

Finance Analytics

Chapter 4. 질적 설명변수를 포함한 회귀모형

권태연

한국외대 국제금융학과

질적 설명변수 (설명변수가 범주형 변수인 경우)

질적설명변수

- 3장의 선형회귀모형: Y변수-정량, X변수-정량인 경우만
- X변수 중 범주형 변수 (질적변수)가 포함된다면?
- 더미변수(dummy variable)을 만들어 질적 변수를 "수치화"하여 정량 변수처럼 이용한다.
- 예: 성별
 - 성별변수 : 범주형, 남/여
 - 성별에대한 더미변수 설정 : Female
 - Female=1, if 성별=여성
 - Female=0, if 성별=남성

임금예제

- WAGE: 시간당임금
- Female: 성별 (Female=1은 여성, Female=0은 남성)
- nonwhite: 인종 (nonwhite=1은 백인이 아님, nonwhite=0은 백인)
- union: 노조가입여부 (Union=1은 노조가입, Union=0은 비노조)

$$Wage_i = B_0 + B_1 Female_i + B_2 nonwhite_i + B_3 Union_i + B_4 Educ_i + B_5 Exper_i + \epsilon_i$$

lm() 추정결과를 이용하여 위의 모수모형을 추정한 결과는

$$\hat{W}age_i = -7 - 3Female_i - 1.6nonwhite_i + 1.1Union_i + 1.37Educ_i + 0.16Exper_i$$

회귀 계수의 해석?

더미변수의 해석

$$\hat{Wage}_i = -7 - 3Female_i - 1.6nonwhite_i + 1.1Union_i + 1.37Educ_i + 0.16Exper_i$$

- Female=1은 여성, Female=0은 남성
- 여성의 시간당 임금에 대한 선형회귀모형

$$\hat{Wage}_i = -7 - 3 - 1.6nonwhite_i + 1.1Union_i + 1.37Educ_i + 0.16Exper_i$$

- 남성의 시간당 임금에 대한 선형회귀모형

$$\hat{Wage}_i = -7 - 1.6nonwhite_i + 1.1Union_i + 1.37Educ_i + 0.16Exper_i$$

더미변수의 해석

$$\hat{Wage}_i = -7 - 3Female_i - 1.6nonwhite_i + 1.1Union_i + 1.37Educ_i + 0.16Exper_i$$

- -3은 여성의 시간당 임금의 선형회귀모형의 절편과 남성 모형의 절편 차이
- 여성노동자는 Control된 다른 모든 조건들(인종, 노조가입여부, 경력, 교육기간)이 모두 같은 남성에 비하여 시간당 임금이 3\$ 적다. (P-value<0.05이므로 이 차이는 0과 다른 매우 유의미한 차이로 하겠다.)
- *nonwhite*, *Union*의 회귀계수에 대한 해석?

더미변수의 해석

$$\hat{Wage}_i = -7 - 3Female_i - 1.6nonwhite_i + 1.1Union_i + 1.37Educ_i + 0.16Exper_i$$

- 백인이 아닌 여성과 백인남성의 평균 시간당 임금 차이에 대한 해석?

범주의 갯수가 3개일 때는?

- 범주형 변수의 범주의 갯수가 2개인 경우에는 더미변수 1개를 만든다.
- 범주형 변수의 범주의 갯수가 3개인 경우에는?
- (예) 결혼 여부
 - 범주 1: 미혼
 - 범주 2: 기혼
 - 범주 3: Divorced or Widowed.
- 더미변수를 2개 만든다.
 - 미혼이면 $D1=0$, $D2=0$
 - 기혼이면 $D2=1$, $D2=0$
 - Divorced or Widowed 이면 $D1=0$, $D2=1$

범주의 갯수가 3개일 때는?

$$\begin{aligned}\hat{Wage}_i = & -7 - 3Female_i - 1.6nonwhite_i + 1.1Union_{4i} + 2D_{1i} - 0.5D_{2i} \\ & + 1.37Educ_i + 0.16Exper_i\end{aligned}$$

- 미혼 노동자의 시간당 임금에 대한 선형회귀모형?
- 기혼 노동자의 시간당 임금에 대한 선형회귀모형?
- Divorced or Widowed 노동자의 시간당 임금에 대한 선형회귀모형?
- $D1$ 의 회귀계수의 의미?
- $D2$ 의 회귀계수의 의미?

범주의 갯수가 m 개인 경우

- $m-1$ 개의 더미변수를 만든다.
- m 개의 더미변수를 만든다면 가정6 다중공선성 문제가 발생하게 된다.

범주형 변수값에 따른 정량변수의 기울기 차이

- Female에 대한 회귀계수는 남성과 여성 노동자의 임금에 관한 회귀모형의 절편의 차이
다른조건이 동일할때 성별에 따른 시간당 평균임금의 차이
- 학업기간이 1년 증가한다면 남성 여성 노동자 모두 1.37달러의 시간당 임금이 평균적으로 증가.
- 만약, 남성의 학업기간 1년증가와 여성의 학업기간 1년증가의 시간당 임금에 미치는 영향이 다르다고 의심된다면?
- $Female \times Educ$ 변수를 생성하여 회귀모형에 추가한다.
- 이때 $Female \times Educ$ 를 interaction (교호작용, 교호효과) term라 한다.

범주형 변수값에 따른 정량변수의 기울기 차이

$$\begin{aligned} Wage_i = & B_1 + B_2 Female_i + B_3 nonwhite_i + B_4 Union_i + B_5 Educ_i \\ & + B_6 Exper_i + B_7 Educ_i * Female_i + u_i \end{aligned}$$

- 여성노동자의 시간당 임금에 대한 회귀모형:

$$\begin{aligned} Wage_i = & B_1 + B_2 + B_3 nonwhite_i + B_4 Union_i + (B_5 + B_7) Educ_i + B_6 Exper_i + u_i \end{aligned}$$

- 남성 노동자의 시간당 임금에 대한 회귀모형:

$$Wage_i = B_1 + B_3 nonwhite_i + B_4 Union_i + B_5 Educ_i + B_6 Exper_i + u_i$$

- B2에 대한 해석 및 B2계수의 추정치가 유의하다면?
- B7에 대한 해석 B7계수의 추정치가 유의하다면?
- R을 이용하여 위의 모형을 적합하고 해석하여 보자.

두 범주형 변수의 교호작용

- 성별과 인종의 교호효과 $Female \times nonwhite$ 를 모형에 반영하였다고 하자.

$$Wage_i = B_1 + B_2 Female_i + B_3 nonwhite_i + B_4 Union_i + B_5 Educ_i + B_6 Exper_i + B_7 Female_i * nonwhite_i + u_i$$

- 유색인 여성노동자의 시간당 임금에 대한 회귀모형.
- 백인 여성노동자의 시간당 임금에 대한 회귀모형.
- 유색인 남성노동자의 시간당 임금에 대한 회귀모형.
- 백인 여성노동자의 시간당 임금에 대한 회귀모형.
- $Female \times nonwhite$ 를 추가함으로써 인종과 성별의 모든 조합별로 비교할 수 있다. 즉 성별과 인종이 모두 다른 그룹도 비교가능하게 할 수 있다.

