높았던 12~14단계의 모형과 검증 데이터의 정확도가 가장 높았던 6단계의 모형이다.

[단계별 모형의 Recall]

모형의 단계	Recall (%)
6단계	85.4
14단계	84.0
12단계	83.6
13, 15단계	83.2

종합적으로 고려해보았을 때, 6단계에 포함된 변수인, X13_1(금융비용대부채비율), X36_1(자기자본비율), X98_1(기업경상이익율), X118_1(총자산변동계수), X129_1(자산대비금융비용증가율), X130_1(자산대비영업외비용증가율)으로 변수를 선택한 모형을 가장 적합한 모형으로 채택할 수 있다.

Ⅱ-나. 다중판별분석 모형을 통해 분석



다음과 같이, 집단변수로 부실여부, 선택변수로 데이터셋 구분을 한 뒤, 이 외의 변수들을 독립변수로 입력한 뒤, 분류의 경우 요약표를 선택하고, 저장시, 예측 소속집단, 판별점수, 집단소속 확률을 선택하고 분석을 진행하였다.

분석 결과로 케이스 처리 요약은 아래와 같다.

분석 케이스 처리 요약

가중되지 않은 케이스		N	퍼센트
유효		2134	79.9
제외	결측되었거나 범위를 벗어난 집단코드	0	.0
	적어도 하나 이상의 결측 판별변수	0	.0
	집단코드와 최소 하나 이상의 결측 판별변수가 누락되었거나 범위를 벗어남	0	.0
	선택되지 않음	536	20.1
	전체	536	20.1
전체		2670	100.0

CLASS=1로 설정한 데이터 셋의 학습 데이터는 79.9%였으며, CLASS=0으로 설정한 데이터 셋의 검증 데이터의 경우, 20.1%로 구성되었다. 이분형 로지스틱 모형을 통한 분석의 경우, 로지스틱 모형과 다르게 X74_1(지급여력도) 변수가 빠지면서, 총 14단계로 이루어져 있으며, 이는 14개의 독립변수가 선택되었다는 것을 의미한다. 단계선택 통계량에 진입된/제거된 변수는 다음 장과 같다.

[단계선택 통계량]

단계	진입된	Wilks 람다							
						정확한 F			
		통계량	자유도1	자유도2	자유도3	통계량	자유도1	자유도2	유의확률
1	자산대비영업외비용증가율	.824	1	1	2132.000	456.651	1	2132.000	.000
2	기업경상이익율	.708	2	1	2132.000	440.096	2	2131.000	.000
3	자기자본비율	.642	3	1	2132.000	396.707	3	2130.000	.000
4	자산대비금융비용증가율	.615	4	1	2132.000	332.550	4	2129.000	.000
5	총자산변동계수	.604	5	1	2132.000	278.941	5	2128.000	.000
6	금융비용대부채비율	.598	6	1	2132.000	238.605	6	2127.000	.000
7	매입채무회전율	.592	7	1	2132.000	208.949	7	2126.000	.000
8	현금흐름7대전기총부채	.588	8	1	2132.000	186.469	8	2125.000	.000
9	매출원가율*매출원가평균증가비	.584	9	1	2132.000	168.242	9	2124.000	.000
10	금융비용부담율증가분	.580	10	1	2132.000	154.024	10	2123.000	.000
11	분식계수2	.577	11	1	2132.000	141.326	11	2122.000	.000
12	매출대비재고자산증가율	.575	12	1	2132.000	130.409	12	2121.000	.000
13	총자본회전율*매출액증가비	.574	13	1	2132.000	121.047	13	2120.000	.000
14	매출원가비율	.572	14	1	2132.000	113.377	14	2119.000	.000

각 단계에서 전체 Wilks의 람다를 최소화하는 변수가 입력됩니다.

a. 최대 단계 수는 30입니다.

b. 입력할 최소 부분 F는 3.84입니다.

c. 제거할 최대 부분 F는 2.71입니다.

d. F 수준, 공차한계 또는 VIN 부족으로 계산을 더 수행할 수 없습니다.

판별모형 분석 후, 분류 결과는 아래와 같다.

[분류결과]

			부실여부(1=부실, 0=건전)	예측 소속집단		전체
				0	1	
선택된 케이스	원래값	빈도	0	911	156	1067
			1	185	882	1067
		%	0	85.4	14.6	100.0
			1	17.3	82.7	100.0
선택되지 않은 케이스	원래값	빈도	0	226	42	268
			1	49	219	268
		%	0	84.3	15.7	100.0
			1	18.3	81.7	100.0

a. 원래의 선택 집단 케이스 중 84.0%이(가) 올바로 분류되었습니다.

b. 원래의 비선택 집단 케이스 중 83.0%이(가) 올바로 분류되었습니다.

학습 데이터의 정확도는 84%, 검증 데이터의 정확도는 83%로 분류된 것을 확인할 수 있다.

앞서서 로지스틱 회귀모형의 결과와 비교할 경우, 로지스틱 모형의 6단계 모형과 비교할 경우, 학습 데이터의 정확도는 0.1%p 미세하게 높지만, 검증 데이터의 정확도의 경우, 2.1%p 정도 낮은 것을 확인할 수 있다.

따라서, bankruptcy.xlsx 데이터를 로지스틱 회귀분석과 다중판별분석 수행 시 로지스틱 회귀분석 모형이 더 적합하다고 볼 수 있다.