# Part III 회귀분석의 제 문제

## 제 7 장 모형의 설정 및 검정

### 1. 모형설정 오류(model misspecification):

- 실증분석에서 회귀모형이 잘못 설정되어 이를 통해 종속변수의 체계적인 움직임을 제대로 분석할 수 없는 경우

- 오류발생 원인

a) 사용변수의 측정오차가 불가피한 경우 (measurement error)

b) 종속변수와 설명변수 간의 관계식 자체가 잘못 설정된 경우(wrong function form)

c) 포함되어야할 적절한 설명변수가 모형에서 제외된 경우:누락변수(omitted variable)

d) 부적합한 설명변수가 모형에 포함된 경우(irrelevant variable)

e) 설명변수간의 공선성(colinearity) 문제가 발생하는 경우

1) 누락변수 (omitted variable)의 경우

- 적절한 모형: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

→ 설명변수 $X_{2i}$ 가 누락된 누락변수의 문제를 가진 모형: $Y_i = \alpha^* + \beta^*_1 X_{1i} + \varepsilon^*_i$

- 누락변수에 의해 야기될 수 있는 문제점:

a) 적절한 모형하에서 회귀계수 $\beta_1$ ($\beta_2$)의 추정은 $X_{2i}$ ($X_{1i}$)의 값을 일정하게 유지(통제, control)시키는 조건하에서 $X_{1i}$ ($X_{2i}$ )가 Y 에 미치는 영향 정도를 계측한다. 그러나, 누락변수의 문제를 포함한 모형하에서 추정된 $\beta_1$($\beta_2$)의 추정은 $X_{2i}$ ($X_{1i}$)의 값을 통제하지 못한 상태에서 도출된 것이다.

b) 잘못된 제약($\beta_2$=0)을 실제모형에 부과한 것과 같은 문제를 야기 한다

→ 잘못 설정된 모형하에서 그 오차항은 $\varepsilon^*_i = \beta_2 X_{2i} + \varepsilon_i$ 의 의미를 갖게되어 선형회귀모형의 기본가정인 $E[\varepsilon^*_i] = 0$ 을 충족시킬 수 없다. 따라서 최소자승법을 적용하면 그 추정치가 BLUE(최량선형불편추정량)이 되지 못한다:

→ 최소자승법에 의한 $\beta^*_1$의 추정량을 산출하면,

$\beta^*_1 = [\ \sum(X_{2i} - X_2)\ (Y_i - Y)/ \sum_{i=1,n} (X_{2i} - X_2)^2] = \beta_1 + \beta_2\ [Cov(X_{1i}, X_{2i})/Var(X_{1i})]$

⇨ $\beta_2\ [Cov(X_{1i}, X_{2i})/Var(X_{1i})] \neq 0$ (누락변수 편의: omitted variable bias)이고 따라서 $E(\beta^*_1) \neq \beta^*_1$ 이기 때문에 $\beta^*_1$ 는 편의추정량(biased estimator)이 된다.

⇨ 누락변수 편의, $\beta_2$ [$Cov(X_{1i}, X_{2i})/Var(X_{1i})$],는 표본크기가 아무리 커도 사라지지 않기 때문에 일치성(consistency)의 특성도 잃게 된다

c) 누락변수의 오류는 $\beta^*_1$ 의 추정량의 분산크기를 감소시켜 일반적인 t-통계치를 이용한 가설검정이나 신뢰구간 설정에서 잘못된 결론을 유발 시킬 수 있다

2) 부적절한 변수(irrelevant variable)의 경우

- 적절한 모형: $Y_i = \alpha + \beta_1 X_{1i} + \varepsilon_i$

→ 부적절한 설명변수 $X_{2i}$ 가 포함된 모형: $Y_i = \alpha^* + \beta^*_1 X_{1i} + \beta^*_2 X_{2i} + \varepsilon^*_i$

- 부적절한 변수의 포함에 의해 야기될 수 있는 문제점

a) 회귀계수($\beta^*_1$)의 추정량에 편의(bias)가 나타나지 않으므로 불편의성에는 영향을 주지 않는다

b) 회귀계수 추정량의 분산, $var(\beta^*_1)$의 크기가 크게 나타나 효율성이 떨어진다

c) 특히 부적절하게 포함된 변수($X_{2i}$)가 다른 설명변수($X_{1i}$)와 서로 밀접한 상관관계가 있는 경우에는, 추정량이 분산이 매우 커져 검정통계치인 t-통계량이 작아짐에 따라 원래 설명력이 높은데도 불구하고 유의성이 낮게 나타나,

귀무가설 $H_0$: $\beta^*_1 = 0$ 을 기각하기 어려워진다.

| | 누락변수의 경우 | 부적합한 변수의 경우 |
|---|---|---|
| 회귀계수의 추정량 | 편의성(bias) 및 비일관성 (inconsistency) | 불편의성 (unbiasedness) 유지 |
| 회귀계수의 분산값 | 작아짐 (하향편의) | 커짐(상향편의로 비효율성: inefficiency) |
| 오차항의 분산값 | 커짐(상향편의) | 거의 변동 없음 |
| t-통계량 | 불확실 | 대체로 하락 |
| 검정 및 구간추정의 결론 | 무의미 | 의미는 유지되나 귀무가설($H_0$: $\beta^*_1 = 0$)을 부당하게 채택할 가능성이 커짐 |

## 2. 모형설정 검정법(specification test)

- 실증분석 결과를 이용하여 결론을 도출하기 전에 사용된 계량모형이 제대로 설정되었는지를 확인하는 작업이 선행되어야 한다.

→ 변수사이의 관계를 선형으로 분석 가능한지 또는 회귀모형의 선정이 올바른지에 대해서 점검이 필요하다.

### 1) 유의성 검정법

- 주로 설명변수의 적합성 여부를 파악하기위해 사용된다
- 설정 모형: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$

→ 변수 $X_{ki}$ 가 이 모형에 포함되어야 하는지 검정하기 위한 가장 쉬운 방법이 변수에 대한 유의성 검정을 통해 유의성이 낮은 변수를 가려내는 방법

⇨ t-통계치 ($\beta_k/s_\beta$)를 이용하여, $H_0$: $\beta_k = 0$ 에 대한 유의성을 검정

⇨ 여러 개의 변수($X_a$, $X_b$, $X_c$)가 동시에 모형에 포함되어야 하는지 여부를 검정: F-검정량을 이용하여, $H_0$: $\beta_a = \beta_b = \beta_c = 0$ 을 판정,

- 주의할 사항은 t-검정 또는 F-검정을 실행하기전에 핵심적인 설명변수로 구성

된 기본모형에 대한 정당한 이론을 근거로 한 확신이 있어야 한다.

2) LM(Lagrnage Multiplier) 검정법

- 기존의 회귀모형에서 누락된 변수가 있는지를 검정하기 위해 적용된다
- 설명변수의 수가 서로다른 두개의 모형중에서 하나를 선택하는 방법으로 제약(restriction)이 반영된 모형을 귀무가설로 하고 제약이 없는 모형을 대립가설로 하여 서로 비교하여 결정한다,

→ 설정 모형 $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ 에서 $X_{2i}$ 변수가 부적합 변수인지 아니면 적합한 변수인지를 분석하고자 한다:

⇨ 제약: $\beta_2 = 0$

⇨ 검정 가설, $H_0$: $\beta_2 = 0$ ($Y_i = \alpha + \beta_1 X_{1i} + \varepsilon_i$) vs. $H_0$: $\beta_2 \neq 0$ ($Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_I$)

⇨ 검정 통계량, LM 통계량 = $nR^2$ (n: 표본관측치의 수, $R^2$ 는 대립가설하의 모형에서 도출한 결정계수의 값)

⇨ LM ~ $\chi^2$ (j) (j 는 자유도: 제약식의 수)

3) 정보 기준(information criteria)

- 모형선정 방법으로 모형의 설명력과 모형의 크기를 동시에 고려하는 방법

→ 새로운 변수의 도입이 모형의 설명력을 향상시키나 반면에 모형의 크기를 확대시켜 자유도의 감소를 초래 한다. 정보기준은 설명변수의 추가에 따른 장단점을 비교하여 적절한 모형을 선택하도록 하는 지표로 사용된다.

- AIC (Akaike Information Criterion)와 SIC (Schwarz Information Criterion)

→ AIC = log(RSS/n)+(2k/n), SIC = log(RSS/n)+[log(n)k/n]

- ⇨ 두가지 기준은 그값이 적을수록 모형의 적합도가 높다는 것을 나타낸다
- ⇨ 설명변수를 추가하면 잔차제곱합(RSS)이 작아져서 정보기준값을 감소시켜, 모형의 설명력을 높이는 통계적 이익(statistical benefit)을 얻을 수 있으나, 동시에 모형추정 비용을 추가시킴으로써 통계적 비용(statistical cost)를 증가시켜 정보기준의 값을 상승 시킨다.

4) 모형의 비선형성 검정

- 모형설정오류는 일반적으로 두가지 경우, 누락변수와 부적합한 변수에 의해

설명되어 져 왔으나, 실제로 오류의 요인은 매우 광범위하며 그중 매우 흔한 설정오류를 범하게 되는 경우가 함수형태 자체를 잘못 설정하는 상황이다

→ 경제이론이나 가설을 계량분석이 가능한 구체적 함수관계로 표현하는 과정에서 가장 편리한 선형함수모형의 적용이 적절함에도 불구하고 이를 무시하고 비선형모형을 고수할 경우 설정오류가 발생할 수 있다

- 설명변수의 비선형항을 모형에 포함시킬 것인지의 여부를 검정하기 위해 LM 검정법을 적용할 수 있다.

→ 회귀모형: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ 에서 비선형항 $X^2_{1i}$ 와 $X^2_{2i}$ 변수가 적합변수인지 아니면 부적합한 변수인지를 LM 검정법을 이용하여 분석하고자 한다:

⇨ 최소자승법을 이용하여 기존의 회귀모형을 추정, 잔차를 도출한다:

$$e_i = Y - \alpha^{\wedge} - \beta^{\wedge}_1 X_{1i} - \beta^{\wedge}_2 X_{2i}$$

⇨ 잔차항 $e_i$ 를 상수항과 설명변수 $X_{1i}$ $X_{2i}$ $X^2_{1i}$ $X^2_{2i}$ 에 대해 회귀분석을 하여, 결정계수 $R^2$ 값을 구한다

$$e_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X^2_{1i} + \gamma_4 X^2_{2i} + \mu_i$$

⇨ 비선형변수가 설명력이 있다면, 잔차 $e_i$ 는 이들과 상관관계를 가지고 있어야 한다. 이를 점검하기위해 LM 검정법을 이용하여 귀무가설 $H_0$:

$\gamma_3 = \gamma_4 = 0$

$LM = n R^2 \sim \chi^2 (j)$ (자유도 j=2: 제약식의 수)

⇨ 만일 귀무가설이 기각되면 종속변수와 설명변수사이에 비선형관계가 존재함을 의미한다.