Part. 02

회귀분석

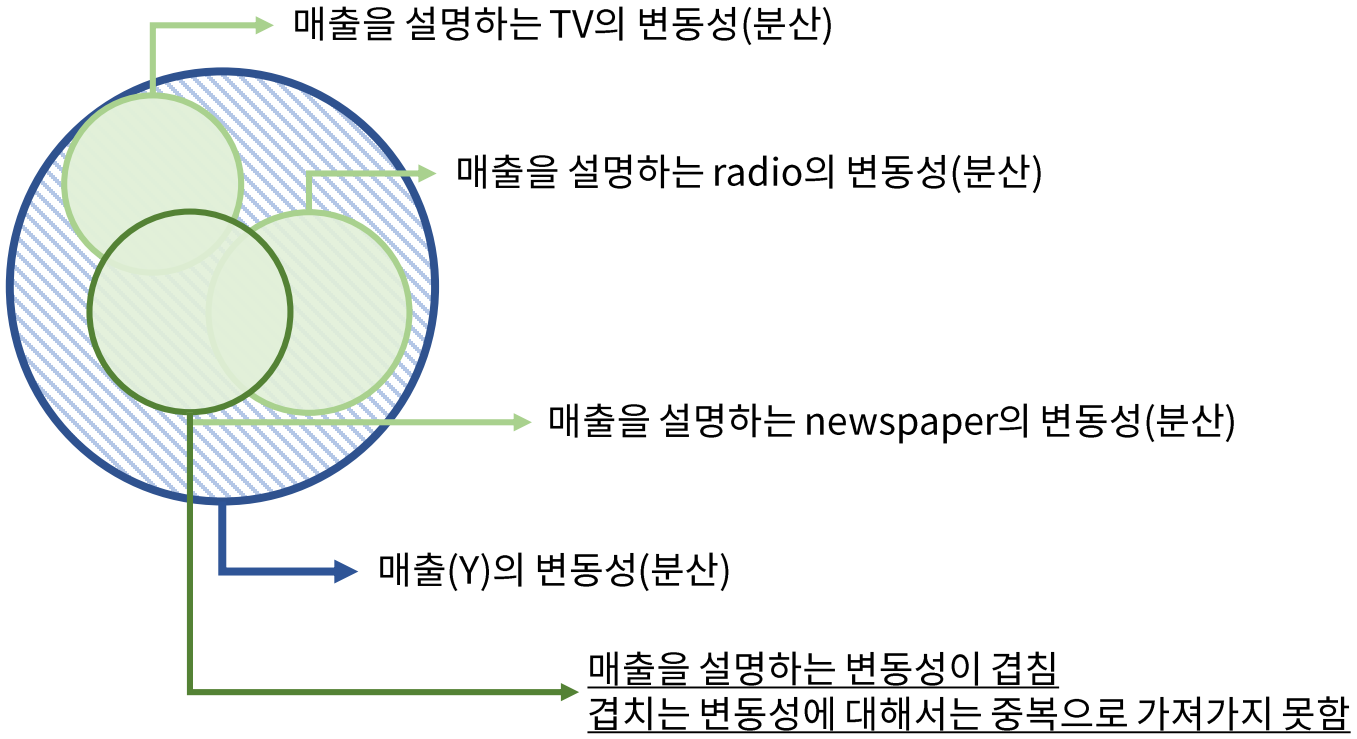
다중공선성 진단 방법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택

I 다중 선형 회귀분석

- 다중공선성(Multicollinearity)



이러한 현상에 대해서 변수들간의 다중공선성(Multicollinearity)이 있다고 한다.

잘못된 변수해석, 예측 정학도 하락 등을 야기시킨다

I 다중 선형 회귀분석

- 다중공선성을 진단하는 방법

- VIF(Variance inflation factor), 변수들간의 Correlation 등으로 진단

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

- 다중공선성을 해결하는 방법

- Feature Selection : 중요 변수만 선택하는 방법
  · 단순히 변수를 제거하는 방법 (correlation 등의 지표를 보고)
  · Lasso
  · Stepwise
  · 기타 변수 선택 알고리즘 (유전알고리즘 등)

- 변수를 줄이지 않고 활용하는 방법
  · AutoEncoder등의 Feature Extraction 기법 (딥러닝기법)
  · PCA
  · Ridge

중요한 Feature(변수)를 뽑는 방법은 데이터사이언스 분야에서도 현재까지 큰 이슈

I 다중 선형 회귀분석

- VIF(Variance inflation factor)

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

VIF가 10 이상인 경우 다중공선성이 있는 변수라고 판단

$$x_1 = \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \varepsilon$$

X1을 종속변수, 나머지 변수를 독립변수로 하여 회귀 모델($f1$) 적합

↓

$R_1{}^2$

$f1$ 의 R^2를 이용하여 VIF_1계산

↓

$$VIF_1 = \frac{1}{1-R_1{}^2}$$

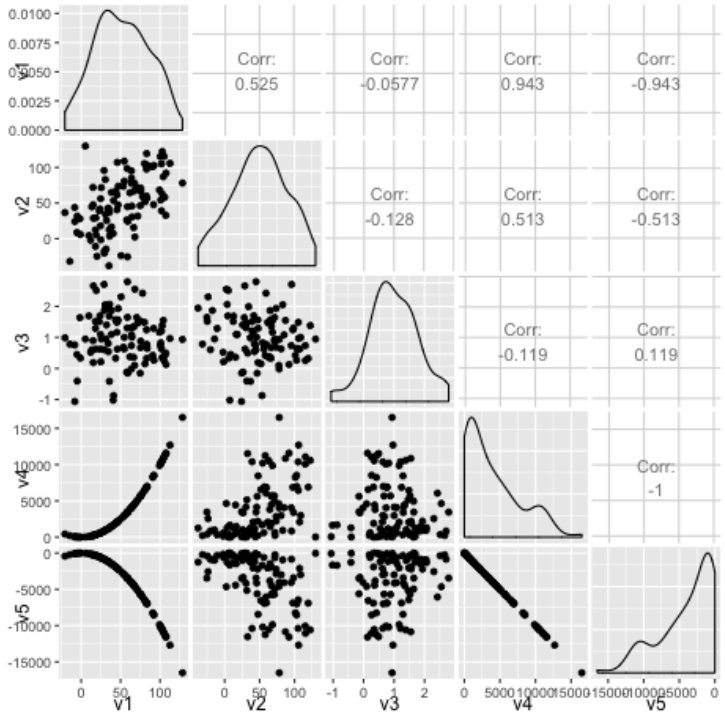
VIF_1의 의미 : 다른 변수의 선형결합으로 X1을 설명할 수 있는 정도

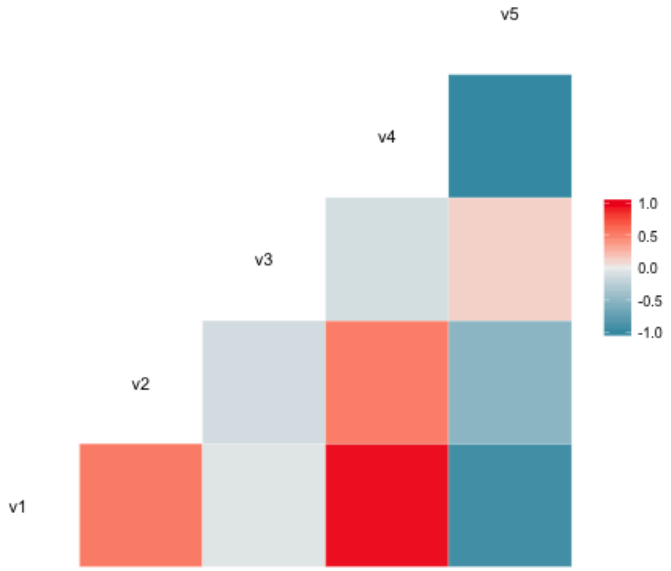
$R^2 > 0.9$ 이상인 경우, $\mathrm{VIF} > 10$

I 다중 선형 회귀분석

- 상관행렬(correlation matrix)

  - 상관행렬 및 산점도를 보고 판단





I 다중 선형 회귀분석

- 다중공선성을 근본적으로 해결하는 방법은 (아직) 없다.

  - 최근 머신 러닝 기법들은 중요 Feature 만을 뽑는 알고리즘 내장

중요변수 100개 vs 모든변수 10000개

↓

중요변수 100개가 무조건 좋음
(예측성능, 학습속도, 활용측면)

↓

머신러닝 기법은 기본적으로 학습데이터 내에서 예측력을 높이기 위해 최대한 많은 변수를 활용하려 함

Part. 02
회귀분석

회귀모델의 성능지표

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택