계량경제학 데이터 분석 과제

- Problem set 8 -

이름 : 김겨레
학번 : 2019314908

0. 분석 개요

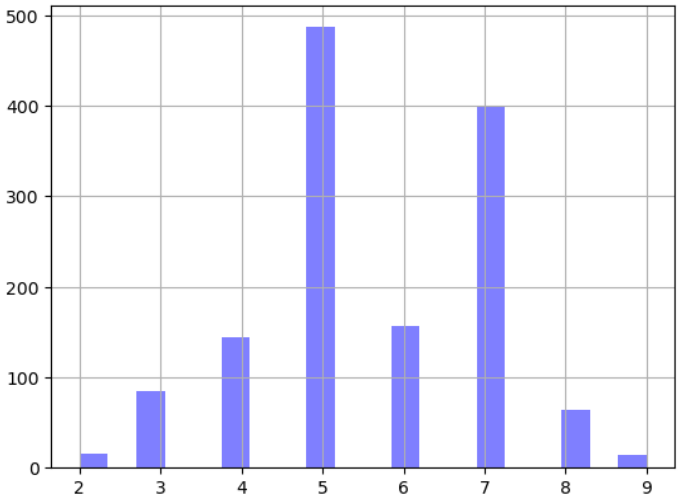
본 데이터 분석의 목적은 교육의 가치에 대해 평가하는 것이다. 시장에서 가치는 가격에 의해 평가된다. 교육은 일반적으로 시장에서 가치를 창출하는 데 필요한 능력을 기르기 위한 목적을 가진다. 그러므로 그 가치를 평가할 수 있는 가격 지표는 '창출한 부가가치의 정도', 즉 소득이 될 수 있다. 하지만 본 데이터 분석에서는 노동시장이 아닌 결혼시장에 주목하여 소득이 아닌 다른 변수를 종속변수로 설정하였다. 일정 이상의 나이에서 결혼의 경험이 있다는 것은, 결혼시장에서 상대 성별로부터 그 가치를 인정을 받은 경험이 있다는 의미이다. 그리고 교육 수준은 결혼시장에서 일어나는 가치평가에서 중요한 변수로 작용할 수 있다. 따라서 교육의 가치를 노동시장 측면을 넘어서 결혼시장의 관점에서 분석하는 것은 충분히 타당성을 가질 수 있다.

본 데이터 분석에는 다양한 통제변수를 고려해 교육이 결혼에 미치는 영향을 분석할 것이다. 또한, 세대(연령)와 성별 등에 따라 영향이 다르게 나타날 수 있음을 고려하여 상호작용항 등을 설정해 세대와 성별에 따라 교육의 효과가 어떻게 나타나는지까지 확인할 것이다.

1. 데이터 설명과 변수 설정

본 연구에서 사용할 데이터는 'ID-8', 서울 지역의 데이터이다. 데이터는 1366명의 정보를 담고 있다. 이중 532명, 즉 38.95%가 여성이다. 또한, 309명, 22.6%가 고용인이고, 절반 이상인 738명, 54.03%가 상용직이다. 교육 연수가 구간별로 제시되어 정확한 평균치는 알 수 없지만, 구간값의 평균치가 5.387994인 것으로 보아 대략 13-14년이 평균적인 교육 연수인 것으로 보인다. 아래 도표는 교육 수준에 따른 인구 분포를 보여주는데, 고등학교(5)와 4년제 대학(7)에 집중되어 있다. 초등학교, 중학교, 고등학교를 합쳐 12년, 대학교 4년을 고려할 때, 아래와 같은 분포가 13-14년의 평균적인 교육 기간을 설명해준다.

<교육 수준(School) 별 인구분포>



<성별에 따라 소득 기초통계량>

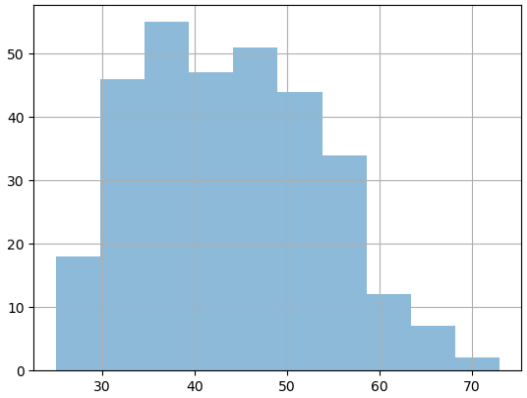
성별	평균	표준편차
여성	7273.372218	74839.693508
남성	12509.085132	97268.437362

결혼 여부 및 교육 수준을 나타내는 변수를 설정하기 위해 기존 데이터에서 “The marital Status of individuals”에서 미혼일 경우 0, 미혼을 제외한 나머지 값을 가질 경우 1의 값을 가지는 ‘결혼 경험 유무 더미변수’를 만들었으며, 4년제 대학 이상(졸업 여부 미포함)이면 1, 아니면 0의 값을 갖는 ‘4년제 대학 이상 여부 더미변수’를 만들었다. 아래 표는 결혼 여부 더미변수를 통해 확인한 기초통계량이다. “The marital Status of individuals”로 에서 나타난 “married” 여성의 평균 소득은 11208.1867로 다른 여성 그룹보다 높았다. 반면, 남성 그룹에서는 미혼의 경우 평균 소득이 13511.0867로 가장 높았고, 그 다음이 ‘married’ 그룹이었다. 여성의 경우, “married” 그룹이 평균 43.76세, 미혼 그룹이 28.53세로, 연령 효과가 클 가능성이 높다.

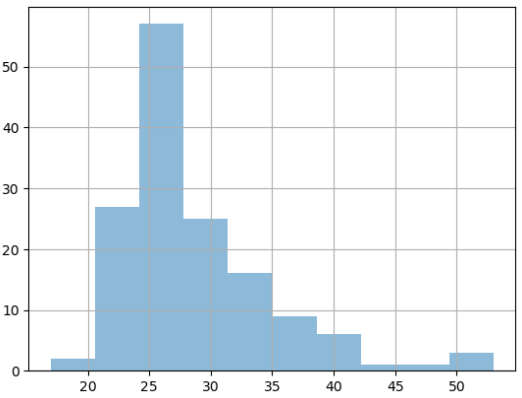
<결혼 여부와 성별에 따른 소득 기초통계량>

성별	여성		남성	
결혼 경험 여부	유	무	유	무
평균	9432.3792	1618.8299	12246.8366	13511.0867
표준편차	87908.0881	972.2331	107003.5063	94640.9549

<’married’ 여성 그룹의 연령분포>

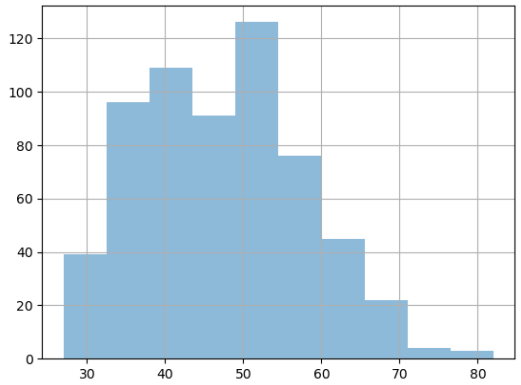


<미혼 여성의 연령 분포>

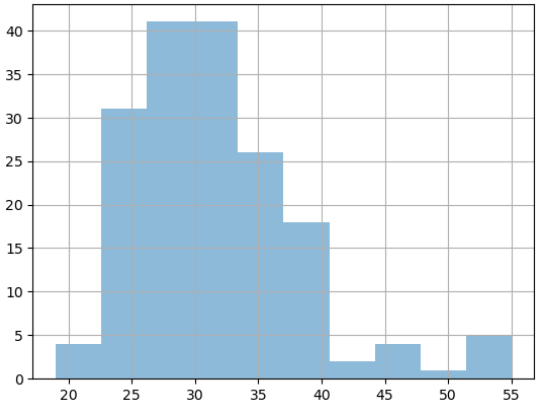


남성 집단의 경우 ‘결혼 여부와 성별에 따른 소득 기초통계량’에서, 미혼 남성이 더 높은 소득을 올리고 있는 것으로 보인다. 하지만 ‘married’ 남성 그룹과 미혼 남성 그룹의 연령분포를 비교하면, 전자에서 60세 이상 인구가 많다. 또한, 60세 이상 남성 인구의 평균연봉은 1426.3919인데 반해, 나머지 집단의 경우 13588,1895이다. 따라서 결혼 여부에 따른 남성 인구 집단 내 평균소득의 차이는 연령 효과에 기인한 것으로 보는 것이 타당하다.

<’married’ 남성 그룹의 연령분포>



<미혼 남성의 연령 분포>



아래는 결혼 경험 여부와 4년제 대학 이상(졸업 여부 미포함)에 따른 인구를 나타낸 표이다. 결혼 연령이 평균적으로 30세 안팎임을 감안하여 30세 이상 그룹을 대상으로도 분포를 확인했다. 4년제 미만에서 결혼 경험의 비율이 높은 것을 확인해볼 수 있다. 또한, 교육 연수와 결혼 여부(더미변수)의 상관계수를 계산해보았을 때, -0.13의 낮은 수준이지만 음의 상관관계가 나왔다. 하지만 이런 분포 형성에는 교육 변수뿐 아니라 성별, 소득 등 다양한 요인이 여기에 개입했을 수 있기에 계량경제학 모형을 통해 추가적인 분석이 필요하다.

<결혼 경험 여부와 4년제 대학 이상>

	4년제 이상	**4**년제 미만
결혼 무	143	177
결혼 유	334	712
30세 이상 그룹		
	4년제 이상	**4**년제 미만
결혼 무	57	68
결혼 유	312	695

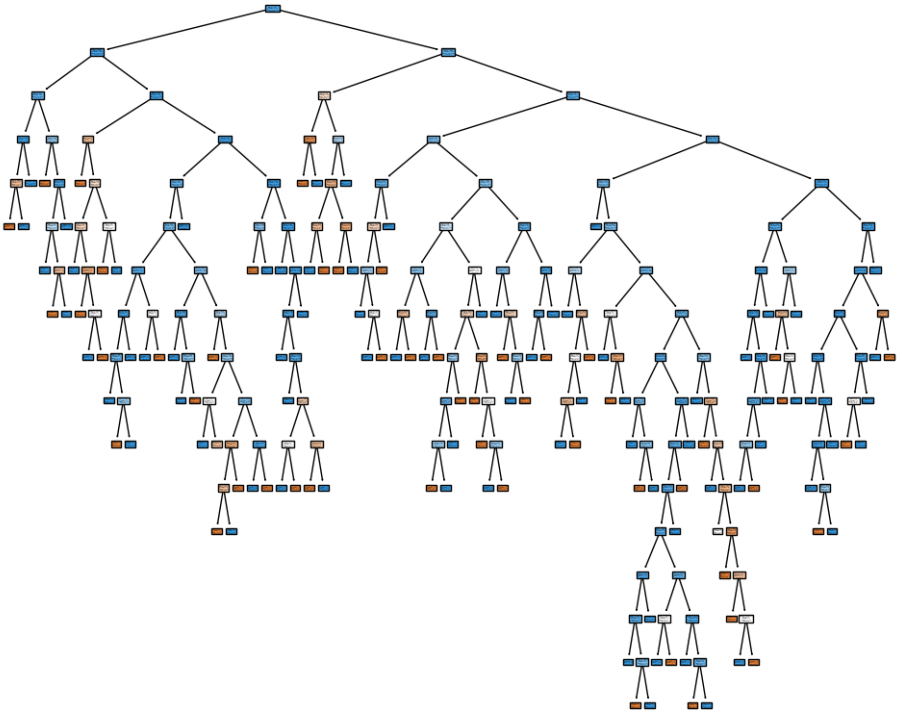
본 데이터 분석에서는 종속변수로 ‘결혼 경험 여부’ 변수를 선택하였다. 교육 요인을 나타내는 독립변수로는, 연속적인 효과를 효과적으로 파악하기 위해 교육연수를 선택하였다. 위에 나타난 것과 같이 나이-소득 변수는 결혼 여부와 연관이 있으므로 통제변수로 선택하였고, 성별에 따라 해당 효과가 다르게 나타나는 나타나는 부분을 통제하기 위해 성별까지 변수로 채택했다. 또한, 30세 이상 그룹을 대상으로 분석을 진행하였다.

2. 계량경제학 모델

본 데이터 분석에서 활용할 종속 변수는 이진변수이다. 따라서 단순 선형회귀모형이 아닌 확률값을 고려한 다른 비선형 모형을 선택하는 것이 바람직하다. 검토한 모형은 Support Vector Machine, Logistic Regression, Random Forest, Naive Bayes이다. 네 가지 모형으로 Cross Validation을 진행한 결과, 네 가지 모형 모두 평균적으로 0.9 수준의 유사한 예측력을 보여주었다. 예측 모형으로 네 가지 모두 적합함을 보여준다. 하지만 Naive Bayes는

상호작용을 고려한 분석에 적합하지 않았고, Support Vector Machine은 Hyper parameter 설정 문제가 있으며, Random Forest는 상호작용을 모델 자체적으로 반영하지만, 아래 그림과 같이 복잡성으로 인해 해석에 문제가 따른다. 결론적으로, Logistic Regression 모형을 활용하였다.

<Random Forest 모형 결과 도출한 Tree>



교육연수과 다른 변수 간 상호작용을 효과적으로 고려하기 위해 Logistic Regression 모형은 아래와 같이 설정하였다.

$$P(marriage = 1) = \beta_0 + \beta_1 ysrsshl + \beta_2 age + \beta_3 wage$$
$$+ \beta_4 age * ysrsshl + \beta_4 sex * ysrsshl + \beta_5 wage * ysrsshl$$

모형 분석에 들어가기에 앞서 교육연수와 소득의 상관성으로 인해 다중공선성 문제가 있을 수 있으므로, VIF 값을 확인했다. 모두 5 이하로, 다중공선성을 신경쓰지 않아도 될 정도로 VIF 값이 작게 나왔기 때문에 분석을 그대로 진행하였다.

Variables	VIF
Age	4.3977
Sex	2.9973
Wage	1.0177
Ysrsshl	4.2919

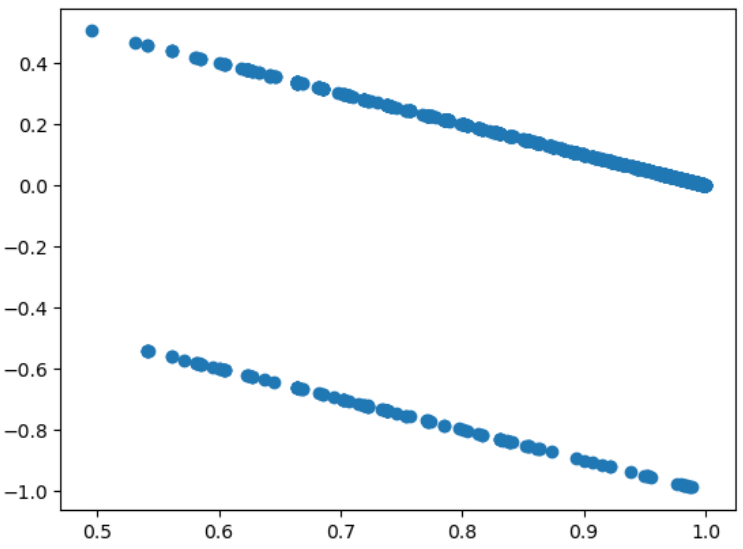
3. 분석 결과

Variable	Coef	Std Err	Z-score	P-value	95% CI (Lower)	95% CI (Upper)
Intercept	-6.1768	1.921	-3.216	**0.001*** **	-9.941	-2.413
age	0.1990	0.046	4.335	**0.000*** **	0.109	0.289
ysrsshl	0.1267	0.313	0.404	0.686	-0.487	0.741
age:ysrsshl	-0.0030	0.008	-0.381	0.703	-0.019	0.013
wage	0.0007	0.000	3.367	**0.001*** **	0.000	0.001
wage:ysrsshl	-9.256e-05	2.74e-05	-3.374	**0.001*** **	-0.000	-3.88e-05
sex	-2.2901	0.703	-3.258	**0.001*** **	-3.668	-0.912
sex:ysrsshl	0.3250	0.109	2.969	**0.003*** **	0.110	0.540
Pseudo R-square : 0.2330 Log-Likelihood : -301.63 Method : MLE						

Pseudo R-square 값은 0.2330으로 낮은 수준이다. 하지만 8개의 변수 중 6개의의 변수가 통계적으로 유의하고, 0.1 이상의 R-square 값을 가진다. 또한, 모형의 log-likelihood 값은 -301.63으로 작은 값을 보인다. 따라서 본 모형은 받아들일 수 있는 모형이다. 모형의 변수에서 독립변수인 교육연수 항은 5% 수준에서 통계적으로 유의하지 않은 것으로 나타났다. 하지만 소득 및 성별과 교육연수의 상호작용항은 통계적으로 유의한 것으로 나타났다.

추정 결과, 성별이 남성일 때, 교육연수의 계수값이 0.3250만큼 증가한다. 즉, 남성의 경우 결혼시장에서 교육의 가치가 더 높게 나타나는 것이다. 소득이 증가할 때는, 교육연수의 계수가 작아지는 것을 확인할 수 있다. 반대로 해석하면, 현재의 소득값이 낮아도 교육의 가치가 결혼시장에서 일정수준 인정 받고 있다는다는 것을 의미한다. 결론적으로, 결혼시장에서 소득은 그 자체로 높은 가치를 인정받지만, 교육은 소득과 달리, 특정 성별에 있어서 높은 가치로 평가받는다.

<Residual의 분포>



하지만 데이터 분석에 활용한 해당 모형은 여러 한계를 가지고 있다. 먼저, 이분산성이 발견되었다. 위 그래프는 Residual의 분포를 보여준다. 우하향 하는 직선의 형태이다. 해당 방향으로 이분산성이 있음을 시사한다. 또한, 직장의 안정성을 보여주는 '직종' 등의 변수가 포함이 되지 않았다는 점에서 변수 설정에서도 한계가 있다. 하지만 그럼에도 교육의 가치가 소득과 성별 집단에 따라 다르게 평가될 수 있다는 결과를 제시했다는 점에서 의미가 있다.