Part. 02

회귀분석

I 다중공선성

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택

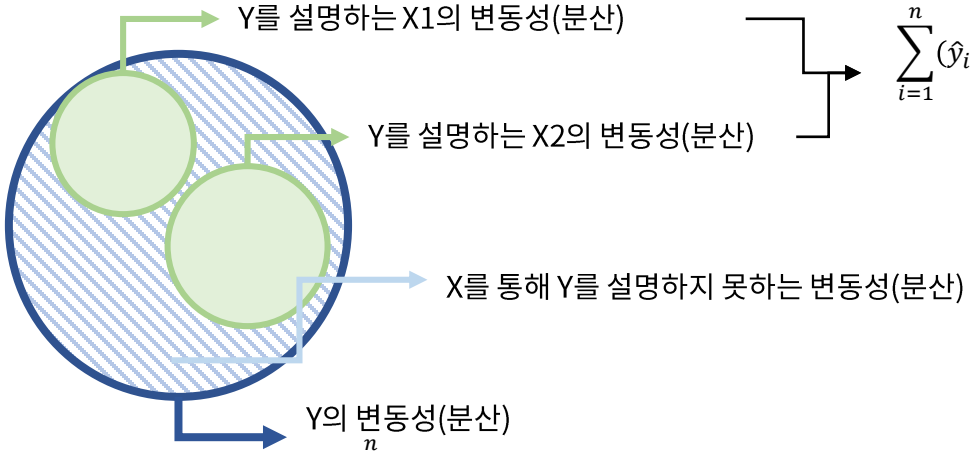
I 다중 선형 회귀분석

- 다중 선형 회귀 모델 검정

귀무가설 : $B_1 = B_2 \ldots B_p = 0$ (모든 회귀계수는 0이다, 즉 변수의 설명력이 하나도 존재 하지 않는다)

대립가설 : 하나의 회귀계수라도 0이 아니다. (즉 설명력이 있는 변수가 존재 한다.)

➡ 기각 하기 너무 쉬운 가설. 변수가 추가 되면 추가 될수록 기각하기 쉬워진다.



$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

② 제곱합의 형태 이기 때문에 변수가 추가되면 추가될수록 자연스레 증가한다.

③ 제곱합의 형태로 검정을 하는 F검정의 특성상 변수가 추가되면 자연스레 기각하기 쉬워진다. (R^2도 커짐)

$\sum_{i=1}^{n}(y_i - \bar{y})^2$

① ➡ 총 변동은 정해져 있다 (바뀌지 않는다)

I 다중 선형 회귀분석

- 다중공선성(Multicollinearity)

  독립변수들이 강한 선형관계에 있을때 다중공선성이 있다고 한다.

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

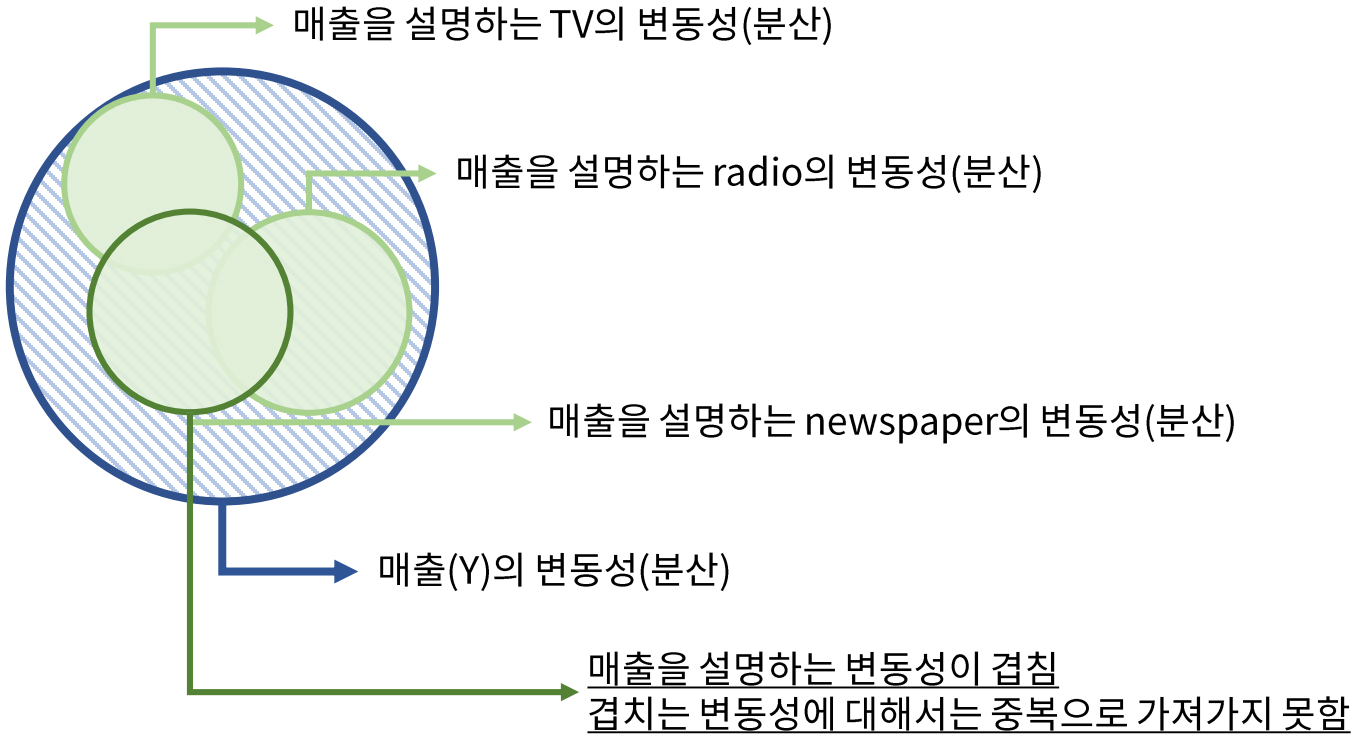
	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	−0.001	0.0059	−0.18	0.8599

이런 현상이 발생하는 이유는?

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

Part. 02

회귀분석

다중공선성 진단 방법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택