*Journal of the Korean Data & Information Science Society*
*2015,* **26**(2), 387–397

http://dx.doi.org/10.7465/jkdi.2015.26.2.387
한국데이터정보과학회지

# 중소기업 청년인턴 이직횟수 결정요인 분석

박성익[1] · 류장수[2] · 김종한[3] · 조장식[4]

[1]경성대학교 국제무역통상학과 · [2]부경대학교 경제학부 ·
[3]경성대학교 경제금융물류학부 · [4]경성대학교 정보통계학과

접수 2015년 2월 23일, 수정 2015년 3월 8일, 게재확정 2015년 3월 18일

**요 약**

본 연구에서는 청년인턴 DB와 고용보험 DB를 사용하여 중소기업 청년인턴의 이직횟수에 영향을 미치는 요인을 분석하였다. 이직횟수는 음수가 아닌 정수 값만 가지는 계수 데이터 (count data)이므로 일반적인 선형회귀모형을 적용하는 것은 문제가 있다. 따라서 계수 데이터에 적합한 회귀모형으로 포아송 회귀모형, 영과잉 포아송 회귀모형, 음이항 회귀모형, 영과잉 음이항 회귀모형 등 4개의 회귀모형을 적용하였다. 분석결과 최적모형으로 영과잉 음이항 회귀모형이 선택되었다. 주요 분석결과를 정리하면 다음과 같다. 첫째, 통제집단 (비인턴집단)에 비해서 처리집단 (인턴집단)이 통계적으로 유의하게 이직경험이 낮게 나타났다. 둘째, 연령이 작을수록 통계적으로 유의하게 이직경험이 낮게 나타났다. 셋째, 여자에 비해서 남자가 유의하게 이직횟수가 높게 나타났다. 마지막으로 기업규모가 클수록 이직횟수가 유의하게 감소하는 것으로 나타났다.

주요용어: 과대산포, 다중대응분석, 영과잉, 음이항 회귀모형, 포아송 회귀모형.

## 1. 서론

중소기업 청년인턴제는 청년 미취업자에게 중소기업 등의 인턴경험을 제공함으로써, 정규직 취업을 촉진하고 중소기업의 인력난을 해소하려는 목적으로 2009년부터 시행되어 해마다 그 규모가 확대되고 있다. 중소기업 청년취업인턴제의 효과에 대해서는 그동안 많은 연구자들에 의해서 분석된 바 있다 (Lee, 2009; Ryu, 2012; Joo 등, 2013; Ryu 등, 2014). 이들 연구는 중소기업 청년인턴제도가 정부의 정책의도와 부합되게 청년들의 고용창출 및 정규직 전환에 긍정적인 효과를 미쳤다는 연구결과를 제시하고 있다. 그런데 중소기업 청년인턴제도는 그 취지가 중소기업의 인력난을 해소하려는 것도 있다는 점을 유념할 필요가 있다. 이러한 관점에서 청년인턴으로 취업한 사람이 타 직장으로 이직하게 되면, 정책에서 의도하는 목적을 달성하였다고 보기에는 어느 정도 한계가 있다. 그럼에도 불구하고, 기존의 연구에서는 청년인턴의 이직을 결정하는 요인에 대한 분석을 소홀히 한 것이 사실이다. 이에 본 연구는 청년인턴의 이직을 결정하는 요인에 대하여 분석하고자 한다.

그런데 이직횟수는 계수 데이터 (count data)로서 음수가 아닌 정수 값만을 가짐으로서 일반적인 선형회귀분석을 적용하기에는 정규성, 등분산성, 선형성 등의 가정을 충족하지 못하는 등 많은 문제점을 갖고 있다. 이러한 문제점을 극복할 수 있는 대안으로 포아송 회귀모형 (Poisson regression model)은

[1] (608-736) 부산광역시 남구 수영로 309 번지, 경성대학교 국제무역통상학과, 교수.
[2] (608-737) 부산광역시 남구 용소로 45, 부경대학교 경제학부, 교수.
[3] (608-736) 부산광역시 남구 수영로 309 번지, 경성대학교 경제금융물류학부, 교수.
[4] 교신저자 : (608-736) 부산광역시 남구 수영로 309 번지, 경성대학교 정보통계학과, 교수.
E-mail: jscho@ks.ac.kr

종속변수가 양의 정수 값만 가질 때 적합시킬 수 있는 가장 기본적인 모형이지만 평균과 분산이 동일해야 한다는 제약이 있다. 그러나 많은 경우 평균보다 분산이 커지는 과대산포 (over-dispersion)의 문제가 발생한다. 이러한 문제가 존재하는 데이터에서 포아송 회귀모형을 적합 시킨다면 모형적합의 효율성이 떨어지게 된다. 한편 음이항 회귀모형은 이분산성 (hetero-skedasticity)을 허용하는 분산함수로 정의되어 평균과 분산이 크게 다른 경우에도 포아송 회귀모형의 단점을 해결해 준다. 즉 음이항 회귀모형은 각 사례의 관측되지 않은 비동질 (heterogeneity) 요소를 모형에 포함시킴으로써 포아송 모형의 문제를 보완해 준다.

한편 계수 데이터에서 포아송 회귀모형이나 음이항 회귀모형을 적합시키는 과정에서 0이 가정된 분포에 비해 과다하게 포함되는 경우가 많이 발생한다. 이와 같은 경우에는 영과잉 (zero inflation)이 반영된 포아송 회귀모형이나 음이항 회귀모형을 적합시켜야 한다. 계수 데이터에 대한 회귀모형과 관련된 연구로는 다음과 같다. Lee (2013)는 영과잉 현상에 적용하기 위한 포아송-로그정규 (zero inflated Poisson-lognormal) 모형을 제시하였으며, Kim과 Eum (2010)은 정상적인 포아송 확률분포보다 0의 값이 과잉 관측되는 경우 0에 관한 확률과 포아송분포의 평균에 대한 추정량을 제시하였다. 그리고 Chun과 Kim (2008)은 음이항 모형을 이용한 SMS 확산요인에 관해 연구를 했으며, Lee (2011)는 경제적 불확실성이 출산 의사결정에 미치는 효과를 음이항 회귀분석을 이용해서 자녀 수를 분석하였다. 그 외에도 Cox (1983), Dean과 Lawless (1989), Kim (2005), Choi와 Ko (2011), Jeong과 Choi (2014), Cho (2014) 등의 연구가 있다.

본 연구에서는 중소기업 청년취업 인턴제도의 참여자들을 대상으로 고용의 질적 측면을 반영하는 이직횟수에 미치는 영향력을 분석하고자 한다. 이직횟수가 계수 데이터이므로, 포아송 회귀모형, 영과잉 포아송 회귀모형, 음이항 회귀모형 및 영과잉 음이항 회귀모형 등 모두 4개의 모형을 적합하여 최적의 모형을 선택한다. 먼저 2장에서는 데이터소개와 간단한 기술통계를 제시하였으며, 3장에서는 의사결정나무분석을 활용하여 이직횟수와의 관련성 있는 독립변수들을 탐색하였다. 그리고 4장에서는 이직횟수에 대한 최적의 회귀모형을 탐색하고, 마지막으로 5장에서는 본 연구의 결론을 제시하였다.

## 2. 데이터 설명

본 연구의 목적은 정부의 정책인 중소기업 청년인턴제도에 참여하여 정규직으로 전환된 사람들의 이직 결정요인을 분석하는 것이다. 그런데 정부의 정책효과를 분석하기 위해서는 유사 실험설계 (quasi-experimental design)를 활용하는 것이 일반적이다. 즉, 인턴사업 참여자와 가능한 한 유사한 특성을 가진 비참여자 집단으로 통제집단을 구성한 후, 사업에 참여한 처리집단과 참여하지 않은 통제집단의 결과를 비교함으로써 정부 사업의 효과를 측정한다. 본 연구에서는 인턴사업 참여자와 동질적인 특성을 가진 통제집단을 구성하기 위하여 청년인턴 DB와 고용보험 DB에서 성향점수매칭법 (propensity score matching method)을 활용하였다. 이에 대한 자세한 내용은 Ryu 등 (2014)의 자료를 참고하기 바란다.

본 연구에서는 분석을 위해 성별 (sex), 연령 (age), 직장소재지 (area), 직장규모 (size), 직종 (type), 이직횟수 ($y$) 등과 청년인턴 참여자 여부 (group) 변수를 사용하였다. 이 변수들에 대한 설명은 아래 Table 2.1과 같다.

**Table 2.1** Variables explanation

| variables | explanation | categories | scale |
|---|---|---|---|
| group | participant or not | 0='no (control)', 1='yes (treatment)' | nominal |
| sex | male dummy | 0='no', 1='yes' | nominal |
| age | The age at the time of the survey point | 0='≤20', 1='21∼25', 2='26∼30', 3='>30' | ordinal |
| area | metropolitan area of company | 0='Capital', 1='Chungchung', 2='Junra', 3='Daekyung', 4='Dongnam', 5='Gangwon/Jeju' | nominal |
| size | number of workers | 0='≤9', 1='10∼49', 2='50∼99', 3='100∼299', 4='>300' | ordinal |
| type | type of jobs | 0='job1', 1='job2', 2='job3', 3='job4', 4='job5' | nominal |
| $y$ | no. of change of jobs | | continuous |

Note: job1=administrative and office work, job2=professional service-related work, job3=simple service-related work, job4=science and engineering-related work, job5=simple production-related work.

다음으로 이직횟수에 대한 막대그래프 결과는 아래 Figure 2.1과 같다.

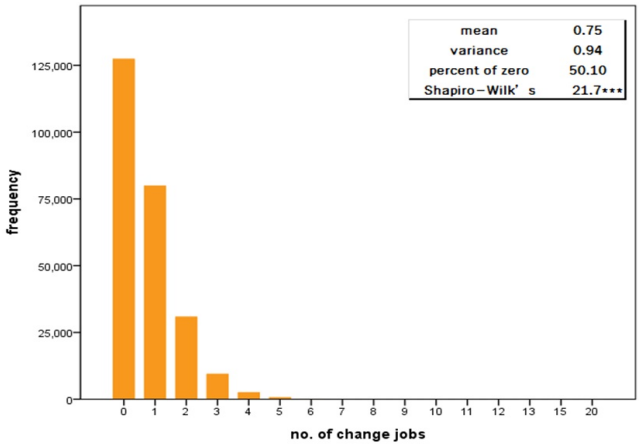


**Figure 2.1** Bar chart for number of change of jobs

Figure 2.1의 결과에 따르면 이직횟수가 0회로 가장 많이 나타났으며 (50.1%), 이직횟수가 증가할수록 빈도는 작아짐을 알 수 있다. 그리고 이직횟수의 평균은 0.75회이고 분산은 0.94회로 나타나서 평균에 비해서 분산이 더 크다는 것을 알 수 있다. 이런 결과는 이직횟수가 평균에 비해서 분산이 큰 과대산포의 문제와 0이 가정된 분포에 비해 과다하게 포함되는 영과잉의 문제가 존재할 가능성이 있음을 시사한다. 또한 이직횟수의 분포형태가 Shapiro-Wilk's 검정결과 정규분포를 따르지 않음을 알 수 있다 (21.7***).

아래 Table 2.2는 이직경험 유무에 따라 처리집단과 통제집단별로 독립변수들에 대한 기술통계 분석 결과를 나타낸 것이다.

**Table 2.2** Descriptive statistic analysis

| variables | category | n | $y=0$ | | $y>0$ | | | |
|---|---|---|---|---|---|---|---|---|
| | | | control | treatment | control | | treatment | |
| | | | $p(y=0)$ | $p(y=0)$ | mean | s.d. | mean | s.d. |
| | total | 252,921 | 49.0 | 52.0 | 1.58 | 0.91 | 1.46 | 0.79 |
| sex | female | 88,436 | 48.1 | 50.2 | 1.54 | .85 | 1.42 | .72 |
| | male | 164,485 | 49.5 | 52.9 | 1.60 | .94 | 1.49 | .82 |
| age | ≤20 | 23,905 | 63.7 | 65.6 | 1.42 | .75 | 1.35 | .66 |
| | 21∼25 | 92,516 | 46.8 | 49.7 | 1.56 | .88 | 1.46 | .77 |
| | 26∼30 | 125,705 | 47.5 | 51.1 | 1.60 | .94 | 1.47 | .80 |
| | >30 | 10,795 | 49.7 | 52.0 | 1.65 | 1.01 | 1.61 | 1.01 |
| area | Capital | 155,267 | 49.5 | 52.9 | 1.54 | .87 | 1.43 | .74 |
| | Chungchung | 19,847 | 47.0 | 51.5 | 1.62 | .94 | 1.52 | .88 |
| | Junra | 18,429 | 48.6 | 50.6 | 1.61 | .92 | 1.53 | .85 |
| | Daekyung | 26,675 | 48.6 | 50.6 | 1.64 | 1.02 | 1.53 | .85 |
| | Dongnam | 28,007 | 48.6 | 49.6 | 1.65 | .99 | 1.49 | .84 |
| | Gangwon/Jeju | 4,696 | 47.6 | 51.3 | 1.60 | .89 | 1.46 | .76 |
| size | <10 | 71,738 | 41.0 | 39.1 | 1.58 | .90 | 1.46 | .76 |
| | 10∼49 | 83,431 | 50.8 | 53.5 | 1.56 | .90 | 1.46 | .79 |
| | 50∼99 | 40,541 | 53.2 | 57.7 | 1.57 | .92 | 1.46 | .80 |
| | 100∼299 | 43,274 | 54.5 | 59.4 | 1.59 | .93 | 1.48 | .81 |
| | ≥300 | 13,937 | 54.5 | 62.5 | 1.59 | .94 | 1.48 | .79 |
| type | job1 | 123,606 | 48.8 | 50.9 | 1.56 | .88 | 1.44 | .74 |
| | job2 | 16,586 | 48.3 | 51.9 | 1.48 | .77 | 1.40 | .69 |
| | job3 | 6,750 | 48.0 | 45.7 | 1.60 | .94 | 1.49 | .81 |
| | job4 | 71,645 | 49.9 | 55.0 | 1.58 | .94 | 1.47 | .82 |
| | job5 | 34,333 | 48.3 | 50.8 | 1.68 | .99 | 1.58 | .91 |

Table 2.2의 결과를 변수별로 살펴보면 다음과 같다. 먼저 이직경험이 없는 경우 ($y = 0$)에 대해서 살펴보면, 청년인턴 미참여자 (control)에 비해서 참여자 (treatment)가 이직경험 없음의 비율이 더 높게 나타났다. 독립변수별로 청년인턴 참여자를 중심으로 이직경험 없음의 비율을 살펴보면, 여자에 비해서 남자가 더 높게 나타났으며, 연령은 20세 이하인 경우를 제외하고는 연령이 증가할수록 증가하는 것으로 나타났다. 광역권별로 보면, 수도권이 가장 높게 나타났으며, 그 다음으로 충청권, 강원/제주권, 전라권, 대경권의 순으로 나타났다. 직장규모별로는 직장규모가 커질수록 이직경험 없음의 비율이 증가함을 수 있다. 또한 직종별로는 직종4가 이직경험 없음의 비율이 가장 높게 나타났으며, 그 다음으로 직종2, 직종1, 직종5의 순으로 나타났다.

한편 이직경험이 있는 경우 ($y > 0$), 평균 이직횟수는 청년인턴 미참여자 (1.58)에 비해서 참여자 (1.46)가 더 낮게 나타났다. 독립변수별로 청년인턴 참여자를 중심으로 평균 이직횟수를 살펴보면, 여자에 비해서 남자가 다소 높게 나타났으며, 연령이 증가할수록 평균 이직횟수가 증가하는 경향을 보이고 있다. 광역권별로 보면 전라권, 대경권이 상대적으로 다른 광역권에 비해서 높게 나타났으며, 직장규모별로는 규모가 증가할수록 평균 이직횟수가 증가하는 것으로 나타났다. 직종별로는 직종5가 가장 높게 나타났고, 그 다음으로 직종3, 직종4의 순으로 나타났다.

한편 Table 2.1의 수준별 독립변수들이 이직횟수와의 상호관련성을 알아보기 위해 다중대응분석 (multiple correspondence analysis)을 실시하였다. 다중대응분석은 개체 (케이스)와 범주에 계량적 수치를 부여함으로써 범주형 데이터를 수량화하는 분석기법으로 내적 일관성의 원리로부터 범주의 수량화를 실시하는 분석기법이다. 여기서 편의상 이직횟수는 없음 (none), 한번 (1), 두 번 이상 (>1) 등 3개의 그룹으로 범주화 하였다. 아래 Figure 2.2는 다중대응 분석을 실시한 결과이다.

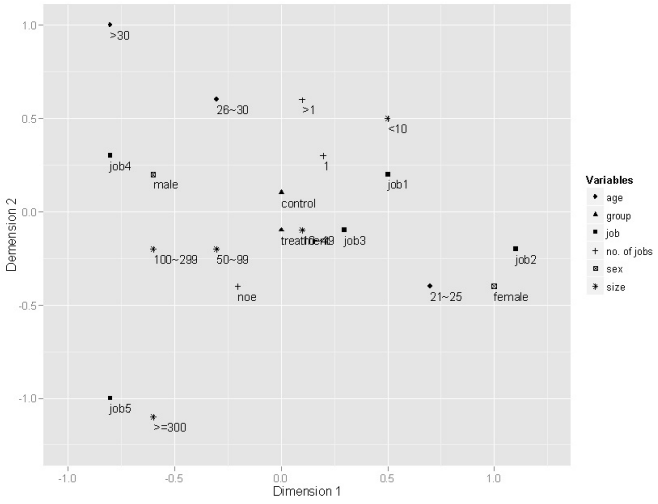


**Figure 2.2** The result of multiple correspondence analysis

위의 Figure 2.2의 결과로부터 다음과 같은 결과를 알 수 있다. 먼저 이직횟수는 Y축을 기준으로 없음은 음의 방향에 위치하고 있으며, 한번과 두 번 이상으로 증가할수록 Y축의 양의 방향으로 위치하고 있음을 알 수 있다. 이와 같은 현상은 기업규모와 연령에서도 나타나는데, 기업규모가 큰 기업이고 연령이 작을수록 Y축의 음의 방향에 위치하고 있으며, 기업규모가 작고 연령이 많을수록 Y축의 양의 방향에 위치하고 있다. 한편 남자는 X축의 음의 방향에, 여자는 X축의 양의 방향에 위치하고 있으며, 직종

1, 2, 3은 X축의 양의 방향에 위치하고 있고, 직종 4, 5는 X축의 음의 방향에 위치하고 있음을 알 수 있다. 이러한 결과를 통해서 이직경험이 없는 경우는 기업규모가 클수록, 청년인턴 미참여자에 비해서 참여자, 여자에 비해서 남자, 연령은 작을수록 상대적으로 가까운 거리에 위치하고 있어서 이들 간의 상호관련성은 높은 것으로 나타났다. 한편 이직경험이 있는 경우는 기업규모가 작을수록, 연령이 높을수록, 그리고 청년인턴 참여자에 비해서 미참여 집단이 상대적으로 가까운 거리에 위치하고 있어서 이들 간의 상호관련성이 높은 것으로 나타났다.

## 3. 의사결정나무 분석

이 절에서는 독립변수들이 이직횟수에 미치는 효과를 분석하기 위해 비모수적인 방법인 의사결정나무 분석을 이용하고자 한다. 즉 의사결정나무 분석을 통해 이직횟수에 영향을 미치는 각 수준별 독립변수들에 대한 고차의 상호작용효과를 분석한다. 이를 위해 지니지수 (Gini index)를 분리기준으로 사용하였으며, 이지분리를 수행하는 CART (classification and regression trees; Breiman 등, 1984) 알고리즘을 사용하였다. 의사결정나무 분석과 관련된 선행연구로는 Cho (2010, 2012), Cho와 Park (2012), Jung과 Min (2013) 등이 있다. 정지규칙으로는 최대나무깊이 (maximum tree depth)는 3으로 설정하였으며, 최소 케이스 수 (minimum number of cases)에서 부모마디 (parent node)는 5,000, 자식마디 (child node)는 1,000으로 설정하였으며, 적절한 가지치기 (pruning)를 병행하였다. 의사결정나무분석 결과는 아래 Figure 3.1과 같다.

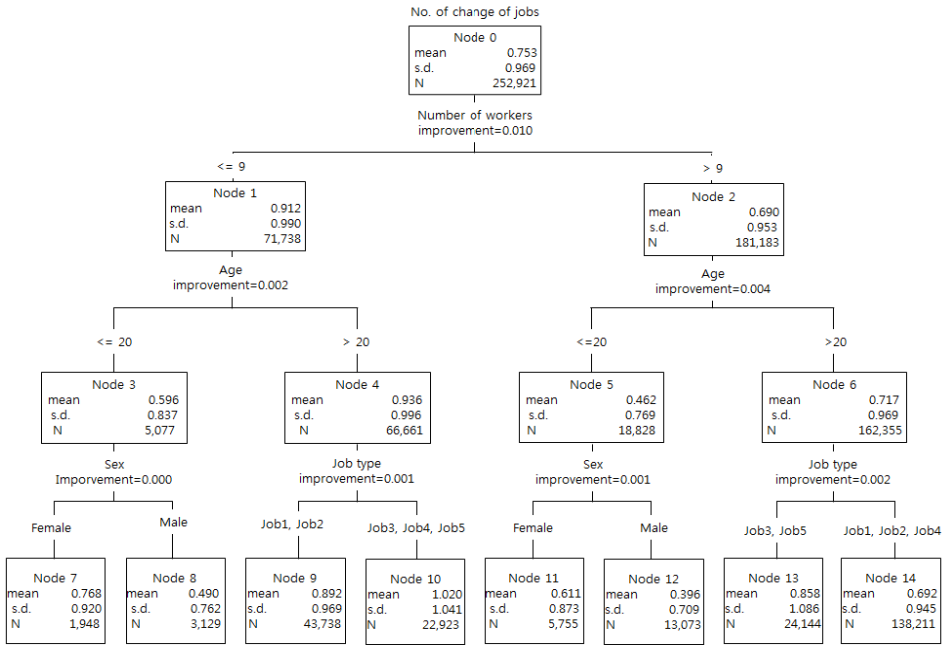


**Figure 3.1** Decision tree analysis for number of change of jobs

Figure 3.1의 결과에 따르면, 이직횟수에 가장 많은 영향을 미치는 변수로는 기업규모이며, 그 다음으로 연령, 성별, 직종 등의 순으로 나타났다. 먼저 기업규모가 9인 이하인 경우를 살펴보면 다음과 같다. 연령이 20세 이하인 집단 (0.596)에 비해서 20세 초과인 집단 (0.936)의 이직횟수가 더 높게 나타났

다. 특히 연령이 20세 이하인 집단 중에서는 여자 (0.768)가 남자 (0.490)에 비해서 이직횟수가 더 높았으며, 20세 초과인 집단 중에서는 직종1과 직종2인 집단 (0.892)에 비해서 직종3, 직종4, 직종5인 집단 (1.0230)의 이직횟수가 더 높게 나타났다.

한편 기업규모가 10인 이상인 경우는 연령이 20세 이하인 경우 (0.462)에 비해서 20세 초과인 경우 (0.717)의 이직횟수가 더 높게 나타났다. 특히 연령이 20세 이하인 경우 중에서는 남자 (0.396)인 경우에 비해서 여자 (0.611)의 이직횟수가 더 높게 나타났으며, 연령이 20세 초과인 집단 중에서는 직종1, 직종2, 직종4인 경우 (0.692)에 비해서 직종3과 직종5인 경우 (0.858), 이직횟수가 더 높은 것으로 나타났다.

## 4. 이직횟수 결정요인 분석

이 장에서는 이직횟수에 대한 결정요인분석을 하고자 한다. 앞의 Figure 1에서 언급했듯이, 이직횟수가 0회의 비율이 50.1%로 가장 높았으며 이직횟수가 증가할수록 그 비율은 큰 폭으로 감소하는 경향을 띠고 있다. 또한 이직횟수의 평균은 0.75회이고 분산은 0.94회로 나타나서, 계수 데이터에서 영과잉과 과대산포의 문제가 존재할 가능성이 있다. 따라서 모형 적합에서 영과잉과 과대산포를 반영할 수 있는 모형으로 적합할 필요성이 있다는 것을 시사한다.

한편 포아송 회귀모형 (Poisson regression model)은 종속변수가 음이 아닌 정수 값만 가질 때 적합할 수 있는 가장 기본적인 모형으로, 독립변수 $X_1, X_2, \cdots, X_k$가 주어질 때 종속변수 $Y_i$의 확률은 다음과 같은 포아송 분포에 의해 구해진다.

$$P(Y_i = y_i | X_1, \cdots, X_k) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!},\ y_i = 0, 1, \cdots \tag{4.1}$$

식 (4.1)에서 평균 모수 $\mu_i$ ($i$번째 조사대상자의 조건부 평균 이직횟수)는 $i$번째 조사대상자의 설명변수 벡터의 함수이며 $E(y_i|\boldsymbol{x_i}) = \mu_i = \exp(\boldsymbol{x_i}'\boldsymbol{\beta})$를 의미한다. 여기서 $\boldsymbol{x_i} = (x_1, \cdots, x_k)'$이고 $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)$이다. 한편 포아송 회귀모형은 모형의 정의에 의해 조건부 평균과 조건부 분산이 동일하다는 제약이 있다. 하지만 실제 얻어지는 계수 데이터는 종종 평균에 비해서 분산이 큰 경향을 보이는 과대산포 (over-dispersion)가 존재하면 모형적합의 효율성이 떨어지게 된다.

음이항 회귀모형 (NB : negative binomial regression)은 포아송 회귀모형의 일반화된 모형으로 이분산성 (hetero-skedasticity)을 반영하는 분산함수로 정의되어 평균과 분산이 크게 다른 경우에도 포아송 회귀모형의 문제점을 해결해 준다. 즉 음이항 회귀모형은 $i$번째 조사대상자의 관측되지 않은 비동질 (heterogeneity) 요소를 모형에 포함시킴으로써 포아송 모형을 일반화시킨 형태이며 과대산포를 보완해 준다. 독립변수 $X_1, X_2, \cdots, X_k$가 주어질 경우 종속변수 $Y_i$가 음이항 분포를 따를 때 식 (4.2)와 같이 주어진다.

$$P(Y_i = y_i | X_1, \cdots, X_k) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i},\ y_i = 0, 1, \cdots. \tag{4.2}$$

여기서 $\alpha$는 과대산포를 나타내는 모수이며 $E(y_i|\boldsymbol{x_i}) = \mu_i = \exp(\boldsymbol{x_i}'\boldsymbol{\beta})$이다.

한편 포아송 회귀모형을 적합하는 과정에서 0이 가정된 분포에 비해 과다하게 포함되는 경우가 많이 발생한다. 이와 같은 경우에 종속변수가 0인 부분과 0이 아닌 부분으로 나누어 베르누이 분포와 포아송 분포를 혼합하여 만든 모형을 적용한다. 즉 확률 $p_i$로 이직횟수가 0의 값을 가지고, 확률 $1 - p_i$로 이직횟수가 포아송 분포를 따른다고 가정한다. 여기서 $p_i$ ($0 \le p_i < 1$)는 0에서 과잉확률을 의미한다. 이와

같은 혼합모형은 아래 식 (4.3)과 같이 나타낼 수 있으며, 영과잉 포아송 회귀모형 (ZIP : zero inflated Poisson)이라 한다.

$$P(Y_i = y_i|X_1, \cdots, X_k) = \begin{cases} p_i + (1-p_i)e^{-\mu_i}, y_i = 0 \\ (1-p_i)\dfrac{e^{-\mu_i}\mu_i^{y_i}}{y_i!},\ y_i = 1, 2, \cdots \end{cases} \tag{4.3}$$

식 (4.3)에서 이직횟수가 없는 경우는 확률 $p_i$를 가지는 베르누이 분포를 따르므로 로지스틱 연결함수 (logit link function)를 사용하여 설명변수를 투입하고, 이직횟수가 있는 경우는 포아송 분포를 따라 로그-선형 모형의 형태로 설명변수를 투입하여 식 (4.4)와 같이 나타낼 수 있다.

$$\log\left(\frac{p_i}{1-p_i}\right) = \gamma_0 + \gamma_1 z_{1i} + \cdots + \gamma_p z_\pi,$$
$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_\pi. \tag{4.4}$$

식 (4.4)에서 윗부분은 이직경험 유무에 대한 로지스틱 부분이고 아래 부분은 이직 경험이 있는 대상자에 대해서 이직횟수에 대한 포아송 부분이다. 각 부분에 투입된 $z_{1i}, \cdots, z_\pi$와 $x_{1i}, \cdots, x_\pi$는 이직경험 유무 및 이직횟수를 설명하기 위한 설명변수들이다. 여기서 $\gamma_i$는 이직을 경험한 일자리 유무에 대한 로지스틱 회귀계수이고, $\beta_i$는 이직횟수에 대한 회귀계수를 나타낸다.

또한 포아송 회귀모형을 적합 시키는 과정에서 0이 과다하게 포함되고, 분산이 크게 되는 과대산포의 문제가 동시에 발생하는 경우 0을 포함하는 분포와 음이항 분포를 혼합해서 만든 영과잉 음이항 회귀모형을 사용한다. 독립변수 $X_1, X_2, \cdots, X_k$가 주어질 경우 종속변수 $Y_i$는 확률 $p_i$로 0의 값을 가지고, 확률 $1-p_i$로 음이항 분포를 따른다고 가정한다. 여기서 $p_i$ $(0 \le p_i < 1)$는 0에서 과잉확률을 의미하며, 이와 같은 혼합모형은 아래 식 (4.5)와 같이 나타내며 영과잉 음이항 분포 (ZINB : zero inflated negative binomial)라고 한다.

$$P(Y_i = y_i|X_1, \cdots, X_k) = \begin{cases} p_i + (1-p_i)(1+\alpha\mu_i)^{-\alpha^{-1}},\ y_i = 0 \\ (1-p_i)\dfrac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i+1)}(1+\alpha\mu_i)\left(\dfrac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{-\alpha^{-1}},\ y_i = 1, 2, \cdots \end{cases} \tag{4.5}$$

식 (4.5)에서 이직경험이 없는 경우와 이직경험이 있는 경우 설명변수는 식 (4.4)와 같은 방법으로 투입하게 된다.

한편 이직횟수에 대한 최적모형은 다음과 같은 절차에 의해서 결정한다. 먼저 ZINB 모형에서 과대산포의 존재여부에 대한 가설검정 $H_1 : \alpha = 0$ 대 $H_a : \alpha > 0$을 한다. 이 경우 귀무가설이 참이면 과대산포가 존재하지 않는 ZIP모형이 선택된다.

만약에 ZINB 모형 하에서 과대산포에 대한 귀무가설이 기각되면, ZINB 모형에서 영과잉의 존재여부에 대한 가설검정 $H_2 : p_i = 0$ 대 $H_a : p_i > 0$을 한다. 이 경우 귀무가설이 참이면 영과잉이 존재하지 않는 NB 모형이 선택된다.

다음으로 ZIP 모형에서 영과잉의 존재여부에 대한 가설검정 $H_3 : p_i = 0$ 대 $H_a : p_i > 0$을 한다. 이 경우 귀무가설이 참이면 영과잉이 존재하지 않는 포아송 모형이 선택된다. 마지막으로 NB 모형에서 과대산포의 존재여부에 대한 가설검정 $H_4 : \alpha = 0$ 대 $H_a : \alpha > 0$을 한다. 이 경우 귀무가설이 참이면 과대산포가 존재하지 않는 포아송 회귀모형이 선택된다. 여기서 사용한 검정통계량은 과대산포 존재여부에 대해서는 LRT (likelihood ratio test), 영과잉 존재여부에 대해서는 Vuoug 검정통계량을 사용하였다. 여기에 대한 자세한 내용은 Dean과 Lawless (1989), Ridout 등 (2001)을 참고하기 바란다.

다음 Table 4.1은 이직횟수에 대한 최적모형을 선택하기 위해 포아송 회귀모형 (P), 영과잉 포아송 회귀모형 (ZIP), 음이항 회귀모형 (NB), 영과잉 음이항 회귀모형 (ZINB) 등 4개의 회귀모형에 대한 분석결과를 제시한 것이다. 여기서 이직경험 유무에 대한 독립변수 $z_{1i}, \cdots, z_{\pi}$는 다중공선성의 문제와 모형의 간결성 원칙을 고려하여 참여여부, 연령 및 기업규모 변수만을 사용하였다.

**Table 4.1** Results of regression analysis for count data

| | | P | | ZIP | | NB | | ZINB | |
|---|---|---|---|---|---|---|---|---|---|
| | | B | s.e. | B | s.e. | B | s.e. | B | s.e. |
| no. of change jobs ($\hat{\beta}$) | | | | | | | | | |
| treatment | | -.12*** | .01 | -.19*** | .01 | -.13*** | .01 | -.14*** | .01 |
| male | | -.01* | .01 | .01* | .01 | -.01** | .01 | -.01* | .01 |
| age | 21-25 | .48*** | .01 | .46*** | .01 | .48*** | .01 | .35*** | .02 |
| | 26-30 | .49*** | .01 | .47*** | .01 | .48*** | .01 | .35*** | .02 |
| | >30 | .52*** | .01 | .50*** | .02 | .51*** | .02 | .37*** | .02 |
| area | Chungchung | .10*** | .01 | .10*** | .01 | .10*** | .01 | .10*** | .01 |
| | Junra | .07*** | .01 | .07*** | .01 | .07*** | .01 | .07*** | .01 |
| | Daekyung | .09*** | .01 | .10*** | .01 | .09*** | .01 | .10*** | .01 |
| | Dongnam | .10*** | .01 | .11*** | .01 | .10*** | .01 | .11*** | .01 |
| | Gangwon/Jeju | .03 | .02 | .03 | .02 | .02 | .02 | .03 | .02 |
| size | 10-49 | -.22*** | .01 | -.20*** | .01 | -.22*** | .01 | -.22*** | .01 |
| | 50-99 | -.29*** | .01 | -.26*** | .01 | -.29*** | .01 | -.29*** | .01 |
| | 100-299 | -.32*** | .01 | -.28*** | .01 | -.32*** | .01 | -.32*** | .01 |
| | >300 | -.34*** | .01 | -.26*** | .01 | -.34*** | .01 | -.34*** | .01 |
| job type | job2 | -.06*** | .01 | -.06*** | .01 | -.06*** | .01 | -.06*** | .01 |
| | job3 | .12*** | .01 | .12*** | .02 | .12*** | .02 | .12*** | .02 |
| | job4 | .02*** | .01 | .02*** | .01 | .02*** | .01 | .02*** | .01 |
| | job5 | .16*** | .01 | .16*** | .01 | .16*** | .01 | .16*** | .01 |
| constant | | -.54*** | .01 | -.37*** | .01 | -.54*** | .01 | -.40*** | .02 |
| inflate ($\hat{\gamma}$) | | | | | | | | | |
| treatment | | | | -.47*** | .05 | | | -.79*** | .17 |
| age | | | | -.01 | .01 | | | -.65*** | .05 |
| size | | | | .00*** | .00 | | | .00 | .00 |
| constant | | | | -1.48*** | .21 | | | -.40*** | .86 |
| $\widehat{\alpha}$ | | | | | | 0.258*** | | 0.243*** | |
| $H_1 : \alpha = 0$ under ZINB | | | | | | | | 1,932.66*** | |
| $H_2 : p_i = 0$ under ZINB | | | | | | | | 5.80*** | |
| $H_3 : p_i = 0$ under ZIP | | | | 21.97*** | | | | | |
| $H_4 : \alpha = 0$ under NB | | | | | | 4,253.21*** | | | |
| log-likelihood | | -296,525.490 | | -295,290.40 | | -294,398.890 | | -294,324.100 | |
| Wald (LR) $\chi^2$ | | 7,794.62*** | | 5,096.76*** | | 6,453.6*** | | 4,697.55*** | |

*, ** and *** mean significant with 0.1, 0.05 and 0.01 level, respectively.

위의 분석결과에 따르면 포아송 회귀모형에서는 영과잉 문제 ($H_3 : p_i = 0$ 하 ZIP)가 통계적으로 유의하게 존재한다는 것을 알 수 있다 (21.97***). 또한 과대산포를 나타내는 모수 추정치는 $\hat{\alpha}$=0.258이며, 통계적으로 유의하게 과대산포가 존재함을 알 수 있다 (4,253.21***). 음이항 회귀모형에서는 과대산포 ($H_4 : \alpha = 0$ 하 NB)는 $\hat{\alpha}$=0.243으로 추정되어 통계적으로 유의하며 (4,253.21***), 영과잉 ($H_2 : p_i = 0$ 하 ZINB) 또한 통계적으로 유의한 것으로 나타났다 (5.80***). 그리고 영과잉 음이항 회귀모형에서 과대산포의 존재 ($H_1 : \alpha = 0$ 하 ZINB)에 대한 검정결과 통계적으로 유의하게 과대산포가 존재함을 알 수 있다 (1,932.66***). 따라서 본 연구에서는 이직횟수를 적합 시킬 수 있는 최적의 회귀모형으로 영과잉 음이항 회귀모형이 선택되었다.

위의 Table 4.1에서 분석결과는 이직경험이 없는 경우와 이직경험이 있는 경우로 나누어 각각 회귀계수 $\hat{\gamma}$와 $\hat{\beta}$를 추정한 결과가 제시되어 있다. 영과잉 음이항 회귀모형을 중심으로 분석결과를 살펴보면 다

음과 같다.

먼저 영과잉 존재여부에 대한 설명변수의 추정결과 ($\widehat{\gamma}$)를 살펴보면, 기업규모를 제외한 독립변수들이 통계적으로 유의함을 알 수 있다. 통제집단에 비해서 처리집단이, 그리고 연령이 많을수록 영과잉이 아닐 가능성이 높은 것으로 나타났다. 이러한 결과는 Figure 2.2의 다중대응분석의 결과와 유사한 것을 알 수 있다.

한편, 이직경험이 있는 경우 회귀계수 추정치 ($\widehat{\beta}$)의 결과를 살펴보면, 대부분의 독립변수들이 이직횟수에 통계적으로 유의하게 영향을 미치는 것을 알 수 있다. 먼저 통제집단에 비해서 처리집단이, 그리고 여자에 비해서 남자의 이직횟수가 작게 나타났다. 그리고 20세 이하 집단에 비해서 연령이 증가할수록 이직횟수가 높게 나타났으며, 지역은 수도권에 비해서 충청권, 전라권, 대경권 및 동남권이 이직횟수가 높은 것으로 나타났다. 또한 기업규모가 클수록 이직횟수가 유의하게 감소하였으며, 직종은 직종1에 비해서 직종3, 직종4, 직종5는 통계적으로 유의하게 이직횟수가 높게 나타났다. 처리집단의 이직횟수가 통제집단에 비하여 적은 것으로 나타난 것은 중소기업 청년인턴제가 긍정적인 효과를 지니고 있는 것을 의미한다. 그리고 연령이 많은 사람의 이직횟수가 많은 이유는, 20세 미만은 고졸일 가능성이 많아 기대수준이 낮으므로 이직의 확률이 낮기 때문인 것으로 생각된다. 이에 비하여 20세 이상은 연령대별로 기대수준에 별 차이가 없어서 이직의 확률에 별 차이가 없기 때문인 것으로 생각된다. 또한 기업의 규모가 클수록 이직을 적게 하는 것은, 기업의 규모가 직장의 환경을 나타내는 대리변수로 볼 수 있는데 직장의 규모가 커지면 환경이 좋아지므로 이직의 확률이 낮아지기 때문인 것으로 판단된다. 이러한 결과는 Figure 2의 의사결정나무 분석의 결과와도 유사함을 알 수 있다.

## 5. 결론

본 연구에서는 이직횟수에 영향을 미치는 요인을 분석하기 위하여 청년인턴 DB와 고용보험 DB를 활용하여 다중대응분석, 의사결정나무분석 및 회귀분석을 실시하였다. 특히 계수 데이터인 이직횟수에 대한 회귀모형으로 포아송 회귀모형, 영과잉 포아송 회귀모형, 음이항 회귀모형, 영과잉 음이항 회귀모형 등 4개의 회귀모형을 적용하였다. 최적모형으로 영과잉 음이항 회귀모형이 선택되었으며 주요 분석결과를 정리하면 다음과 같다.

첫째, 청년인턴 미참여자에 비해서 참여자들이 통계적으로 유의하게 이직경험이 낮게 나타났다. 이는 정부의 중소기업 청년인턴 제도가 중소기업의 인력난을 해소하려는 목적을 일정 부분 달성하고 있다는 것을 의미한다.

둘째, 여자에 비해서 남자가, 20세 이하의 집단에 비해서 연령이 높아질수록 유의하게 이직횟수가 높게 나타났다.

셋째, 수도권에 비해서 충청권, 전라권, 대경권 및 동남권이 유의하게 이직횟수가 높은 것으로 나타났다.

마지막으로 기업규모가 클수록 이직횟수가 유의하게 감소하였으며, 직종별로도 이직횟수에 유의한 차이가 존재하는 것으로 나타났다.

한편 본 연구에서는 이직횟수에 대한 결정요인을 분석함에 있어서 수집이 가능한 제한된 독립변수만을 사용했다는 점에서, 본 연구의 결과를 지나치게 일반화하는 것은 무리가 있음을 지적하고자 한다. 따라서 더 많은 독립변수를 활용한 이직횟수에 대한 결정요인 분석은 향후 과제로 남겨둔다.

## References

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*, Wadsworth, Belmont.

Cho, J. S. (2010). A study on equating method based on regression analysis. *Journal of the Korean Data & Information Science Society*, **21**, 513-521.

Cho, J. S. (2012). Inflow and outflow analysis of double majors using social network analysis. *Journal of the Korean Data & Information Science Society*, **23**, 693-701.

Cho, J. S. (2014). A study on determinant factor for number of stop-out. *Journal of the Korean Data Analysis Society*, **16**, 201-210.

Cho, K. H. and Park, H. C. (2012). A study on decision tree creation using marginally conditional variables. *Journal of the Korean Data & Information Science Society*, **23**, 299-307.

Choi, J. H., Ko, I. M. and Cheon, S. Y. (2011). A zero-inflated model for insurance data. *The Korean Journal of Applied Statistics*, **24**, 485-494.

Chun, H. J. and Kim, H. I. (2008). A Study on SMS Expansion Using Negative Binomial Regression. *Journal of the Korean Data Analysis Society*, **10**, 2065-2074.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika*, **70**, 269-274.

Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, **84**, 467-472.

Jeong, J. P. and Choi, J. H. (2014). Poisson regression and negative binomial regression model fit for traffic accidents. *Journal of the Korean Data Analysis Society*, **16**, 165-172.

Joo, M. H., Kim, D. S., Kim, B. W., Choi, J. I. and Kim, J. H. (2013). *The study on the employment effect of the youth-intern project*, Korea Employment Information Service.

Jung, H. J. and Min, D. K. (2013). The study of foreign exchange trading revenue model using decision tree and gradient boosting. *Journal of the Korean Data & Information Science Society*, **24**, 161-170.

Kim, K. M. (2005). A study on the zero-inflated Poisson regression model. *Journal of the Korean Data Analysis Society*, **7**, 497-505.

Kim, K. M. and Eum, H. J. (2010). Comparisons of the estimators for the zero-inflated Poisson distribution. *Journal of the Korean Data Analysis Society*, **12**, 1113-1124.

Lee, K. Y. (2009). *The evaluation of the youth-intern project*, The Monthly Labor Review, The Korea Labor Institute.

Lee, J. K. (2011). The Effect of the Economic Background Risk on Fertility Decision: Using Censored and Zero Inflated Count Data Regression. *Journal of the Korean Data Analysis Society*, **13**, 1521-1531.

Lee, D. H. (2013). Simulated maximum likelihood estimator for the zero inflated Poisson lognormal regression model. *Journal of the Korean Data Analysis Society*, **15**, 1347-1360.

Ridout, M., Hinde, J. and Demetrio, C. G. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**, 219-223.

Ryu, J. S., Park, S. I., Cho, J. S., Kim, J. H., Ha, B. C. and Kim, J. H. (2012). *The current state and the employment effect of the youth-intern project*, Human Resource Development Institute at Pukyung National University.

Ryu, J. S., Park, S. I., Cho, J. S. and Jung, H. J. (2014). *The performance evaluation of the youth-intern project*, Human Resource Development Institute at Pukyung National University.

*Journal of the Korean Data & Information Science Society*
*2015,* **26**(2), 387–397

http://dx.doi.org/10.7465/jkdi.2015.26.2.387
한국데이터정보과학회지

# The study on the determinants of the number of job changes

Sungik Park[1] · Jangsoo Ryu[2] · Jonghan Kim[3] · Jangsik Cho[4]

[1]International Trade and Commerce, Kyungsung University
[2]Division of Economics, Pukyong National University
[3]Division of Economics, Finace and Logistics, Kyungsung University
[4]Department of Informational Statistics, Kyungsung University

Received 23 February 2015, revised 8 March 2015, accepted 18 March 2015

**Abstract**

In this paper, the determinants of the number of job changes in the SMEs (small and medium enterprises) youth-intern project is analysed, utilizing SMEs youth-intern DB and employment insurance DB. Since the number of job changes are count data which take integer values other than negative values, general linear regression analysis becomes inappropriate. Therefore, four models such as Poisson regression model, zero inflated Poisson regression model, negative binomial regression model and zero inflated negative binomial regression model are tried to fit count data. A zero inflated negative binomial regression model is selected to be the best model. Major results are the followings. First, the number of job changes is shown to be significantly smaller in the treatment group than in the control group. Second, the number of job changes turns out to be significantly smaller in the young-age group than in the old-age group. Third, it is also shown that the number of job changes of man is significantly greater than that of woman. Lastly, the number of job changes in the bigger firm is shown to be significantly less than that of the smaller firm.

*Keywords*: Multiple correspondence analysis, negative binomial regression model, over-dispersion, Poisson regression model, zero inflation.

[1] Professor, Department of International Trade and Commerce, Kyungsung University, Busan 608-736, Korea.
[2] Professor, Division of Economics, Pukyong National University Busan 608-737, Korea.
[3] Professor, Division of Economics, Finace and Logistics, Kyungsung University, Busan 608-736, Korea.
[4] Corresponding author: Professor, Department of Informational Statistics, Kyungsung University, Busan 608-736, Korea. E-mail : jscho@ks.ac.kr