제8강 가변수

1. 모형
2. 유형
3. 복수 가변수

1. 모형

(1) 필요성

-가변수(dummy variable)는 독립변수 중의 일부가 성질을 달리하는 질적인 자료로 되어 있을 경우 사용됨

-가변수란 질적인 면을 나타내 주는 변수이며 가변수를 사용하면 더욱 정확한 통계적 추론을 할 수 있음

-질적인 면을 고려해야 함에도 불구하고 이를 고려하지 않을 경우 이는 모형내 있어야 할 변수를 뺀 경우이고 이 경우 추정량은 불편성을 가지지 못함

(2) 종류

-가변수는 질적 범주를 구별하기 의해 사용되는 변수이므로 모형에서 어느 한 질적 변수의 2개의 질적 범주를 구분하기 위해서 하나의 가변수가 통상적으로 쓰임

-둘 이상의 질적 변수나 어느 한 질적 변수의 두 개 이상의 질적 범주를 모형이 포함하고 있을 경우 두 개 이상의 가변수가 필요함

(예1)전시와 소비시의 소비행태

$$C_t = \beta_0 + \beta_1 Y_t + u_t$$ (보통회귀식)

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 D_t + u_t$$ (가변수가 포함된 회귀식)

단, $D_t = \begin{cases} 1, t\text{가 평화시} \\ 0, t\text{가 전쟁시} \end{cases}$

(예2)성별 및 인종별 임금격차(두 개의 질적 변수를 가진 경우)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$ (보통회귀식)

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 D_{2i} + u_i$$ (가변수가 포함된 회귀식)

단, $D_{1i} = \begin{cases} 1, i\text{가 남자} \\ 0, i\text{가 여자} \end{cases}$

$D_{2i} = \begin{cases} 1, i\text{가 백인} \\ 0, i\text{가 유색인} \end{cases}$

(예3)학력별 임금격차(두 개 이상의 질적 범주를 가진 경우)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$ (보통회귀식)

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{Hi} + \beta_3 D_{Ci} + u_i$$ (가변수가 포함된 회귀식)

단, $D_{Hi} = 0, D_{Ci} = 0,$ i가 중졸 이하

$D_{Hi} = 1, D_{Ci} = 0,$ i가 고졸

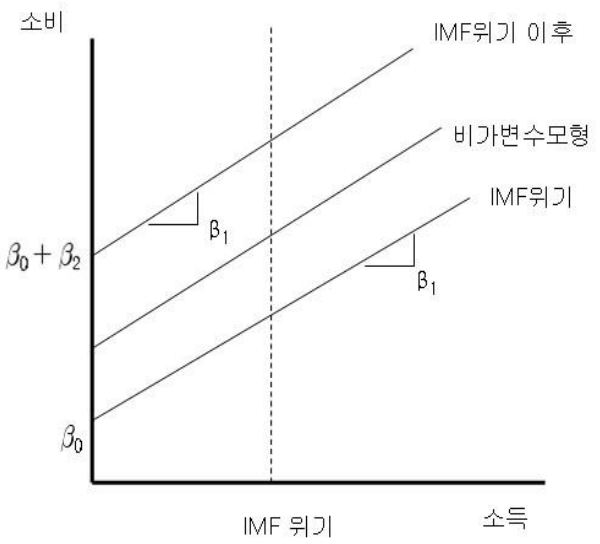
$D_{Hi} = 0, D_{Ci} = 1,$ i가 대졸 이상

2. 유형

(1)절편(평균)의 변화를 나타내는 가변수

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 D_t + u_t$$

단, $D_t = \begin{cases} 1, t\text{가 } IMF\text{위기 이후} \\ 0, t\text{가 } IMF\text{위기} \end{cases}$



자료 구조

연도	C(조 원)	Y(조 원)	D
.	.	.	.
1995	214	408	1
1996	245	458	1
1997	271	502	1
1998	252	492	0
1999	289	542	0
2000	330	600	0
2001	364	649	1
.	.	.	.
2009	577	1,068	1

-위의 가변수모형을 추정한 후 가변수가 통계적으로 유의할 경우 위 식의 추정은 다음의 두 식을 개별적으로 추정한 결과와 동일함

(IMF위기) $C_t = \beta_0 + \beta_1 Y_t + u_t$

(IMF위기 이후) $C_t = (\beta_0 + \beta_2) + \beta_1 Y_t + u_t$

(가설검정)
-가변수에 대한 가설검정에서 귀무가설은 $\boldsymbol{H_0}: \boldsymbol{\beta_2} = \mathbf{0}$으로 IMF위기와 IMF위기 이후의 소비수준에 차이가 없다는 가설임
-귀무가설을 기각하면(즉, $\boldsymbol{\beta_2}$가 통계적으로 유의하면) IMF위기와 IMF위기 이후의 소비수준에 차이가 있다고 결론을 내림
-따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 하는데 가변수를 포함한 식을 추정하면 이와 동일한 결과를 얻을 수 있음

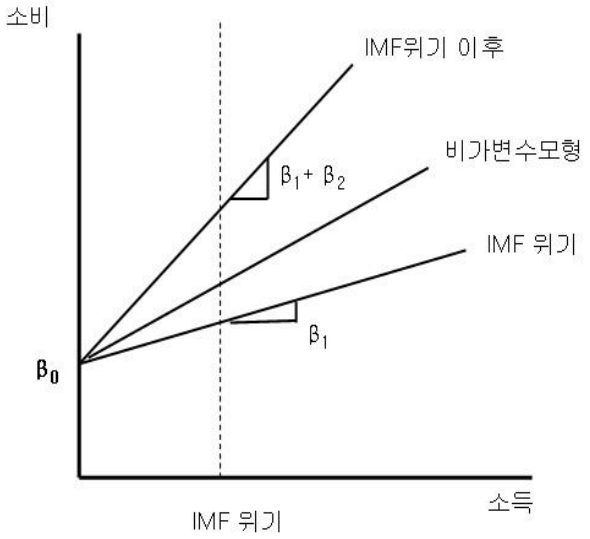
(회귀계수에 대한 해석)
-$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$: 한계소비성향으로 IMF위기의 한계소비성향과 IMF위기 이후의 한계소비성향이 같음
-$\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$: IMF위기의 절대소비수준
-$\widehat{\boldsymbol{\beta}}_{\mathbf{2}}$: IMF위기 이후 소비수준과 IMF위기 소비수준의 차이
-따라서 IMF위기 이후 소비수준은 $\widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \widehat{\boldsymbol{\beta}}_{\mathbf{2}}$

```
> m1.lm<-lm(c~d+y)

> summary(m1.lm)

Call:
lm(formula = c ~ d + y)

Residuals:
    Min      1Q  Median      3Q     Max
-2307.8  -806.9  -375.7  1054.6  3019.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.854e+03  4.432e+03   2.223   0.0368 *
d           3.419e+03  6.979e+02   4.898 6.74e-05 ***
y           5.092e-01  4.014e-02  12.685 1.36e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1448 on 22 degrees of freedom
Multiple R-squared:  0.8841,    Adjusted R-squared:  0.8735
F-statistic: 83.87 on 2 and 22 DF,  p-value: 5.092e-11
```

(2)기울기(한계)의 변화를 나타내는 가변수

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 D_t Y_t + u_t$$

단, $D_t = \begin{cases} 1, t\text{가 } IMF\text{위기 이후} \\ 0, t\text{가 } IMF\text{위기} \end{cases}$



자료 구조

연도	C(조 원)	Y(조 원)	D	D*Y
.	.	.	.	.
1995	214	408	1	408
1996	245	458	1	458
1997	271	502	1	502
1998	252	492	0	0
1999	289	542	0	0
2000	330	600	0	0
2001	364	649	1	649
.	.	.	.	.
2009	577	1,068	1	1,068

-위의 가변수모형을 추정한 후 가변수가 통계적으로 유의할 경우 위 식의 추정은 다음의 두 식을 개별적으로 추정한 결과와 동일함

(IMF위기) $C_t = \beta_0 + \beta_1 Y_t + u_t$

(IMF위기 이후) $C_t = \beta_0 + (\beta_1 + \beta_2) Y_t + u_t$

(가설검정)
-귀무가설은 $H_0: \beta_2 = 0$으로 IMF위기와 IMF위기 이후의 한계소비성향에 차이가 없다는 가설임
-귀무가설을 기각하면(즉, β_2가 통계적으로 유의하면) IMF위기와 IMF위기 이후의 한계소비성향에 차이가 있다고 결론을 내림
-따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 하는데 가변수를 포함한 식을 추정하면 이와 동일한 결과를 얻을 수 있음

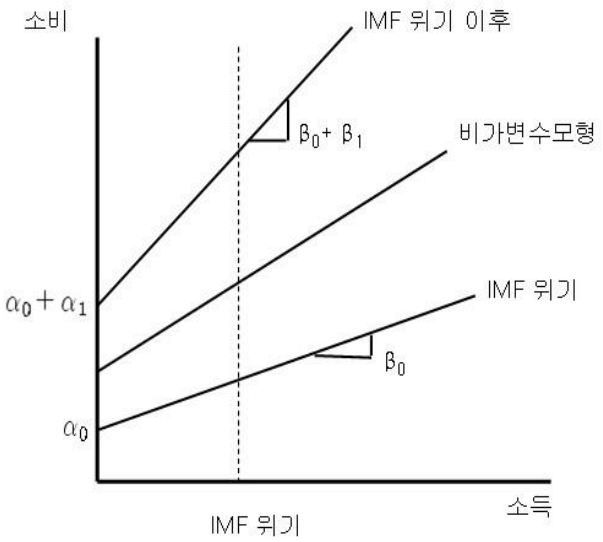
(회귀계수에 대한 해석)
-$\hat{\beta}_0$: 절대소비수준으로 IMF위기의 절대소비수준과 IMF위기 이후의 절대소비수준은 같음
-$\hat{\beta}_1$: IMF위기의 한계소비성향
-$\hat{\beta}_2$: IMF위기 이후 한계소비성향과 IMF위기 한계소비성향의 차이
-따라서 IMF위기 이후 한계소비성향의 값은 $\hat{\beta}_1 + \hat{\beta}_2$

```
> m2.lm<-lm(c~y+dy)

> summary(m2.lm)

Call:
lm(formula = c ~ y + dy)

Residuals:
    Min      1Q  Median      3Q     Max
-2230.1  -874.9  -292.3  1064.8  2923.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.043e+04  4.237e+03   2.463   0.0221 *
y            5.377e-01  4.235e-02  12.698 1.33e-11 ***
dy          -3.398e-02  6.652e-03  -5.108 4.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1416 on 22 degrees of freedom
Multiple R-squared:  0.8891,    Adjusted R-squared:  0.879
F-statistic:  88.2 on 2 and 22 DF,  p-value: 3.117e-11
```

(3)절편과 기울기의 동시변화를 나타내는 가변수

$$C_t = \alpha_0 + \alpha_1 D_t + \beta_0 Y_t + \beta_1 D_t Y_t + u_t$$

단, $D_t = \begin{cases} 1, t\text{가 } IMF\text{위기 이후} \\ 0, t\text{가 } IMF\text{위기} \end{cases}$



자료 구조

연도	C(조 원)	Y(조 원)	D	D*Y
.	.	.	.	.
1995	214	408	1	408
1996	245	458	1	458
1997	271	502	1	502
1998	252	492	0	0
1999	289	542	0	0
2000	330	600	0	0
2001	364	649	1	649
.	.	.	.	.
2009	577	1,068	1	1,068

-위의 가변수모형을 추정한 후 가변수가 통계적으로 유의할 경우 위 식의 추정은 다음의 두 식을 개별적으로 추정한 결과와 동일함

(IMF위기) $C_t = \alpha_0 + \beta_0 Y_t + u_t$

(IMF위기 이후) $C_t = (\alpha_0+\alpha_1) + (\beta_0+\beta_1) Y_t + u_t$

(가설검정) 3종류의 가설검정이 가능함

① $H_0: \alpha_1 = 0$ (IMF위기와 IMF위기 이후의 소비수준에 차이가 없다)
-귀무가설을 기각하면(즉, α_1이 통계적으로 유의하면) IMF위기와 IMF위기 이후의 소비수준에 차이가 있다고 결론을 내림
-따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 하는데 가변수를 포함한 식을 추정하면 이와 동일한 결과를 얻을 수 있음

② $H_0: \beta_1 = 0$ (IMF위기와 IMF위기 이후의 한계소비성향에 차이가 없다)
-귀무가설을 기각하면(즉, β_1이 통계적으로 유의하면) IMF위기와 IMF위기 이후의 한계소비성향에 차이가 있다고 결론을 내림
-따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 하는데 가변수를 포함한 식을 추정하면 이와 동일한 결과를 얻을 수 있음

③ $H_0: \alpha_1 = \beta_1 = 0$ (IMF위기와 IMF위기 이후의 소비수준 및 한계소비성향에 차이가 없다)
-귀무가설을 기각하면(즉, α_1과 β_1이 동시에 통계적으로 유의하면) IMF위기와 IMF위기 이후의 소비수준 및 한계소비성향에 차이가 있다고 결론을 내림
-따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 하는데 가변수를 포함한 식을 추정하면 이와 동일한 결과를 얻을 수 있음

(회귀계수에 대한 해석)

-$\hat{\alpha}_0$: IMF위기의 절대소비수준
-$\hat{\alpha}_1$: IMF위기 이후의 절대소비수준과 IMF 위기의 절대소비수준의 차이
-$\hat{\beta}_0$: IMF위기의 한계소비성향
-$\hat{\beta}_1$: IMF위기 이후의 한계소비성향과 IMF 위기의 한계소비성향의 차이

Regression Results of using Dummy Variable

	Dependent variable:			
	c			
	(1)	(2)	(3)	(4)
d		3,418.648***		-12,172.840
		(697.938)		(9,656.625)
y	0.400***	0.509***	0.504***	0.476***
	(0.047)	(0.040)	(0.038)	(0.044)
dy			0.034***	0.152
			(0.007)	(0.094)
Constant	23,033.120***	9,854.362**	10,433.910**	13,476.090**
	(4,980.694)	(4,432.226)	(4,236.569)	(4,827.492)
Observations	25	25	25	25
R2	0.758	0.884	0.889	0.897
Adjusted R2	0.747	0.874	0.879	0.882
Residual Std. Error	2,047.203 (df = 23)	1,447.710 (df = 22)	1,415.773 (df = 22)	1,397.192 (df = 21)
F Statistic	71.886*** (df = 1; 23)	83.870*** (df = 2; 22)	88.199*** (df = 2; 22)	60.903*** (df = 3; 21)
Note:				*p<0.1; **p<0.05; ***p<0.01

```
> jointHo<-c("d","dy")

> linearHypothesis(m3.lm, jointHo)
Linear hypothesis test

Hypothesis:
d = 0
dy = 0

Model 1: restricted model
Model 2: c ~ d + y + dy

  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     23 96393877
2     21 40995058  2  55398818 14.189 0.0001262 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

가설검정 요약

구분	1단계	2단계	3단계	4단계
Case 1	α1=0 기각 β1=0 허용	가변수 모형 1		
Case 2	α1=0 허용 β1=0 기각	가변수 모형 2		
Case 3	α1=0 기각 β1=0 기각	가변수 모형 3		
Case 4	α1=0 허용 β1=0 허용	α1=β1=0 허용	가변수가 없는 모형	
Case 5	α1=0 허용 β1=0 허용	α1=β1=0 기각	α1=0 기각 and β1=0 허용	가변수 모형 1
Case 6	α1=0 허용 β1=0 허용	α1=β1=0 기각	α1=0 허용 and β1=0 기각	가변수 모형 2
Case 7	α1=0 허용 β1=0 허용	α1=β1=0 기각	α1=0 기각 and β1=0 허용 또는 β1=0 기각 and α1=0 허용	가변수 모형 1 및 가변수 모형 2 중 결정계수가 큰 모형

3. 복수 가변수

(1)한 개 정성변수에 여러 범주의 경우

-한 개 정성변수에 범주가 n개(n≥3)인 경우 n-1개의 가변수를 생성

(예)학력을 중졸이하, 고졸, 대졸이상으로 구분할 경우 High, College 가변수를 생성하면 다음과 같이 3개의 범주로 구분이 됨

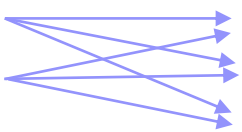
-중졸 이하 : High=0, College=0
-고졸 : High=1, College=0
-대졸 이상 : High=0, College=1

(2)여러 정성변수의 경우

-정성변수가 2개 이상인 경우 각각의 정성변수에 대해 가변수를 생성

(예)성별 및 학력별에 따른 임금구조를 알고자 할 경우 성별 가변수 Gender(남자=1, 여자=0)와 학력 가변수 High, College를 생성

-여자 : Gender=0
-남자 : Gender=1



-중졸 이하 : High=0, College=0
-고졸 : High=1, College=0
-대졸 이상 : High=0, College=1

(예)근무연수(나이, X) 외에 성별(Gender) 및 학력별(High, College) 임금구조를 분석하고자 할 경우 다음의 모형을 설정할 수 있음

$$Y_i = \alpha + \beta X_i + \gamma Gender_i + \delta High_i + \pi College_i + u_i$$

(해석)위 회귀식에서 다음과 같이 임금구조를 도출할 수 있음

중졸이하 여자의 경우 : $Y_i = \alpha + \beta X_i + u_i$

중졸이하 남자의 경우 : $Y_i = (\alpha + \gamma) + \beta X_i + u_i$

고졸 여자의 경우 : $Y_i = (\alpha + \delta) + \beta X_i + u_i$

고졸 남자의 경우 : $Y_i = (\alpha + \gamma + \delta) + \beta X_i + u_i$

대졸이상 여자의 경우 : $Y_i = (\alpha + \pi) + \beta X_i + u_i$

대졸이상 남자의 경우 : $Y_i = (\alpha + \gamma + \pi) + \beta X_i + u_i$

구분	여자	남자
중졸 이하	−1.363(reference)	(−1.363+0.658)=−0.705
고졸	(−1,363+0.389)	(−0.705+0.389)
대졸 이상	(−1,363+0.982)	(−0.705+0.982)

Regression Results of using Dummy Variable

	Dependent variable:
	income
age	0.052***
	(0.014)
gender	0.658***
	(0.209)
high	0.389
	(0.239)
college	0.982***
	(0.241)
Constant	-1.363**
	(0.596)
Observations	85
R2	0.409
Adjusted R2	0.379
Residual Std. Error	0.646 (df = 80)
F Statistic	13.820*** (df = 4; 80)
Note:	*p<0.1; **p<0.05; ***p<0.01