빅데이터의이해 HW13

– employee_data.sav의 회귀분석 –

201614010 영어영문학과(휴먼ICT) 박도준

회귀분석은 여러 변수들 간의 관계를 함수식으로 모형화 하여 한 개의 종속변수를 한 개 이상의 독립변수들의 선형 함수식으로 표현하는 분석 기법이다. 종속변수를 설명하는 독립변수가 한 개인 경우 단순선형회귀모형을 적용하며, 독립변수들이 두 개 이상 있는 경우 여러 개의 독립변수들을 동시에 고려하는 다중회귀모형을 적용한다. 본 과제에서는 employee_data.sav 데이터 파일의 다수의 변수들(피교육, 최초급여, 근무월수, 경력, 소수민족여부, 성별, 경영직여부, 관리직여부)이 현재급여에 미치는 인과관계에 대해 알고자 한다. 따라서 다중회귀분석을 적용하여 분석을 수행하고, 어떤 모형이 변수들 간의 관계를 가장 잘 대변하는 최적의 모형인지 알아보도록 하자.

회귀분석에 사용되는 독립변수와 종속변수는 모두 구간, 비율척도 이어야 한다. 하지만, 독립변수의 경우 0과 1로 구분되는 더미(dummy)변수의 형태로 변형시킨다면 범주형 변수도 사용이 가능하다. 따라서 본 회귀분석에서 독립변수로 적용하려는 여덟 가지의 변수 중 범주형 변수에 속하는 소수민족여부, 성별, 경영직여부, 관리직여부는 더미변수의 형태로 바꾸어 사용하도록 한다.

또한 회귀모형이 타당한 지를 확인하기 위해 두 가지 결과에 대한 검정이 이루어져야 한다. 첫 번째는 독립변수들이 종속변수에 선형적 영향을 주는지 알기 위해 실행하는 모형 전체에 대한 적합성 검정으로, 아래와 같은 가설을 설정할 수 있다.

귀무가설 H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$
(모든 독립변수들이 현재급여에 영향을 주지 않는다.)
대립가설 H_1: 적어도 하나의 $\beta_i \neq 0$, $i = 1, \ldots, 8$
(적어도 하나의 독립변수가 현재급여에 영향을 준다.)

두번째는 회귀계수에 대한 검정이다. 이는 회귀모형에 대한 검정에서 유의하다는 결론이 내려지면 각각의 독립변수들이 종속변수에 영향을 주는지를 알아보기 위한 검정으로, 아래와 같이 가설을 설정한다.

귀무가설 H_0: $\beta_i = 0$ (독립변수 X_i가 현재급여에 영향을 주지 않는다.)
대립가설 H_1: $\beta_i \neq 0$ (독립변수 X_i가 현재급여에 영향을 준다.)

모형 전체에 대한 적합성 검정은 분산분석을 통해 얻어지는 F값에 따라 귀무가설의 기각 여부를 결정하며, 회귀계수에 대한 검정은 검정통계량 T값에 의해 귀무가설의 기각여부를 결정한다.

이제 아래 표를 통해서 회귀분석의 결과를 확인해보도록 하자.

변수	b	표준오차(s.e.)	베타	t	유의확률
상수	-8455.59	3169.44		-2.668	.008
최초급여	1.34	.073	.618	18.397	.000
경영직	11255.52	1364.21	.252	8.251	.000
입사전 타 근무월수	-22.28	3.57	-.136	-6.248	.000
현 근무월수	148.62	31.35	.088	4.741	.000
관리직	6736.87	1629.55	.092	4.134	.000
피교육년수	499.31	159.99	.084	3.121	.002
남자	1870.06	760.48	.055	2.459	.014

[표 1] 현재급여에 대한 추정된 회귀모형

[표 1]은 단계별 회귀방법을 적용하여 얻은 모형 중 7단계에서 소수민족여부를 제외한 모든 변수가 입력되어 얻은 회귀분석의 결과이다. 위 모형은 R^2값이 0.843으로 7개의 모형 중 가장 높았으며, F 변화량의 유의확률은 0.014로 변화량 역시 유의함을 보여주었다. 분산분석표를 통해 얻은 F 통계량은 357.821로 7단계의 모형 중 가장 낮은 값을 가졌고, 유의확률이 0.000으로 나타났다. 따라서 모형 전체에 대한 적합성 검정의 귀무가설(모든 독립변수들이 현재급여에 영향을 주지 않는다)을 기각하게 된다. T 통계량에 대한 유의확률은 모두 0.005보다 작은 것으로 나타났다. 따라서 회귀계수 검정에 대한 귀무가설(독립변수 X_i가 현재급여에 영향을 준다) 역시 기각하게 된다. 자기상관이 있는지를 말해주는 Durbin-Watson 통계량 값은 1.843으로 나타났는데, 이 값은 2에 가까우므로 자기상관이 존재하지 않음을 보여준다.

[표 1]의 유의확률을 보면 모든 회귀계수의 유의확률이 0.05보다 작게 나타나는데, 이는 유의수준 5%에서 유의하다는 사실을 말해준다. 다음으로 베타계수를 보면 최초급여(0.618)가 가장 높으며, 다음으로 경영직 범주(0.252), 입사전 타 근무월수(-0.136) 순으로 높음을 볼 수 있는데, 이는 현재급여에 가장 많은 영향력을 미치는 순서로 해석된다. 입사 전 타 근무월수의 베타계수는 음수 값을 갖는데, 이는 현재급여에 음의 방향으로 영향을 미치는 것을 나타낸다.