빅데이터의이해 HW11

- employee_data.sav의 상관관계 분석 -

201614010 영어영문학과(휴먼ICT) 박도준

employee_data.sav은 사원들의 정보를 담고 있는 데이터 셋이다. 이 중 수치형 자료(구간·비율척도)에 해당하는 현재급여, 최초급여, 피교육년수, 현 근무월수, 입사전 타 근무월수, 총 5개의 수치형 자료를 선정하여 위 변수들 간의 관련성을 상관관계 분석을 통해 알아보고자 한다.

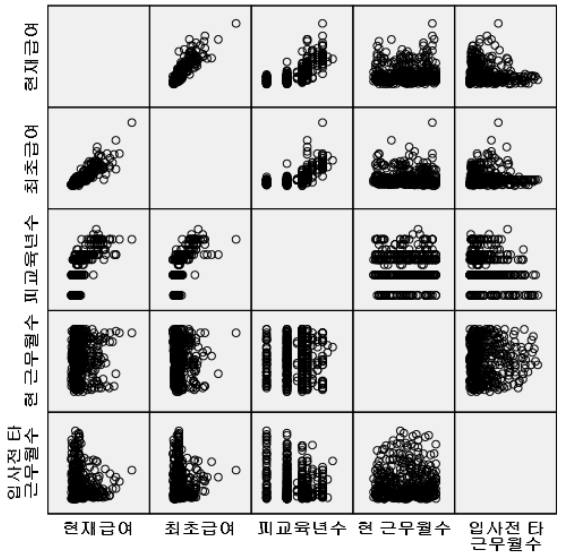

그림1: 행렬산점도 분석 결과

그림 1은 상관관계 분석의 대상이 되는 변수들의 행렬산점도이다. 위 그래프를 보면 먼저 현재급여와 최초급여 간에 선형 관계가 존재함을 확인해 볼 수 있다. 또한 피교육년수와 최초급여, 피교육년수와 현재급여 간에도 선형 관계가 있다고 여겨질 만한 그래프의 양상을 띄고 있는데, 정확한 상관관계를 파악하기 위한 상관관계 분석이 필요해 보인다. 위 데이터 셋은 표본의 수가 474개로서 대표본에 해당하기에, 피어슨 상관계수 분석을 수행하는 것이 적절하다.

피어슨 상관계수(유의확률)

	현재급여	최초급여	피교육년수	현 근무월수	입사전 타 근무월수
현재급여		.880** (.000)	.661** (.000)	.084 (.067)	-.097* (.034)
최초급여			.633** (.000)	-.020 (.668)	.045 (.327)
피교육년수				.047 (.303)	-.252** (.000)
현 근무월수					.003 (.948)
입사전 타 근무월수					

***. 상관이 0.01 수준에서 유의합니다(양쪽).*

**. 상관이 0.05 수준에서 유의합니다(양쪽).*

표 1: 피어슨 상관 계수 분석 결과

표 1은 현재급여, 최초급여, 피교육년수, 현 근무월수, 입사전 타 근무월수 간의 피어슨 상관관계 분석 결과를 보여주고 있다. 행렬산점도에서 가장 명확한 선형적 상관관계를 보여준 최초급여와 현재급여는 유의수준 0.01에서 상관계수 0.808, 유의확률 0.000으로 두 변수 간에 양의 상관관계가 있음을 보여준다. 다음으로 높은 상관계수를 살펴보면, 피교육년수와 현재급여, 피교육년수와 최초급여가 있는데, 각각 유의수준 0.01에서 0.661(0.000), 0.633(0.000)의 상관계수(유의확률)를 가지고, 따라서 양의 상관관계가 있다고 판단해 볼 수 있다. 뿐만 아니라 입사전 타 근무월수와 현재급여의 상관계수(유의확률)가 유의수준 0.05에서 -0.97(0.34), 입사전 타 근무월수와 피교육년수의 상관계수(유의확률)가 유의수준 0.01에서 0.252(0.000)으로 분석됨으로 음의 상관관계가 있다고 결론 지을 수 있다.

피어슨 상관계수(유의확률)

	Diff	피교육년수	현 근무월수
diff		.582** (.000)	.147** (.001)
피교육년수			.047 (.303)
현 근무월수			

**. 상관이 0.01 수준에서 유의합니다(양쪽).

표 2: diff, 피교육년수, 현 근무월수 간의 피어슨 상관계수 및 유의확률

표 2에서는 현재급여와 최고급여의 차이를 새로운 변수 diff를 만들어서 추가하여 diff, 피교육년수, 현 근무월수, 총 3개의 변수를 선택하여 피어슨 상관관계를 분석하였다. diff와 피교육년수의 상관계수는 유의수준 0.01에서 0.582, 유의확률 0.000로서 선형적 관련성이 있음을 알 수 있다. 또한 diff와 현 근무월수의 상관계수가 유의수준 0.001에서 0.147, 유의확률 0.001을 가짐으로 유의미한 상관관계가 있음을 확인해 볼 수 있다.

편상관계수(유의확률)

제어변수		diff	피교육년수	현 근무월수
최초급여	Diff		.281 (.000)	.214 (.000)
	피교육년수			.077 (.093)
	현 근무월수			

표 3: 최초급여를 통제한 조건에서의 diff, 피교육년수, 현 근무월수 간의 편상관계수 및 유의확률

표 3은 diff, 피교육년수, 현근무월수 변수에 유의미한 관련성을 가질 것으로 여겨지는 최초급여를 통제하고 세 변수 간의 상관계수를 구한 결과이다. 피교육년수와 diff 변수의 상관관계는 유의수준 0.001에서 상관계수 0.281, 유의확률 0.000을 보여준다. 최초급여를 통제하지 않은 상황에서의 상관계수 0.582보다 다소 낮게 나타났지만, 제어변수를 통제하였음에도 유의미한 상관관계가 있음을 보여주고 있다. 또한 현 근무월수와 diff변수 간의 상관관계는 유의수준 0.001에서 0.214, 유의확률 0.000을 갖는다. 이는 제어변수를 통제하기 전인 상관계수 0.147, 유의확률 0.001보다 다소 높아진 결과이다. 이를 통해 현근무월수와 diff변수는 제어변수를 통제하였을 때 보다 더 명확한 상관관계를 가진다는 결론을 얻을 수 있다.