회귀분석팀

6팀

신성민
신유정
김찬영
윤주희
이혜인

INDEX

다중공선성이란?

정의 **문제점**

다중공선성의 문제점

- 회귀 계수의 분산/표준편차가 커져서 T검정 통계량이 낮아짐

  If, T검정 통계량이 낮아지면 귀무가설($\beta_i = 0$) 채택할 확률이 높아짐!

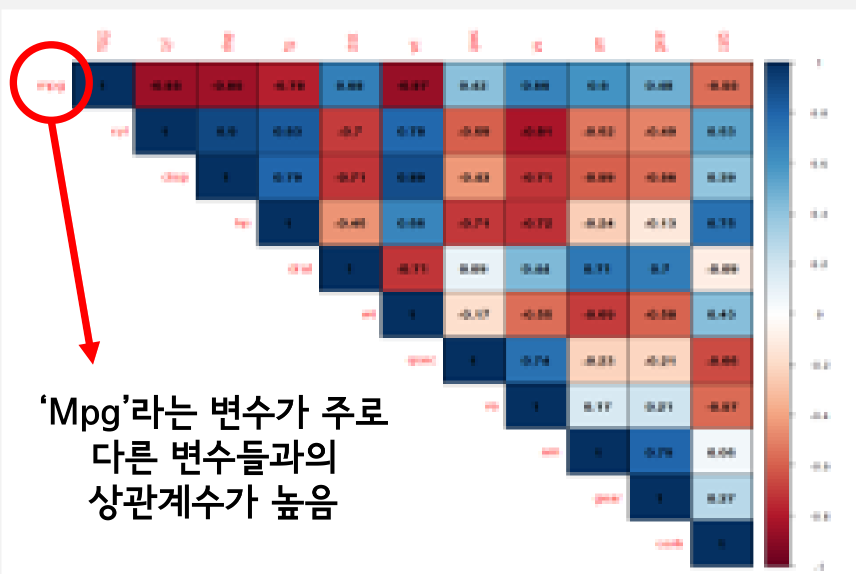
- 회귀계수의 추정치가 데이터의 작은 변화와 변수 추가/제거에 민감하게 반응하게 됨

다중공선성의 판별법 산점도와 상관계수 분산확대인자 상태지수

산점도와 상관계수

- 변수마다의 상관계수 표현한 그래프로 확인

<실제 Correlation Plot 예시>



- 색이 짙을수록 상관계수 절대값 높음
- 예시와 같이 전반적으로 상관계수가 다 높으면 다중공선성 의심 가능

※R의 Corrplot패키지 이용하면 됨!

다중공선성의 판별법

산점도와 상관계수 **분산확대인자** 상태지수

분산확대인자 (VIF) 값의 해석

$$VIF_j = \frac{1}{1-R_j^2}, (j = 1, 2, 3, \ldots, p)$$

- R^2가 1에 근사할수록 VIF값은 커짐
- 따라서 VIF값이 클수록 해당 변수로 인한 다중공선성 의심 가능
- 보통 10이상일 경우, 심각한 다중공선성 의심 가능
- 만약 VIF가 1이라면 다중공선성에 문제가 전혀 없는 변수

<Mtcar데이터 활용한 VIF예시 in R>

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear
7.613515	14.592594	12.632396	7.684396	3.260755	11.101134	7.489241	4.930285	4.970200	4.421571

※VIF값을 얻으려면 car 패키지 안에 있는 ‘vif’를 이용하면 됨!

다중공선성의 판별법

산점도와 상관계수 분산확대인자 **상태지수**

상태지수

- P개의 예측변수들의 상관행렬은 p개의 고유값 가짐
- 이를 내림차순으로 정렬하면 $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_p$
- 이때, 상태지수란! 최대 고유값 나누기 i번째 고유값의 루트!

$$\left(k_i = \sqrt{\frac{\lambda_1}{\lambda_i}} \quad (i = 1,2,3,\ldots,p) \right)$$

<Mtcar데이터의 고유값>

```
eigen() decomposition
$values
 [1] 6.60840025 2.65046789 0.62719727 0.26959744 0.22345110 0.21159612 0.13526199 0.12290143 0.07704665 0.05203544 0.02204441
```

다중공선성 해결

PCA 변수선택

PCA ; Principal Component Analysis

측정된 변수들의 선형 조합에 의해 대표적인 주성분을 만들어 차원을 줄이는 방법

➡ 기존 변수를 **선형결합**(linear combination)해 새로운 변수를 만들어 낸다

➡ 이때 생긴 새로운 변수들은 다중공선성 문제에서 자유롭다

주성분끼리는 직교(orthogonal)하기 때문에-!

다중공선성 해결

PCA 변수선택 1)척도 2)종류

Mallows C_p

$$Cp = \frac{SSEp}{\hat{\sigma}^2} + (2p - n) = SSE_p * \frac{n-P}{SSE} + (2p - n) = n - p + 2p - n = p$$

P: 설명 변수 개수 / n: 데이터 개수

- SSE_p 가 FM의 SSE와 비슷하다면 Cp ≈p

➢ C_p의 값이 p에 근사한 경우 좋은 회귀모델로 판별

다중공선성 해결

PCA　변수선택　1) 척도　2) 종류

AIC

$$AICp = n\ln(SSE_p/n) + 2p$$

- 모형 선택에 있어서 정확도(적합)와 간명성(적은 변수) 사이의 상충을 잘 조절하기 위한 것
- 변수의 개수로 페널티 부여
- AIC값에서 큰 차이를 보이면 유의미한 차이가 존재한다고 간주

➢ 작을수록 좋은 모델

다중공선성 해결

PCA 변수선택 1) 척도 2) 종류

BIC

$$BICp = n\ln(SSE\,p/n) + p(\ln n)$$

- AIC의 페널티 부여 방식을 수정한 방법
- n>8인 경우 BIC 기준이 더 큰 페널티를 주게 된다.
- 작을수록 좋은 모델

다중공선성 해결

PCA 변수선택 1)척도 2)종류

1. 전진적 선택 절차(Forward Selection)

2. 후진적 선택 절차(Backward Selection)

3. 단계적 선택 절차(Stepwise Selection)

다중공선성 해결

정의 Ridge Lasso Elastic Net

Shrinkage Method?

각 변수의 $\hat{\beta}_j$을 수축시켜, 분산을 낮추는 방법!

Shrinkage penalty

이런 감성이랄까..

$\boxed{\hat{\beta}_j}$

구성

LSE식 + Penalty term

다중공선성의 해결 Part 2

다중공선성 해결

정의 **Ridge** Lasso Elastic Net

Ridge regression 수식

tuning parameter, $\lambda \geq 0$

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \underbrace{\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2}_{\text{LSE 식}} + \lambda \underbrace{\sum_{j=1}^{p} \boxed{\beta_j^2}}_{\text{penalty}} \right\}$$

특징

모든 $\hat{\beta}_j$의 효과를 일정 비율로 감소시켜 0에 가깝게 만들어 준다!

4 다중공선성의 해결 Part 2

다중공선성 해결

정의 Ridge **Lasso** Elastic Net

Lasso regression 수식

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \underbrace{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{LSE 식}} + \overset{\text{tuning parameter, } \lambda \ge 0}{\lambda} \underbrace{\sum_{j=1}^{p}\boxed{|\beta_j|}}_{\text{penalty}} \right\}$$

특징

모든 $\hat{\beta}_j$의 효과를 감소시켜 아예 0으로 만들어 준다! 즉, 변수 선택이 가능하다!

다중공선성 해결

정의 Ridge Lasso **Elastic Net**

Elastic Net 수식

tuning parameter, $\lambda \geq 0$

$$\underbrace{\frac{\sum_{i=1}^{n}(y_i - x_i^J\hat{\beta})^2}{2n}}_{\text{LSE 식}} + \lambda\left(\underbrace{\frac{1-\alpha}{2}\sum_{j=1}^{m}\boxed{\hat{\beta}_j^2}}_{\text{Ridge penalty}} + \underbrace{\alpha\sum_{j=1}^{m}\boxed{|\hat{\beta}_j|}}_{\text{Lasso penalty}}\right)$$

tuning parameter, $0 \leq \alpha \leq 1$

특징

Ridge와 Lasso가 공존하는 감성이다!