제 8 장 다중공선성(Multicollinearity)

1. 다중공선성 (Multicollinearity)의 성격과 문제점

1) 완전공선성(perfect multicollinearity)

- 회귀모형, $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$에서 두설명변수가 완전한 선형관계($X_{1i} = \lambda X_{2i}$, $\lambda \neq 0$)인 경우 각 회귀계수에 대해 최소자승추정치를 구할 수 없게 되며, 이러한 경우를 완전 공선성(perfect multicollinearity)이라 한다.
- 일반적으로 k 개의 설명변수를 포함한 회귀모형에서 설명변수간의 관계가

  $\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k = 0$, ($\lambda_1 \neq 0$ $\lambda_2 \neq 0$ $\ldots \lambda_k \neq 0$)의 경우 설명변수사이에 완전공선성이 나타난다고 한다.

2) 다중공선성(multicollinearity)

- 일반적으로 회귀모형에서 설명변수간에 정확한 선형관계(완전 공선성)는 나타나지 않으며, 단지 그 상관관계가 높게 나타나는 문제가 발생하는데, 이경우를 다중공선성(multicollinearity)라 한다: $\rho(X_i\ X_j) \approx \pm 1$

3) 다중공선성이 추정에 미치는 효과: $Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$

a) 추정량, $\hat{\beta}_j$,(j=1..k)이 매우 불안정하게 되어 최소자승추정량 얻을 수 없게 된다

b) $\hat{\beta}_j$의 분산값이 매우 커지게 된다

→ $\hat{\beta}_j$의 표준오차($s_{\hat{\beta}}$)값이 커짐에 따라 t-검정 통계량이 매우 적게되어 귀무가설 (H_0: $\beta_j = 0$)을 기각할 가능성이 부당하게 희박해진다

→ 비록 결정계수 R^2 값이 높고 모형의 적합성(유의성)에 대한 F 값이 높음에도 불구하고 개별 추정량의 통계적 유의성이 없게 나타난다

c) 회귀계수의 추정치와 통계치가 자료의 크기 변화나 설명변수의 누락이나 부적절한 변수의 포함등에 의해 민감하게 변하게 된다

2. 다중공선성의 점검

1) 높은 R^2 값과 낮은 t-검정치

- R^2 값은 크고 개별 회귀계수에 대한 t-검정치가 낮은 경우, 각 설명변수들의 종속변수에 대한 설명력은 높으나 각계수의 추정치의 표준오차가 매우 크다는 것을 의미한다. 이는 설명변수간에 다중공선성이 높을 때 나타나는 현상이다

2) 설명변수들간의 높은 상관계수 값(0.9 이상)

3) 설명변수를 하나 추가하거나 또는 제거함에 따라 추정치값들이 크게변할 때,

4) 관측치(n)의 수를 추가하거나 제거함에 따라 추정치값들이 크게변할 때,

5) 중요한 설명변수의 표준오차가 매우 커거나 추정치의 부호가 예상과 다를 때

6) 각 추정치에 대한 유의성 검정통계량인 t-통계치가 아주 작음에도 불구하고 귀무가설, H_0: $\beta_1 = \beta_2 = \ldots = \beta_j = 0$ 에 대한 검정통계량 F-통계치가 매우 클 때 (종합적 영향력에 비해 개별 설명변수의 영향력이 미미하여 독립적으로 개별 설명변수의 영향력을 분리 추정하기 힘든 경우)

3. 다중공선성에 대한 대책

- 다중공선성을 내포하고 있는 회귀모형의 경우 회귀분석의 목적이 개별개수에 대한 추정이 아니라 종속변수 전체에 대한 예측에 있다면 다중공선성은 문제가 되지 않는다
- 만일 회귀분석 목적이 사용되는 자료의 특성상 개별 설명변수의 종속변수에 대한 효과나 영향을 분리 추정일 경우 이러한 개별변수에 대한 분리 추정을 할 수 없게 되는 문제점이 있으므로 이를 완화 시키거나 해결할 필요가 있다

1) 다중공선성 유발변수의 탐색 및 제거

→ 심각한 다중공선성을 유발시키는 설명변수를 확인한 다음 이들 변수중 추정에 해로운 변수를 제거하는 방법이다

a) 다중공선성 유발변수의 탐색

i) 각 설명변수를 다른 모든 설명변수 들에 대하여 회귀분석 한 다음 R^2 값을 서로 비교하여 가장 큰 R^2 값을 갖는 회귀방정식상의 선형함수관계가 높은 변수가 다중공선성 문제를 유발시키는 것으로 판정 할 수 있다

ii) 설명변수 간의 상관계수를 산출한 다음 H_0: $\rho_{ij} = 0$ 에 대한 t-검정통계량, $t = (\rho_{ij} \sqrt{n-k}) / (\sqrt{1-\rho_{ij}^2})$ 값을 이용하여 귀무가설을 검정한다. 만일 귀무가설이 기각되면, 두변수는 다중공선성을 유발시키는 것으로 볼 수 있다.

iii) 특정 설명변수를 제외시켰을때의 R^2 값이 제외시키기 이전의 R^2 값에 비해 감소하지 않을 경우 이 설명변수에 의해 다중공선성이 높아진 것으로 볼 수 있다.

b) 다중공선성 유발변수의 제거

i) 선형상관관계가 높거나 높은 다중공선성을 유발시키는 것으로 확인된 두 설명변수와 종속변수의 상관계수를 산출, 비교하여 상관계수값이 작은 설명변수를 제거한다. 이는 종속변수에 대한 추가적인 설명능력이 상대적으로 작다는 것을 의미하기 때문이다.

ii) β 계수(βcoefficient)나 탄력성(elasticity)를 이용한다

→ β 계수(βcoefficient) = $\hat{\beta}_i (s_{xi} / s_y)$, $\hat{\beta}_i$는 최소자승 추정치

→ 탄력성(elasticity) = $\hat{\beta}_i (X_i / Y)$,

→ β 계수나 탄력성 값을 비교하여 절대치가 작은 변수를 제외 시킨다

- 일단 회귀모형내에 도입된 변수는 합당한 이론이나 가설에 입각하여 채택된 것이므로 무작정 변수를 제거하면 다중공선성 문제는 완화될 수 있으나 때로는 더 심각한 "설정모형 오류"를 유발할 수 있기 때문에 주의해야 한다.

2) 자료와 모형의 보완 및 변형

- 다중공선성문제를 해소하는 궁극적이고 최선이 되는 방책은 표본관측치를 추가적으로 확보하거나 모형을 변형시켜 기존의 표본관측치값에 의해 야기된 다중공선성을 완화시키는 것이다

a) 일차분차(first difference) 또는 비율(ratio)을 사용하는 방법

→ 기본회귀모형, $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ 대신에, 일차분차식, $\Delta Y_i = \alpha + \beta_1 \Delta X_{1i} + \beta_2 \Delta X_{2i} + \varepsilon_i$ 또는 비율식 $(Y_i / X_{1i}) = \beta_1 + \beta_2 (X_{2i} / X_{1i}) + (\alpha / X_{1i}) + \varepsilon'_i$ 을 이용하여 추정함으로써 다중공선성을 완화시킬 수 있다

b) 추가적인 표본관측치를 확보하여 활용하는 방법

→ 추가로 확보하는 자료들은 가능한 기존의 설명변수의 평균값과 상이한 값을 많이 포함시켜야한다