제 9 강

더미 변수
Dummy Variables

개요

- 다중회귀모형의 첫번째 가정 (MR1):

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + \varepsilon_t,\ t = 1,\ldots,T$$

  - 모형의 모수들, β_k 는 각각의 관측치에 대해 동일함을 의미하기도 함

$$\frac{\Delta E(y_t)}{\Delta x_{1k}} \text{ (other variables held constant) } \cdots = \frac{\partial E(y_t)}{\partial x_{Tk}}$$

- 다중회귀모형에서 모수들이 표본의 일부분에 대해 다른 경우를 다룰 수 있으려면?

개요

더미변수(모의변수, 이원변수)

- "1" 과 "0" 과 같은 값을 갖는 변수들
- 변수들이 이원화(dichotomized)되어 있음을 나타냄
  "presence" or "absence", "yes" or "no", etc.
- 변수들이 질(quality)적인 변수 혹은 특성(attribute)을 나타내는 변수
  "male" or "female",
  "black" or "white",
  "urban" or non-urban"
  "before" or "after"
- 그러한 질 혹은 특성은 여러 범주를 가질 수 있음
  "<10" or "10≤, ≤20" or "20 <"
  "North" or "south" or "east" or "west"
  ⇒ 몇 개의 더미변수를 이용하여 나타낼 수 있음

절편 더미변수

$$y_t = \beta_1 + \beta_2 X_t + \beta_3 G_t + \varepsilon_t$$

For male: $G_t = 1$.
For female: $G_t = 0$.

y_t = wage rate per hour

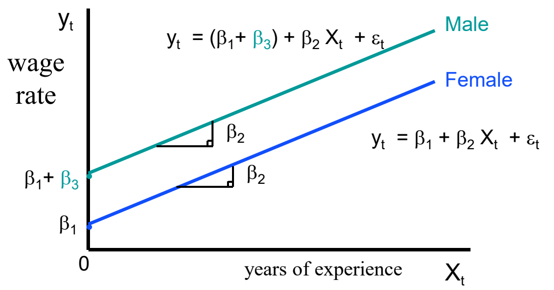
X_t = years of experience

- 초임에 있어서 여성 노동자들에 대한 차별여부를 검정 — H_0: $\beta_3 = 0$ H_1: $\beta_3 > 0$
- 초임에 있어서 여성과 남성간에 차이가 있는가를 검정 — H_0: $\beta_3 = 0$ H_1: $\beta_3 \neq 0$

절편 더미변수

$y_t = (\beta_1 + \beta_3) + \beta_2 X_t + \varepsilon_t$: Male Workers

$y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$: Female Workers



기울기 더미변수

$$y_t = \beta_1 + \beta_2 X_t + \beta_4 G_t X_t + \varepsilon_t$$

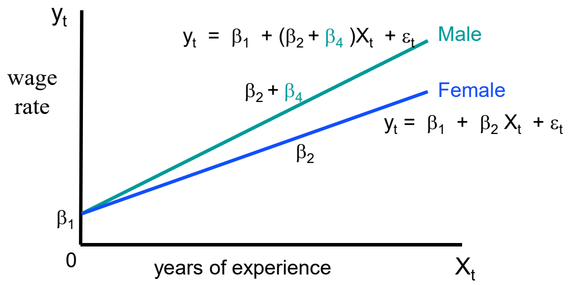
For male: $G_t = 1$.
For female: $G_t = 0$.

- 남성 여성 모두 동일한 초임 β_1 을 갖지만 그들의 경력에 따른 임금률은 다른 비율로 증가함 (차이 = β_4).
- $\beta_4 > 0$ 은 남성의 임금률이 여성의 임금률에 비해 빠르게 증가함을 의미함

기울기 더미변수

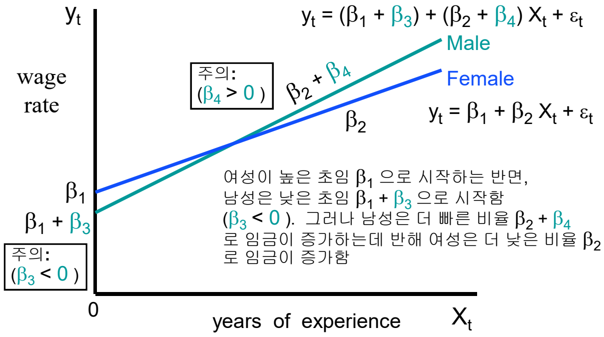
$y_t = \beta_1 + (\beta_2 + \beta_4)X_t + \varepsilon_t$: Male Workers

$y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$: Female Workers



기울기 및 절편의 차이

$y_t = \beta_1 + \beta_2 X_t + \beta_3 G_t + \beta_4 G_t X_t + \varepsilon_t$



더미변수 trap

- 만약 한 질적 변수의 두 범주를 확인하기 위해 두 개의 더미변수를 다음과 같이 도입을 하는 경우

$Y_t = \beta_1 + \beta^*_1 D1_t + \beta^{**}_1 D2_t + \beta_2 X_t + \varepsilon_i$

where $D1_t = 1$ if female
$= 0$ otherwise

where $D2_t = 1$ if male
$= 0$ otherwise

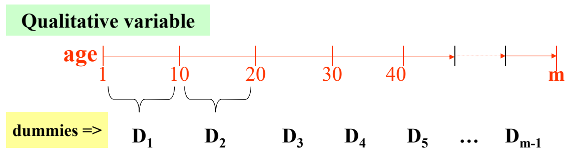
- 이 모형은 D1간의 D2가 완전한 공선성으로 인해 추정될 수 없음

∵ D1 = 1 - D2
or D2 = 1 - D1
or D1 + D2 = 1 (상수항) (Perfect collinearity)

더미변수 trap

- 일반적 원칙 : 완전한 (다중)공선성을 피하기 위해서는

1 직절한 변수가 **"m"** 개의 범주를 가지는 경우, 오직 **"m-1"**개의 더미변수들을 도입해야 함



더미변수 trap

2 어떤 범주에 대해 0의 값이 부여되는 경우, 이 범주는 통제된 범주(controlled category), 혹은 누락집단, 참조집단(reference group)이라고 함

- 더미 변수들의 모수들은 참조집단 대비(relative to) 종속변수 수준의 기대되는 차이를 나타냄

$Y_t = \beta_1 + \beta^*_1 D1_t + \beta_2 X_t + \varepsilon_i$

β^*_1은 남자 대비 여성의 임금의 기대되는 차이를 나타냄

더미변수의 활용

질적인 요소들간의 상호작용

- 둘 혹은 그 이상의 질적 변수들을 도입하는 경우 그들간의 상호작용에 관심을 기울여야 할 수도 있음
  - 성별 뿐 아니라 인종을 임금 방정식에 고려하는 경우

For White: $R_t = 1$.	For non-White: $R_t = 0$.
For Male: $G_t = 1$.	For Female: $G_t = 0$.

  - 단순히 임금방정식에 성별 더미와 인종더비를 포함하는 것으로는 이들 질적 요소들간의 상호작용을 파악할 수 없음

더미변수의 활용

질적인 요소들간의 상호작용

– "white" 이고 "male" 인 경우에 대한 특별한 취급은 개별적 인종과 성별 더미를 통해 파악되지 않음.

상호작용이 없음:
성별 임금격차는 인종에 의존하지 않음

$$y_t = \beta_1 + \beta_2 X_t + \delta_1 R_t + \delta_2 G_t + \varepsilon_t$$

상호작용이 존재:
성별 임금격차가 인종에 의존함

$$y_t = \beta_1 + \beta_2 X_t + \delta_2 R_t + \delta_1 G_t + \gamma R_t G_t + \varepsilon_t$$

더미변수의 활용

질적인 요소들간의 상호작용

$$E(Y_t) = \begin{cases} (\beta_1 + \delta_1 + \delta_2 + \gamma) + \beta_2 X_t & white-male \\ (\beta_1 + \delta_1) + \beta_2 X_t & white-female \\ (\beta_1 + \delta_2) + \beta_2 X_t & nonwhite-male \\ \beta_1 + \beta_2 X_t & nonwhite-female \end{cases}$$

δ_1 : 인종의 효과를 측정

δ_2 : 성별의 효과를 측정

γ : "white" 이고 "male." 인 효과를 측정

더미변수의 활용

여러 가지 범주를 갖는 질적 변수

- 많은 질적 요소들은 두 개 이상의 범주들을 가짐.
  - 예를 들면, 국가의 지역(북, 남, 동, 서), 교육수준(고졸 미만, 고졸, 대졸, 대학원졸 이상)
  - 각 범주에 대해서 개별적인 더미 변수를 만들 수 있음

더미변수의 일반적인 활용

여러 가지 범주를 갖는 질적 변수

– 예: 교육수준에 대한 더미 변수를 다음과 같이 정의

$$E_0 = \begin{cases} 1 & \text{less than high school} \\ 0 & \text{otherwise} \end{cases} \qquad E_1 = \begin{cases} 1 & \text{high school diploma} \\ 0 & \text{otherwise} \end{cases}$$

$$E_2 = \begin{cases} 1 & \text{college degree} \\ 0 & \text{otherwise} \end{cases} \qquad E_3 = \begin{cases} 1 & \text{postgraduate degree} \\ 0 & \text{otherwise} \end{cases}$$

$$WAGE = \beta_1 + \beta_2 EXP + \delta_1 E_1 + \delta_2 E_2 + \delta_3 E_3 + \varepsilon$$

더미변수의 일반적인 활용

여러 가지 범주를 갖는 질적 변수

$$E(WAGE) = \begin{cases} (\beta_1 + \delta_3) + \beta_2 EXP & \text{postgraduate degee} \\ (\beta_1 + \delta_2) + \beta_2 EXP & \text{college degree} \\ (\beta_1 + \delta_1) + \beta_2 EXP & \text{high school diploma} \\ \beta_1 + \beta_2 EXP & \text{less than high school} \end{cases}$$

- 더미 변수들의 모수들은 참조집단(여기서는 고졸미만 집단) 대비 기대되는 임금 격차들을 나타냄

더미변수의 일반적인 활용

시간에 따른 변화를 통제

– 월별 더미
  - 해당 월이 8월인 경우 AUG=1, 그렇지 않은 경우 AUG=0
  - JAN, FEB,…등도 마찬가지로 만들 수 있음

– 계절 더미(Seasonal Dummies)

– 연도별 더미(Annual Dummies)

– 체제 변화(Regime Changes, Structural changes)

$$ITC = \begin{cases} 1 & 1962-1965, 1970-1986 : \text{Investment Tax Credit} \\ 0 & otherwise \end{cases}$$

$$INV_t = \beta_1 + \delta ITC_t + \beta_2 GNP_t + \beta_3 GNP_{t-1} + \varepsilon_t$$

개별적 질적 효과에 대한 검정

계량경제학 9.19

men: $D_t = 1$; women: $D_t = 0$

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 D_t + \beta_4 D_t X_t + \varepsilon_t$$

H_0: $\beta_3 = 0$ vs. H_1: $\beta_3 > 0$ or (H_1: $\beta_3 \neq 0$) intercept

초임 차별(이)에 대한 검정 $\frac{b_3 - 0}{Se(b_3)} \sim t_{T-4}$

H_0: $\beta_4 = 0$ vs. H_1: $\beta_4 > 0$ or (H_1: $\beta_4 \neq 0$) slope

임금 증가율 차별(이)에 대한 검정 $\frac{b_4 - 0}{Se(b_4)} \sim t_{T-4}$

복합적 질적 효과에 대한 검정

계량경제학 9.20

Testing: H_o: $\beta_3 = \beta_4 = 0$
H_1 : otherwise

intercept and slope

$$\frac{(SSE_R - SSE_U) / 2}{SSE_U / (T - 4)} \sim F_{2,T-4}$$

$$SSE_U = \sum_{t=1}^{T} (y_t - b_1 - b_2 X_t - b_3 D_t - b_4 D_t X_t)^2$$

and

$$SSE_R = \sum_{t=1}^{T} (y_t - b_1^* - b_2^* X_t)^2$$

두 회귀식의 등가성에 대한 검정

계량경제학 9.21

F-검정

I. 더미변수를 이용한 접근

men: $D_t = 1$; women: $D_t = 0$

$$y_t = \beta_1 + \beta_2 X_t + \beta_3 D_t + \beta_4 D_t X_t + \varepsilon_t$$

H_o: $\beta_3 = \beta_4 = 0$ vs. H_1: otherwise

y_t = wage rate X_t = years of experience

⇒ F 검정

두 회귀식의 등가성에 대한 검정

계량경제학 9.22

F-검정

$$F = \frac{(SSE_R - SSE_{U1})/J}{SSE_{U1}/(T-K)}$$

J = # restrictions

K=unrestricted coefs.

$J = 2$ $K = 4$

두 회귀식의 등가성에 대한 검정

계량경제학 9.23

Chow 검정

II. 세번의 회귀분석을 통한 접근

남성과 여성의 차이가 없는 경우
모든 사람에 대해:
$y_t = \beta_1 + \beta_2 X_t + \varepsilon_t \rightarrow SSE_R$

남성과 여성의 차이가 있는 경우
남성만: $y_{tm} = \delta_1 + \delta_2 X_{tm} + \varepsilon_{tm} \rightarrow SSE_m$
여성만: $y_{tw} = \gamma_1 + \gamma_2 X_{tw} + \varepsilon_{tw} \rightarrow SSE_w$
⇒ Chow 검정

두 회귀식의 등가성에 대한 검정

계량경제학 9.24

Chow 검정

Let $SSE_{U2} \equiv SSE_m + SSE_w$

$$F = \frac{(SSE_R - SSE_{U2})/J}{SSE_{U2}/(T-2J)}$$

J = # coefficients

$J = 2$

In fact, $SSE_{U1} \equiv SSE_{U2}$. 두 가지 검정은 동일한 F값을 낳음

초우(Chow) 검정

구조적 안정성에 대한 검정

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + \varepsilon_t, \quad t = 1, \ldots, T$$

H_0 : no structural change between N_1 obs and T-N_1 obs

H_1 : yes

- N_1 이후 경제적 관계가 구조적으로 변화했는가를 검정하고자 함
- 절차:

1. 관측치의 수가 T 인 표본을 다음의 두 집단으로 나눔
   - 집단1 은 N_1 의 관측치로 구성.
   - 집단2 는 나머지 $T_2 = T - N_1$ 의 관측치로 구성.
2. 두 부분 집단들에 대해 각각 OLS를 돌리고 SSE_1 과 SSE_2 를 얻음

초우(Chow) 검정

구조적 안정성에 대한 검정

3. 전체 표본 (T)에 대해 OLS를 돌리고 SSE_R 를 얻음

4. F값을 계산: $F^* = \dfrac{(SSE_R - SSE_1 - SSE_2) / K}{(SSE_1 + SSE_2) / T\text{-}2K}$

5. 적절한 유의수준에 대해 임계값 $F^c_{K,\,T\text{-}2K}$ 을 계산함

   If $F^* > F^c$ ==> reject H_0

   It means that there is a structural change in the sample.