Multicollinearity(다중공선성)\_study

PYB

2021 5 3

### **다중공선성**

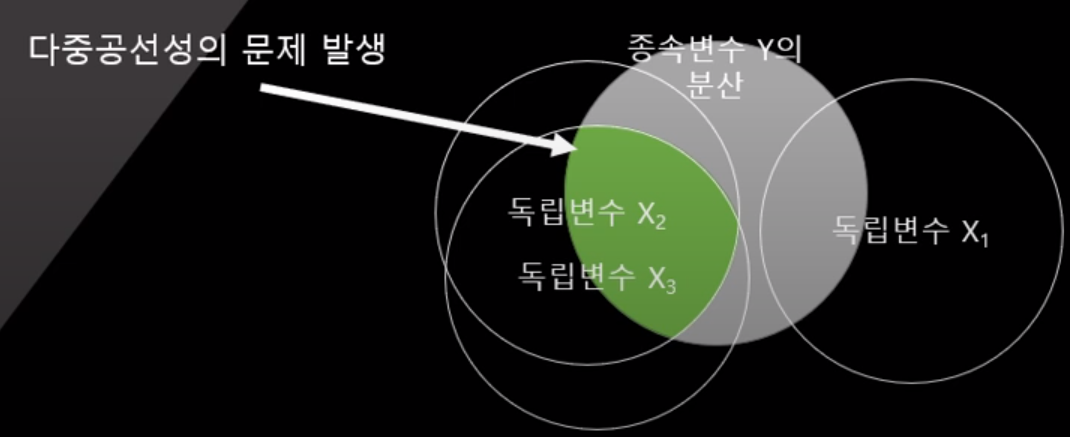
*y = a + b1xx1 + b2xx2 + b3xx3*  
세 개의 독립변수를 통해 종속변수 y의 분산을 얼마나 설명하느냐의 문제

#### **다중공선성이란?**

* 상관관계가 매우 높은 독립변수들끼리 동시에 모델에 포함될 때 발생
* 독립변수 간에 상관성이 높으면 두 변수가 동일한 변수라고 봐도 무방하다.

#### **왜 문제가 되는가?**

* 독립변수들끼리 너무 많이 겹처서 발생하는 문제이다. 이것은 종속변수 y의 분산을 설명할 때 과한 설명을 하게 된다.



사진

#### **다중공선성을 어떻게 찾아낼까?**

**[산포도 & 상관계수]**

* 두 독립변수의 산포도를 보았을 때, 심각하게 상관관계가 높아보인다면 상관계수를 확인한다.
* 상관계수가 만약 0.9를 넘는다면(혹은 0.9보다 작아도) 다중공선성에 문제가 있다.

**[분산팽창지수(VIF: Variance Inflation Factor)]**

* VIF가 크다는 것은 다중공선성이 크다는 의미이다.
* 일반적으로 연속형 변수인 경우, 10보다 크면 문제가 있다고 판단한다.
* 더미변수는 3 이상이라면 다중공선성을 의심해야한다.

#### **해결방법**

* 높은 다중공선성에도 불구하고 유의하다면 그래도 두어도 좋다.
* 유의해야할 변수가 유의하지 않다면 변수 중 누군가를 빼야한다.
* 다른 여러 방법이 있으나 일반적으로 완벽하지 않다.
* 문제는 강제로 분산을 제거하는 방법이 더 큰 문제를 만들 수 있다.

library(car) # vif 함수용

## Loading required package: carData

test <- read.csv('test\_vif2.csv')  
test

## 학생번호 시험점수 아이큐 공부시간 등급평균  
## 1 1 100 125 30 3.9  
## 2 2 95 104 40 2.6  
## 3 3 92 110 25 2.7  
## 4 4 90 105 20 3.0  
## 5 5 85 100 20 2.4  
## 6 6 80 100 20 2.2  
## 7 7 78 95 15 2.1  
## 8 8 75 95 10 2.1  
## 9 9 72 85 0 1.5  
## 10 10 65 90 5 1.8

# test에 있는 학생정보는 불필요하므로 제외  
test <- test[-1]

# 회귀 모델 생성  
regression\_model <- lm(test$시험점수~.,data=test)  
summary(regression\_model)

##   
## Call:  
## lm(formula = test$시험점수 ~ ., data = test)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.3146 -1.2184 -0.4266 1.5516 5.6358   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 50.30669 35.70317 1.409 0.2085   
## 아이큐 0.05875 0.55872 0.105 0.9197   
## 공부시간 0.48876 0.17719 2.758 0.0329 \*  
## 등급평균 7.37578 8.63161 0.855 0.4256   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.952 on 6 degrees of freedom  
## Multiple R-squared: 0.9155, Adjusted R-squared: 0.8733   
## F-statistic: 21.68 on 3 and 6 DF, p-value: 0.001275

summary를 보면 아이큐와 등급평균의 p value 값이 유의수준 0.05보다 크다는 것을 알 수 있다.  
이것은 결론적으로 회귀식에서 아이큐와 등급평균이 유의미하지 않다는 것을 의미한다.  
이러한 경우, 아이큐와 등급평균 두 개의 데이터는 불필요해진다. 하지만 정말 과연 그럴까?

이때 팽창계수를 확인한다.

# 팽창계수 확인  
vif(regression\_model)

## 아이큐 공부시간 등급평균   
## 22.643553 2.517786 19.658264

아이큐와 등급평균이 22, 19로 매우 높은 값을 나타낸다는 것을 알 수 있다.  
이렇게 **팽창계수가 높다는 것은 다중공선성이 있다**는 뜻이고, **다중공선성이 있다함은 독립변수들 중에 서로 영향을 주는 것이 있다**는 뜻이다.  
따라서 우리는 서로 영향을 주는 변수들 중 하나만을 남겨 독립변수로서의 성능을 할 수 있도록 해줘야한다.  
팽창계수를 확인하여 독립변수인지 아닌지를 판단하는 방법에 대해서, 현업 기준으로는 팽창계수가 보통 10보다 큰 것을 골라내고, 엄격하게는 5보다 큰 것, 느슨하게는 15~20보다 큰 것을 골라낸다.

우선적으로 팽창계수가 가장 큰 컬럼부터 지워나가며 다시 회귀모델을 만들고, 또 거기서 팽창계수가 큰 것을 지워나가는 것을 반복하며 회귀모델을 만들어가야 한다.

또한 상관계수로로 판단할 수 있다.

cor(test[c('아이큐', '공부시간', '등급평균')])

## 아이큐 공부시간 등급평균  
## 아이큐 1.0000000 0.7710712 0.9736894  
## 공부시간 0.7710712 1.0000000 0.7300546  
## 등급평균 0.9736894 0.7300546 1.0000000

상관계수를 확인해보면 아이유와 공부시간의 상관계수는 0.771, 아이큐와 등급평균은 0.974, 공부시간과 등급평균은 0.73으로 상관계수가 높은 편에 속한다. 아이큐와 등급평균의 상관관계가 0.9이상으로 매우 높으므로 다중공선성이 있음을 의심해야하는데, 위의 vif() 함수를 통해서 확인할 수 있었다.

**[판단법]**

1. 상관계수 >= 0.9 > 연구자가 직업 선택하여 분석에서 제외
2. 공차(Tolerance < 0.1 > VIF 값이 가장 높은 독립변수 제외하여 조건에 맞을 때까지 다중회귀분석 실시
3. VIF(분산팽창지수) >= 10 > VIF 값이 가장 높은 독립변수 제외하여 조건에 맞을 때까지 다중회귀분석 실시
4. 상태지수 >= 16 > 회귀분석에 투입된 독립변수의 수가 과하게 많은 경우를 생각하여 독립변수 개수 점검

이 중 하나라도 해당되면 **다중공선성이 있다**고 판단한다.