다중공선성

위키백과, 우리 모두의 백과사전.

다중공선성(多重公線性)문제(Multicollinearity)는 통계학의 회귀분석에서 독립변수들 간에 강한 상관관계가 나타나는 문제이다. 독립변수들간에 정확한 선형관계가 존재하는 **완전공선성**의 경우와 독립변수들간에 높은 선형관계가 존재하는 **다중공선성**으로 구분하기도 한다. 이는 회귀분석의 전제 가정을 위배하는 것이므로 적절한 회귀분석을 위해 해결해야 하는 문제가 된다.

진단법 [편집]

1. 결정계수 R^2값은 높아 회귀식의 설명력은 높지만 식안의 독립변수의 P-value값이 커서 개별 인자들이 유의하지 않는 경우가 있다. 이런 경우 독립변수들 간에 높은 상관관계가 있다고 의심된다.
2. 독립변수들간의 상관계수를 구한다.
3. 분산팽창요인(Variance Inflation Factor)를 구하여 이 값이 10을 넘는다면 보통 다중공선성의 문제가 있다.

해결법 [편집]

1. 상관관계가 높은 독립변수중 하나 혹은 일부를 제거한다.
2. 변수를 변형시키거나 새로운 관측치를 이용한다.
3. 자료를 수집하는 현장의 상황을 보아 상관관계의 이유를 파악하여 해결한다.
4. PCA(Principle Component Analysis)를 이용한 diagonal matrix의 형태로 공선성을 없애준다.

다중공선성

다수의 독립변수가 서로 지나치게 높은 상관관계를 가지면서
회귀계수 추정의 오류가 발생하는 문제

"상관계수 r≥0.9"

↑ '높은 상관관계'에 대한 기준.

이렇게 독립변수 간의 상관성이 높으면

상관관계가 높은 경우

↑ 두 독립변수 간의 상관관계가 높으면($|r| \geq 0.9$), 두 변수는 동일한 변수일 가능성이 높다. 즉, 두 독립변수는 서로 '독립'적인 관계가 아닐 가능성이 높다.

똑같은 변수 두 개를 집어넣고

계수[a]

모형		비표준화 계수 B	비표준화 계수 표준화 오류	표준화 계수 베타	t	유의확률	공선성 통계량 공차	공선성 통계량 VIF
1	(상수)	25.540	2.408		10.605	.000		
	TV광고	.798	.057	.592	14.107	.000	1.000	1.000
2	(상수)	37.257	6.269		5.943	.000		
	TV광고	.772	.058	.573	13.345	.000	.949	1.054
	리플렛광고	-.198	.098	-.087	-2.023	.044	.949	1.054

a. 종속변수: A브랜드_만족도

회귀분석 마다 회귀계수 부호(±) 변화가 제각기

$$+\beta \text{ or } -\beta$$

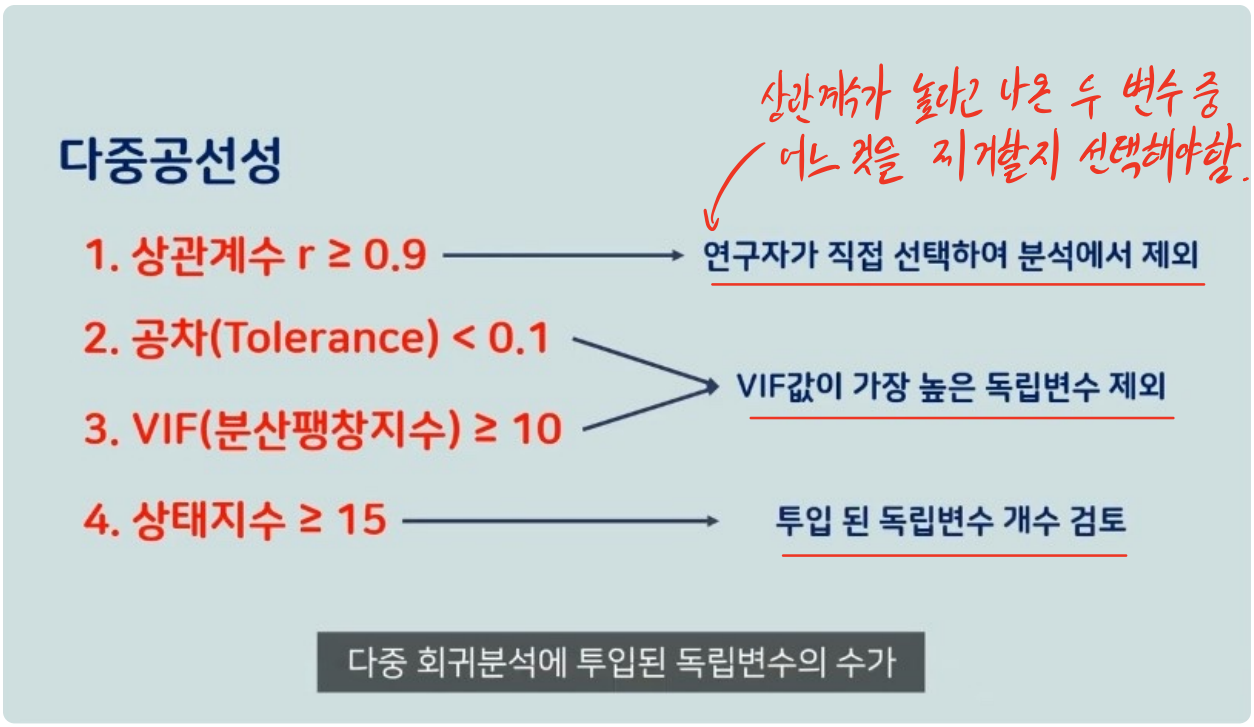
※ 'A'라는 독립변수가 단순회귀분석에서는 회귀계수가 양수로 나타나는데, 다중회귀분석에서는 회귀계수가 음수로 나타난다면, 다중공선성을 의심해 봐야한다.

만약

다중공선성

1. 상관계수 r ≥ 0.9
2. 공차(Tolerance) < 0.1
3. VIF(분산팽창지수) ≥ 10
4. 상태지수 ≥ 15

이 네 개 조건중 하나라도 해당되면 '다중공선성'이 있는 것이다.


다중공선성
상관계수가 높다고 나온 두 변수 중
어느 것을 제거할지 선택해야함
1. 상관계수 r ≥ 0.9
연구자가 직접 선택하여 분석에서 제외
2. 공차(Tolerance) < 0.1
3. VIF(분산팽창지수) ≥ 10
VIF값이 가장 높은 독립변수 제외
4. 상태지수 ≥ 15
투입 된 독립변수 개수 검토
다중 회귀분석에 투입된 독립변수의 수가

다중공선성(Multicollinearity)

- 다중공선성이란
  - 상관관계가 매우 높은 독립변수들이 동시에 모델에 포함될 때 발생
  - 왜 문제가 되는가?
    - 만약 두 변수가 완벽하게 다중공선성에 걸려 있다면
    - 즉, $x_1 = b_2 \times x_2$ 라면
    - 같은 변수를 두 번 넣은 것임 → 최소제곱법 계산상 어려움
  - 완벽한 다중공선성이 아니어도 문제인가?
    - 그렇다. 회귀계수의 표준오차가 비정상적으로 커짐
    - 왜 문제인가?
      - 회귀계수의 유의성은 t-값에 의해 결정됨
      - t-값은 회귀계수를 표준오차로 나누어서 계산 됨
      - 결국, 다중공선성으로 인해 표준오차가 비정상적으로 커지면 t-값이 작아져서 유의해야할 변수가 유의하지 않게 됨

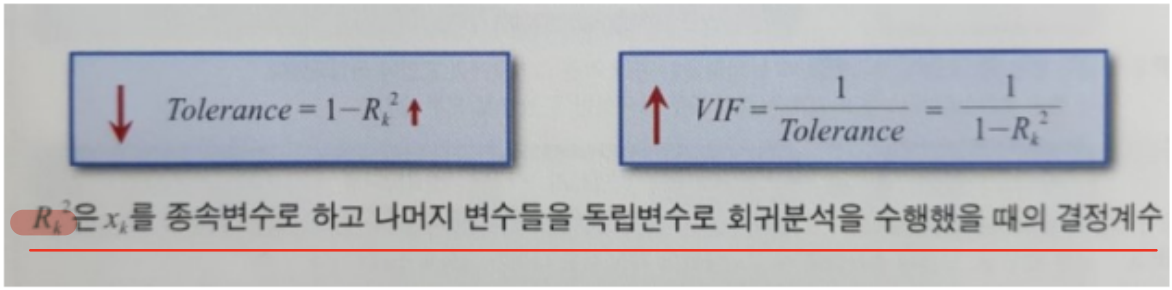
다중공선성(Multicollinearity)

- 다중공선성을 어떻게 찾아낼까?
  - ① 산포도 & 상관계수
    - 두 독립변수의 산포도를 보았을 때, 심각하게 상관관계가 높아보인다면
    - 상관계수를 확인
      - 상관계수가 만약 0.9를 넘는다면 (0.9보다 작아도) 다중공선성의 문제가 있을 수 있음
  - ② 허용/공차 (tolerance)를 확인
    - Tolerance란 한 개의 독립변수를 종속변수로 나머지 독립변수를 독립변수로 하는 회귀분석을 했을 때 나오는 R^2를 이용해 $(1 - R^2)$ 를 의미 함
    - 만약, R^2 가 1 이라면 독립변수 간에 심각한 상관관계가 있음을 의미
      - 이 경우 tolerance인 $(1 - R^2)$ 는 "0"이 될 것임
      - 따라서 tolerance가 0이면 완벽한 상관성을 의미하여 다중공선성이 심각함을 의미

다중공선성(Multicollinearity)

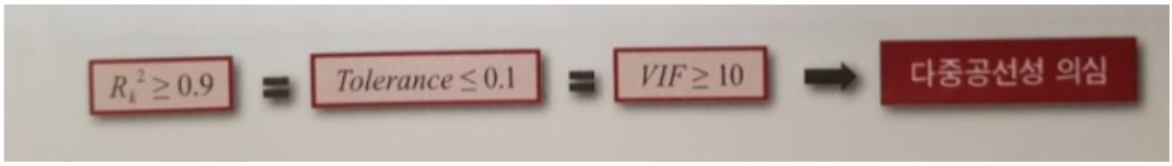
- 다중공선성을 어떻게 찾아낼까?
  - 분산팽창지수 (VIF: Variance Inflation Factor)
    - VIF = 1 ÷ tolerance = 1 ÷ (1 − R^2)
    - VIF가 크다는 것은 다중공선성이 크다는 의미
    - 일반적으로 10보다 크면 문제가 있다고 판단함
    - 그러나 이는 연속형변수의 경우에 해당된다고 보아야 함
    - 만약 더미변수의 VIF가 3 이상이라면 이 경우 다중공선성을 의심해야 함
  - 상태지수 (Condition Index)
    - 흔하게 사용되지는 않음
    - 100 이상이면 심각한 다중공선성이 존재

회귀모형에 다중공선성이 있는 경우, 각 변수들의 공차한계 tolerance와 분산팽창요인 variance inflaction factor, VIF를 통해 파악할 수 있다.



[출처] 배정민 (2012). <<닥터 배의 술술보건의학통계>>. 한나래출판사. 180p

여기서 **R^2k는 회귀식의 독립변수들 중 Xk만을 종속변수로 하고, 나머지 변수들은 그대로 독립변수로 회귀분석을 수행하였을 때의 결정계수를 의미**한다. **R^2k이 크다는 것은 이미 다른 변수들에 의해 Xk가 거의 설명되고 있음을 의미**한다. 예를들어 R^2k가 0.9 이상이라면 Xk는 이미 다른 변수들에 의해 90%이상 설명되고 있으므로 불필요한 변수일 가능성이 매우 높다. 이 경우 공차한계는 0.1 이하가 되고, 분산팽창요인 VIF는 10 이상이 된다.



[출처] 배정민 (2012). <<닥터 배의 술술보건의학통계>>. 한나래출판사. 180p

분산팽창요인 VIF는 해당 변수에 의해 회귀계수 추정량의 분산이 팽창되어 있는 정도를 의미하므로, 값이 클수록 회귀식의 신뢰도는 감소한다. 일반적으로 **VIF가 10 이상이면 다른 변수와 다중공선성**이 있는 것으로 간주한다.