

05

다중 공선성 진단 방법

[다중공선성이란?]

독립변수 중 강한 상관관계가 있는 변수는 어떻게 처리해야

할까요?

- 삭제 or 주성분 or 능형회귀(ridge)를 사용합니다.

00:00

5.5 . 다중공선성이 존재하는가?

- 다중공선성은 회귀분석에서 독립변수들 간에 강한 상관관계가 나타나는 문제
- 다중공선성의 문제가 존재하면 정확한 회귀계수 추정이 어려움
- 독립변수들간의 상관관계를 파악한 후, 다중공선성이 발생하면 변수를 제거하거나 주성분, Ridge회귀 등을 적용하여 문제를 해결
- 다중공선성을 검사하는 방법

- 1) 독립변수들 간의 상관계수를 구하여 상관성을 직접 파악 0.9 이상이면 다중공선성이 있다고 판단
- 2) 허용 오차를 구했을 때 0.1이하이면 다중공선성 문제가 심각하다고 할 수 있음
 $\text{허용오차} = (1 - R^2)$: 한 독립변수의 분산 중 다른 독립변수들에 의해서 설명되지 않는 부분을 의미함, 즉 그 값이 작을 수록 공선성은 높다고 볼 수 있음.
- 3) 분산팽창요인(VIF)은 허용오차의 역수로 그 값이 클수록 독립변수들 간의 상관성이 높다. 일반적으로 VIF가 10 이상이면 공선성의 문제가 심각하다고 본다.

```

In [1]: 1 import pandas as pd
        2
        3 # 데이터 불러오기
        4 Cars = pd.read_csv('../data/Cars93.csv')
  
```

[다중공선성 진단 방법 실습]

03:45

```

In [2]: 1 Cars
Out[2]:

```

	Manufacturer	Model	Type	Min.Price	Price	Max.Price	MPG.city	MPG.highway	AirBags	DriveTrain	Passengers
0	Acura	Integra	Small	12.9	15.9	18.8	25	31	None	Front	...
1	Acura	Legend	Midsized	29.2	33.9	38.7	18	25	Driver & Passenger	Front	...
2	Audi	90	Compact	25.9	29.1	32.3	20	26	Driver only	Front	...
3	Audi	100	Midsized	30.8	37.7	44.6	19	26	Driver & Passenger	Front	...
4	BMW	535i	Midsized	23.7	30.0	36.2	22	30	Driver only	Rear	...
...
88	Volkswagen	Eurovan	Van	16.6	19.7	22.7	17	21	None	Front	...
89	Volkswagen	Passat	Compact	17.6	20.0	22.4	21	30	None	Front	...
90	Volkswagen	Corrado	Sporty	22.9	23.3	23.7	18	25	None	Front	...
91	Volvo	240	Compact	21.8	22.7	23.5	21	28	Driver only	Rear	...