2021.05.17.
김은영

제4장 회귀진단

4.7 다중공선성의 탐색

■ 다중공선성(multicollinearity)
- 설명변수들이 높은 선형종속관계에 있는 경우 다중공선성이 발생한다.
- 회귀계수의 추정과 추론에 심각한 문제를 초래한다.

■ 다중공선성의 존재여부를 판정하는 방법에 대해 살펴보자.

(1) 분산팽창인수(Variance Inflation Factor; VIF)

■ R_k^2: k-번째 설명변수를 반응변수로 하여 나머지 설명변수에 대하여 회귀분석을 하였을 때 얻게 되는 결정계수의 값

[예] $X = \{X_1, X_2, X_3, X_4\}$

• R_1^2: $X_1 \leftarrow \{X_2, X_3, X_4\}$

• R_2^2: $X_2 \leftarrow \{X_1, X_3, X_4\}$

• R_3^2: $X_3 \leftarrow \{X_1, X_2, X_4\}$

• R_4^2: $X_4 \leftarrow \{X_1, X_2, X_3\}$

$$VIF_k = \frac{1}{1-R_k^2} \qquad (k=1, \cdots, p-1)$$

■ k-번째 설명변수가 다른 설명변수들과 밀접한 관계에 있으면 R_k^2은 1에 가깝게 되고 VIF_k의 값은 커지게 된다.

■ VIF_k 값들 중 제일 큰 값이 다중공선성의 존재여부를 알려 주는데, 일반적인 기준으로 VIF_k의 값 중 가장 큰 값이 10보다 크면 다중공선성의 가능성이 있다고 판단한다.

(2) 조건수(Condition Number)

■ 임의의 $n \times p$ $(p < n)$ 행렬 X를 비정칙치분해(Singular Value Decomposition; SVD)하여 만들어진 비정칙치(singular value)를 $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$라 할 때

행렬 X의 조건수

$$\kappa(X) = d_1 / d_p$$

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$: $X^t X$의 고유치 → $\lambda_i = d_i^2$ ⇒ 조건수 $\kappa(X) = d_1/d_p = (\lambda_1/\lambda_p)^{1/2}$

[예제 4.2] VIF와 조건수를 확인해 봅시다.

4.8 자기상관과 더빈-왓슨 검정

■ 오차항의 독립성 여부를 판정하는 것은 매우 중요한 문제다.

- 시간에 따라 관측되는 자료들은 값들 간에 밀접한 관계를 가지게 된다.
- 모형의 구축단계에서 시간과 관련된 중요한 변수가 생략되면 설정된 회귀모형에서 오차항들이 서로 독립이 아니면서 상관관계를 가지게 된다.
- 이 경우 회귀계수의 불편성은 만족하나 최소분산성은 만족하지 못하게 되며, 독립성 가정 하에 수행되는 신뢰구간의 계산, 가설검정 등 여러 가지 추론의 결과가 정확하지 않게 되는 문제가 발생하게 된다.

■ 일차적으로 잔차분석에서 잔차의 산점도를 이용하여 오차항의 독립성 여부를 판정할 수 있다.
[단점] 그림으로 판단하는 주관성이 강하다.

→ 오차항의 독립성에 대한 검정을 객관적으로 실행할 수 있는 방법이 있을까?

⇒ 더빈-왓슨 검정(Durbin-Watson test)
• 오차항들 간에 일차자기상관관계가 있는지 조사하는 방법
• 상기 관계가 있다고 판정되면 오차항의 독립성에 대한 가정이 만족된다고 할 수 없다.

다음과 같은 단순선형회귀모형을 생각해 보자.

$$y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad \epsilon_i \sim i.i.d.\, N(0,\ \sigma^2)$$

$y_1,\ y_2,\ \cdots,\ y_n$: 독립 → y_i와 y_j 간 어떤 상관관계가 없음 → $Cov(y_i,\ y_j) = 0 \ \ \forall i \neq j$

이제 어떤 시간에 따라 관측되는 자료 y_t가 있다고 생각해 보자.

이 경우 y_{t-1}와 y_t 간에 상관관계가 존재할 수 있다.

$y_t = \beta_0 + \beta_1 X_t + \epsilon_t$

$\epsilon_t = \rho\epsilon_{t-1} + \delta_t \quad \rightarrow \; AR(1)$ 모형. 여기서, $|\rho| < 1$, $\delta_t \sim i.i.d.\ N(0, \sigma_\delta^2)$, ϵ_t와 δ_t: 독립

$$
\begin{aligned}
\epsilon_t &= \rho\epsilon_{t-1} + \delta_t \\
&= \rho(\rho\epsilon_{t-2} + \delta_{t-1}) + \delta_t = \rho^2\epsilon_{t-2} + \delta_t + \rho\delta_{t-1} \\
&= \rho^2(\rho\epsilon_{t-3} + \delta_{t-2}) + \delta_t + \rho\delta_{t-1} = \rho^3\epsilon_{t-3} + \delta_t + \rho\delta_{t-1} + \rho^2\delta_{t-2} \\
&\qquad\qquad\qquad\qquad \vdots \\
&= \sum_{j=0}^{\infty} \rho^j \delta_{t-j}
\end{aligned}
$$

$E(\epsilon_t) = E\left(\sum_{j=0}^{\infty} \rho^j \delta_{t-j}\right) = \sum_{j=0}^{\infty} \rho^j E(\delta_{t-j}) = 0 \qquad$ 여기서, $\delta_t \sim i.i.d.\ N(0, \sigma_\delta^2)$

$$
\begin{aligned}
Cov(\epsilon_t,\ \epsilon_{t-l}) &= Cov\left(\sum_{j=0}^{\infty} \rho^j \delta_{t-j},\ \sum_{k=0}^{\infty} \rho^k \delta_{t-l-k}\right) \\
&= Cov\left(\sum_{j=0}^{\infty} \rho^j \delta_{t-j},\ \sum_{j=l}^{\infty} \rho^{j-l} \delta_{t-j}\right) \qquad \text{여기서, } (k=0 \rightarrow j=l) \rightarrow \; j = l + k \\
&= E\left(\sum_{j=0}^{\infty} \rho^j \delta_{t-j},\ \sum_{j=l}^{\infty} \rho^{j-l} \delta_{t-j}\right) - E\left(\sum_{j=0}^{\infty} \rho^j \delta_{t-j}\right) E\left(\sum_{j=l}^{\infty} \rho^{j-l} \delta_{t-j}\right) \\
&\qquad\qquad \text{여기서, } E\left(\sum_{j=0}^{\infty} \rho^j \delta_{t-j}\right) = 0,\ E\left(\sum_{j=l}^{\infty} \rho^{j-l} \delta_{t-j}\right) = 0 \\
&= E\left(\sum_{j=l}^{\infty} \rho^{2j-l} \delta_{t-j}^2\right) = \sum_{j=l}^{\infty} \rho^{2j-l} E(\delta_{t-j}^2) \qquad \text{여기서, } E(\delta_{t-j}^2) = Var(\delta_{t-j}) = \sigma_\delta^2 \\
&= \sigma_\delta^2 (\rho^l + \rho^{2+l} + \rho^{4+l} + \cdots) \\
&= \sigma_\delta^2 \rho^l (1 + \rho^2 + \rho^4 + \cdots) \qquad \text{여기서, } |\rho| < 1 \\
&= \sigma_\delta^2 \rho^l \left(\frac{1}{1-\rho^2}\right)
\end{aligned}
$$

$\Rightarrow \; Var(\epsilon_t) = Cov(\epsilon_t,\ \epsilon_t) = \dfrac{\sigma_\delta^2}{1-\rho^2} \equiv \sigma_\epsilon^2 \quad \leftarrow \; l = 0$

: 오차항 간에 자기상관이 있는 경우에도 등분산성은 여전히 만족됨을 보여준다.

■ 시간 l만큼 떨어져 있는 두 오차항 간의 상관계수

$$Corr(\epsilon_t,\ \epsilon_{t-l}) = \frac{Cov(\epsilon_t,\ \epsilon_{t-l})}{\sqrt{Var(\epsilon_t)\,Var(\epsilon_{t-l})}} = \frac{\sigma_\delta^2 \rho^l \left(\frac{1}{1-\rho^2}\right)}{\sigma_\delta^2 \left(\frac{1}{1-\rho^2}\right)} = \rho^l$$

■ 오차항 간에 일차자기상관이 존재하는지에 대한 더빈-왓슨 검정통계량을 구하기 위하여 다음의 관계식을 생각해 보자.

$$\begin{aligned} E\left[(\epsilon_t - \epsilon_{t-1})^2\right] &= E(\epsilon_t^2 + \epsilon_{t-1}^2 - 2\epsilon_t \epsilon_{t-1}) \\ &= \sigma_\epsilon^2 + \sigma_\epsilon^2 - 2Cov(\epsilon_t,\ \epsilon_{t-1}) = \sigma_\epsilon^2 + \sigma_\epsilon^2 - 2\sigma_\delta^2 \rho \left(\frac{1}{1-\rho^2}\right) \\ &= 2\sigma_\epsilon^2 - 2\frac{\rho \sigma_\delta^2}{1-\rho^2} \end{aligned}$$

■ 위 식에서 일차자기상관계수 ρ의 값에 따라 아래 관계식들이 성립한다.

$$\begin{aligned} \rho = 0 \quad &\rightarrow \quad E\left[(\epsilon_t - \epsilon_{t-1})^2\right] = 2\sigma_\epsilon^2 \quad \Rightarrow \quad DW = 2 \\ \rho > 0 \quad &\rightarrow \quad E\left[(\epsilon_t - \epsilon_{t-1})^2\right] < 2\sigma_\epsilon^2 \quad \Rightarrow \quad DW < 2 \\ \rho < 0 \quad &\rightarrow \quad E\left[(\epsilon_t - \epsilon_{t-1})^2\right] > 2\sigma_\epsilon^2 \quad \Rightarrow \quad DW > 2 \end{aligned}$$

■ 위 관계식은 모수들을 포함하므로 추정치를 사용하여 우변 부등식의 대소 관계를 파악하면 역으로 일차자기상관계수 ρ에 대한 정보를 알 수 있을 것이다.

$H_0 : \rho = 0$ ← 오차항 간의 자기상관관계에 대한 검정

■ DW(Durbin-Watson) 검정통계량

$$DW = \frac{\frac{1}{t-1}\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\frac{1}{t}\sum_{t=1}^{n} e_t^2} \rightarrow DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \Rightarrow DW \approx 2(1-\hat{\rho})$$

만약 DW의 값이 2 근처이면 자기상관이 없고, 2보다 작으면 양의 자기상관관계, 2보다 크면 음의 자기상관관계가 있음을 알 수 있다.

■ DW 검정통계량의 기각역 [지정교재 부록 B <표 7> 참조]

$H_0 : \rho = 0$, $H_1 : \rho > 0$에 대하여

- $DW < d_L$ $\Rightarrow$ 귀무가설을 기각할 수 있다.
- $DW > d_U$ $\Rightarrow$ 귀무가설을 기각할 수 없다.
- $d_L \le DW \le d_U$ $\Rightarrow$ 판정 보류

[예제 4.2] 자기상관에 대한 검정을 해 봅시다.