금융인을 위한

통계와 데이터 분석 입문

회귀분석 이해하기

한국금융연수원
KOREA BANKING INSTITUTE

학습 내용

1. 선형회귀분석 모형이란?
2. 회귀계수, 결정계수 의미 및 해석
3. 회귀진단
4. 다중회귀분석에서 고려할 부분 : 변수선택법, 다중공선성
5. 파이썬을 활용한 회귀분석

회귀분석

- **연속형/수치형 변수가 다른 변수들에 의해 설명되고 어떻게 예측될 수 있는지를 알아보기 위한 통계적 방법**

예

소득과 신용카드 사용액의 관계
: 소득이 증가하면 신용카드 사용액이 늘어날까?

- 회귀분석에서 변수 구분
  - 종속변수(Y)
  - 독립변수(X) : 설명변수

회귀분석

- **연속형/수치형 변수가 다른 변수들에 의해 설명되고 어떻게 예측될 수 있는지를 알아보기 위한 통계적 방법**

예

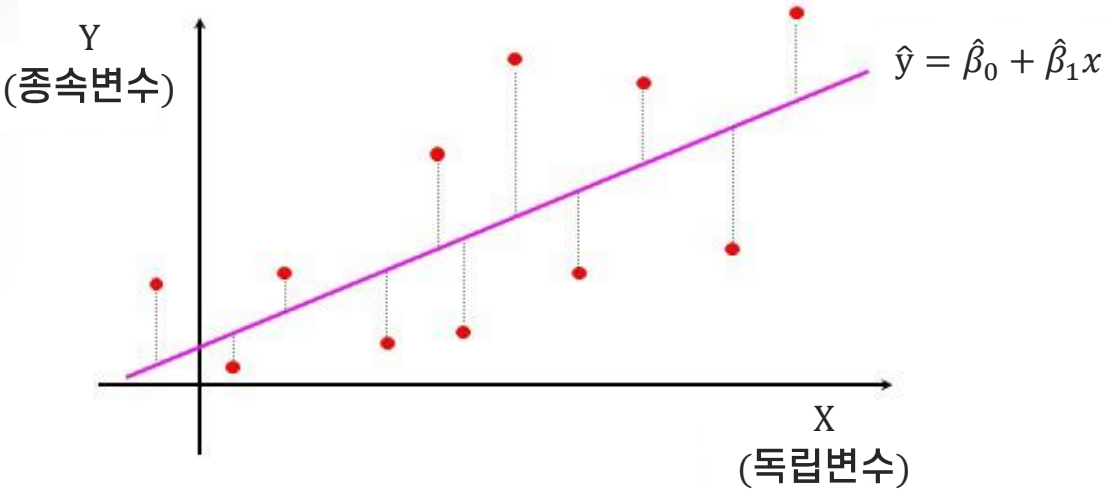
소득과 신용카드 사용액의 관계
: 소득이 증가하면 신용카드 사용액이 늘어날까?

- 독립변수의 개수에 따른 구분
  - **단순회귀** : 독립변수가 **하나**인 회귀분석
  - **다중회귀** : 독립변수가 **여러 개**인 회귀분석

선형회귀모형(linear regression model)

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$
  - 종속변수와 독립변수는 **선형** 관계
  - X_i들은 독립
  - 회귀계수 β_i들은 미지의 모수값이며 잔차의 제곱합을 최소화하는 값으로 추정됨
  - 오차 ε들은 독립이며 동일한 분포 - 정규분포 $N(0, \sigma^2)$를 가짐
    - Y도 X가 주어졌다는 가정하에 정규분포를 따름 : 연속형 변수
    - 추정된 β들도 정규분포를 따르므로 유의성 검정 및 신뢰구간 추정이 가능함

선형회귀모형(linear regression model)



회귀계수의 해석

회귀계수 β_i

다른 독립변수들이 고정되었을 때 독립변수 X_i가 1단위 증가할 때 종속변수 Y가 변화하는 양을 나타냄

- $\beta_i > 0$
  : 독립변수 X_i의 값이 증가하면 종속변수의 값도 증가함
- $\beta_i < 0$
  : 독립변수 X_i의 값이 증가하면 종속변수의 값은 감소함

회귀계수의 해석

- 회귀계수의 절대적인 크기는 영향력을 결정하는데 큰 의미가 없음

예

동일한 데이터라도 독립변수의 단위를 바꾸면 회귀계수의 크기가 변동하나 영향력이 변하는 것은 아님

회귀계수의 유의성

- 회귀계수에 대한 유의성 검정을 통하여 독립변수가 종속변수에 의미있는 영향을 주는지를 검토해 보아야 함

가설

$H_0 : \beta_i = 0 \; vs. \; H_1 : \beta_i \neq 0$

→ **귀무가설의 의미 :**
기울기가 0, X_i 가 Y를 설명하지 못함

- **p-value < α** 이면, H_0를 기각하고 H_1을 채택함
  : β_i는 통계적으로 유의함(즉, X_i는 Y를 설명함)
- **p-value > α** 이면, H_0를 기각하지 못함
  : β_i는 통계적으로 유의하지 않음
  (즉, X_i는 Y를 설명하지 못함)

회귀계수의 유의성

- y절편에 해당하는 β_0을 제외한 모든 회귀계수에 대한 유의성 검정 결과에 따라서 해당하는 독립변수를 모형에 포함할지 말지 결정함

결정계수

결정계수(R^2 , r-squared)

추정된 회귀식이 반응변수의 변이를 얼마나 설명하느냐,
즉 회귀직선의 적합도를 측정하는 가장 대표적인 측도

- 1(100%)에서 회귀식이 설명하지 못한 부분을 빼는 방식으로 계산됨
- $R^2 = 1 - \frac{\sum_1^n (y_i - \hat{y_i})^2}{\sum_1^n (y_i - \bar{y})^2} = \frac{\sum_1^n (\hat{y_i} - \bar{y})^2}{\sum_1^n (y_i - \bar{y})^2}$
- $0 \leq R^2 \leq 1$: 값이 클수록 설명력이 큼

→ 종속변수와 독립변수(들) 간의 관계를 회귀계수의 가설검정, 결정계수 등으로 파악할 수 있음

회귀진단

- 회귀식을 적합시킨 후 남은 잔차를 가지고 회귀진단함
- 이상치(outlier)와 영향치(influential observation) 유무를 확인함
  - 이상치(outlier)
    - 특정한 관찰치의 y값이 나머지 자료들의 일반적인 x-y의 trend에서 벗어나는 경우
  - 영향치(influential observation)
    - 회귀식(회귀계수들)에 영향을 주는 관찰치
  - 잔차값으로 계산된 통계량들과 plot들을 통해서 확인하고 분석에 포함할지 여부를 검토함

회귀진단

- **오차항에 대한 가정(독립이고 동일하며 $N(0, \sigma^2)$를 따름)에 위배되는지 확인함**
  - 잔차 normal QQ plot 등 잔차 plot 검토, 정규성 검정, 자기상관(autocorrelation) 여부 검토 등

예

시계열 자료의 경우 잔차들의 자기상관 관찰됨

변수선택법

- 독립변수들의 수가 많을 때 이들 변수 모두가 종속변수를 설명하는데 필요하지 않을 수 있음
- 변수선택법
  - 전진선택법(forward selection)
    - 고려하고 있는 변수들 중에서 중요한 독립변수부터 **차례대로 모형에 추가**해 나가는 방법
  - 후진선택법(backward selection)
    - 고려하고 있는 독립변수들을 모두 모형에 포함시킨 후에 중요도가 떨어지는 변수들부터 **차례대로 제외**시켜 나가는 방법
  - 단계적 선택법(stepwise selection)
    - 전진선택법과 후진선택법이 결합된 것으로 독립변수의 **추가와 제거**를 반복적으로 함으로써 전진선택법과 후진선택법의 단점을 보완함

다중공선성

다중공선성(collinearity)

독립변수들 간에 선형관계가 있을 때를 말함

- 다중공선성이 있는 경우 분산의 추정량이 커지면서 **회귀계수 추정량의 유의성이 낮게** 나올 수 있음
- 분산확대인자(VIF) 등 통계량 값으로 다중공선성 여부를 알아낼 수 있음

해결방법

선형관계가 있는 독립변수들 중 설명력이 낮은 변수를 회귀모형에서 제외시킴

학습정리

- 연속형 변수를 다른 변수들과의 관계를 통해서 예측하려고 하는 방법을 **회귀분석**이라고 함

- 종속변수(Y)와 독립변수(X)들 간의 선형관계가 있을 때 적절한 모형임

- 오차들은 **독립**이며 **동일한 분포** - **정규분포** $N(0, \sigma^2)$를 가진다는 가정임

- **회귀계수 β_i** 는 **다른 독립변수들이 고정**되었을 때 **독립변수 X_i 가 1단위 증가**할 때
  종속변수 Y 가 변화하는 양으로
  유의성 검정을 통하여 독립변수가 종속변수에 의미있는 영향을 주는지를 검토해 보아야 함

학습정리

- **결정계수**(R^2, r-squared)는 추정된 회귀식이
  반응변수의 변이를 얼마나 설명하는지를 말해주는 통계량

- 종속변수와 독립변수(들) 간의 관계를
  회귀계수의 가설검정, 결정계수 등으로 파악할 수 있음

- 회귀식 적합 후 **회귀진단**을 통해 모형에 대한 가정을
  위반하지 않았는지 확인해야 함

- **다중회귀분석**의 경우는 **변수선택법**을 통해
  의미있는 독립변수들을 선택할 수 있고
  추가적으로 **다중공선성** 여부를 확인해야 함

- 파이썬을 활용한 회귀분석이 가능함