회귀분석

with python™

01. 회귀분석

◆ 정의

✓ 하나 또는 그 이상의 독립변수들이 종속변수에 미치는 영향을 추정할 수 있는 통계 기법

✓ 독립변수의 개수가 하나이면 단순 선형회귀분석

✓ 독립변수의 개수가 두 개 이상이면 다중 선형회귀분석

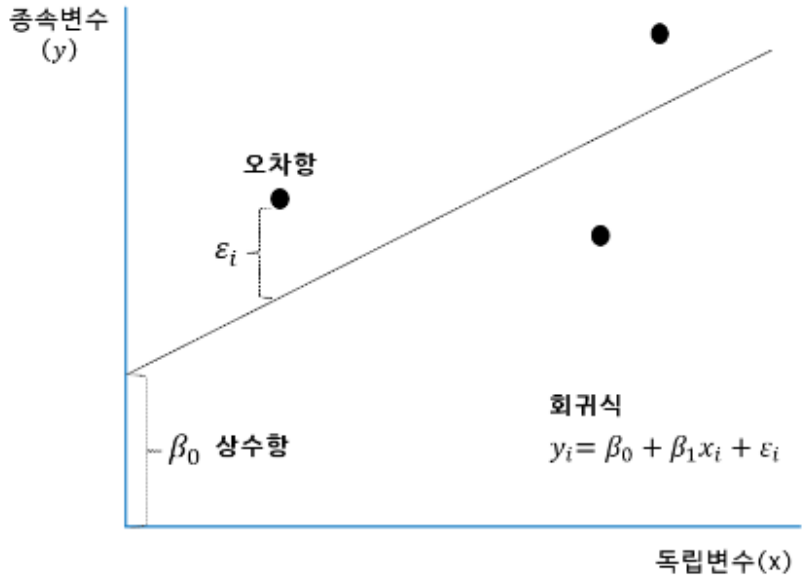
영향을 받는 변수 (y)	반응(Response)변수 = 종속(Dependent)변수 = 결과(Outcome)변수
영향을 주는 변수 (x)	설명(Explanatory)변수 = 독립(Independent)변수 = 예측(Predictor)변수

01. 회귀분석

◆ 단순선형회귀분석

✓ 하나의 독립변수가 종속변수에 미치는 영향을 추정할 수 있는 통계 기법

✓ 최소제곱법 : 측정값을 기초로 제곱합을 만들고 그것이 최소인 값을 구하여 처리하는 방법, 잔차제곱합이 가장 적은 선을 선택



주요 용어	설명
y_i	i번째 종속변수 값
x_i	i번째 독립변수 값
β_0	선형 회귀식의 절편(=상수항)
β_1	선형 회귀식의 기울기
ϵ_i	오차항, 독립적이며 $N(0,\ s^2)$ 분포

01. 회귀분석

◆ 선형회귀분석의 가정

가정	설명
선형성	✓ 입력변수와 출력변수의 관계가 선형이다.
등분산성	✓ 오차의 분산이 회귀식의 적합값(fiited values)와 무관하게 일정하다. ✓ 잔차플롯(산점도)을 활용하여 잔차와 회귀식의 적합값에 아무런 관련성이 없게 무작위적으로 고루 분포되어야 등분산성 가정을 만족한다.
독립성	✓ 입력변수와 오차는 관련이 없다. ✓ 자기상관(독립성)을 알아보기 위해 Durbin-Waston 통계량을 사용하며 주로 시계열 데이터에서 많이 활용
비상관성	✓ 오차들끼리 상관이 없음
정상성(정규성)	✓ 오차의 분포가 정규분포를 따른다. Q-Q Plot, Kolmogorov-Smirnov 검정, Shapiro-Wilk 검정

01. 회귀분석

◆ **회귀분석에서의 검토사항**

✓ **회귀계수들이 유의미한가?**

- 해당 계수의 $t - statistics$의 $p - value$이 0.05보다 작으면 해당 회귀계수가 통계적으로 유의하다고 볼 수 있다.

✓ **모형이 얼마나 설명력을 갖는가?**

- 결정계수(R^2)를 확인한다. 결정계수는 0~1 값을 가지며, 높은 가질수록 추정된 회귀식의 설명력이 높다.

01. 회귀분석

◆ 회귀분석에서의 검정

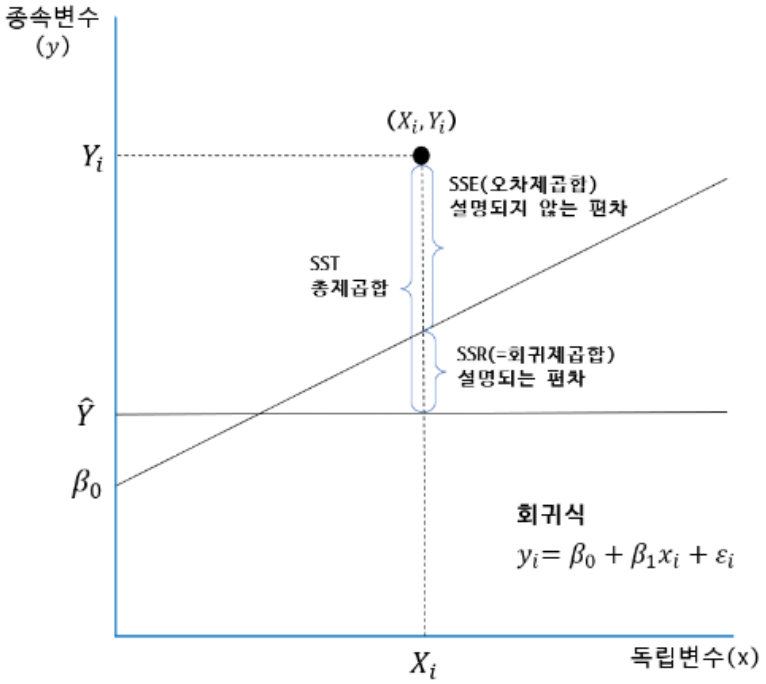
✓ 단순 회귀 분석의 결정계수는 상관계수 값의 제곱과 같음

검정	설명
F-검정	회귀식(모형)에 대한 검증
t-검정	회귀계수에 대한 검증
결정계수(R^2)	모형의 설명력은 1에 가까울수록 모델의 설명력이 높다고 할 수 있음 ($0 \leq R^2 \leq 1$)

01. 회귀분석

◆ 결정계수

✓ 회귀분석의 결정계수는 1에 가까울수록 모델의 설명력이 높음



주요 용어	설명
$\hat{Y}$	관찰값(데이터의 평균)
Y_i	관찰값
SST(Sum of Square Total)	종속변수값의 총 변동 $SST = SSR + SSE$
SSR(Sum of Square Regression)	회귀선에 의해 설명되는 변동
SSE(Sum of Square Error)	회귀선에 의해 설명되지 않는 변동

01. 회귀분석

◆ 다중선형회귀분석

- ✓ 두 개 이상의 독립변수가 하나의 종속변수에 미치는 영향을 추정하는 회귀분석
- ✓ 다중선형회귀식은 다음과 같음

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- ✓ 가설 설정

가설	설명	표현
귀무가설	각 회귀계수의 기울기는 0이다.	$\beta_1 = \beta_2 = \beta_3 \ldots = \beta_k = 0$
대립가설	적어도 한 개의 독립변수 회귀계수의 기울기는 0이 아니다.	$\beta_1 \neq \beta_2 \neq \beta_3 \ldots \neq \beta_k \neq 0$

01. 회귀분석

◆ 다중선형회귀분석

- ✓ 두 개 이상의 독립변수가 하나의 종속변수에 미치는 영향을 추정하는 회귀분석
- ✓ 다중선형회귀식은 다음과 같음

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- ✓ 가설 설정

가설	설명	표현
귀무가설	각 회귀계수의 기울기는 0이다.	$\beta_1 = \beta_2 = \beta_3 \ldots = \beta_k = 0$
대립가설	적어도 한 개의 독립변수 회귀계수의 기울기는 0이 아니다.	$\beta_1 \neq \beta_2 \neq \beta_3 \ldots \neq \beta_k \neq 0$

01. 회귀분석

◆ 다중선형회귀분석 모형의 통계적 유의성

- ✓ 모형의 통계적 유의성은 F-통계량으로 확인
- ✓ 유의수준 5% 하에서 F-통계량의 p-값이 0.05보다 작으면 추정된 회귀식은 통계적으로 유의하다고 볼 수 있음
- ✓ F 통계량이 크면 p-value가 작아지고 이렇게 되면 귀무가설을 기각한다. 즉, 모형이 유의하다고 결론지을 수 있음

변동요인	제곱합	자유도	평균 제곱	F값(f)
회귀(SSR)	회귀제곱합	k	$MSR = \frac{SSR}{k}$	$f = \frac{MSR}{MSE}$
오차(SSE)	오차제곱합	$n-k-1$	$MSE = \frac{SSE}{(n-k-1)}$	
전체변동(SST)	총 제곱합	$n-1$		

n: 관측값의 개수, k: 독립변수의 개수(단순회귀분석 : $k=1$)
MSE(Mean Square Error, 오차제곱평균) 산출, MSR($Mean\ Square\ Regression$, 회귀제곱평균) 산출

01. 회귀분석

◆ **다중선형회귀분석 회귀계수의 유의성**

- ✓ 회귀계수의 유의성은 단순회귀분석의 회귀계수 유의성 검토와 같이 회귀계수 t-통계량을 통해 확인한다.
- ✓ 모든 회귀계수의 유의성이 통계적으로 검증되어야 선택된 변수들의 조합으로 모형을 확인할 수 있음.

◆ **다중선형회귀분석 다중공선성(Multicollinearity)**

- ✓ 회귀분석에서 독립변수들 간에 강한 상관관계가 나타나는 문제
- ✓ 정확한 회귀계수의 추정이 어려워지며, 각 독립변수의 회귀계수가 종속변수에 미치는 영향력을 올바로 설명하지 못하게 됨

◆ **다중공선성 검사 방법**

- ✓ 분산팽창요인(VIF) : 4보다 크면 다중공선성이 존재한다고 볼 수 있고, 10보다 크면 심각한 문제가 있는 것으로 해석할 수 있음

01. 회귀분석

◆ **다중회귀분석의 변수 선택 방법**

- ✓ 이상적으로는 독립변수의 조합에 따라 만들어지는 선형모델 모두를 비교하여 최고의 모델 선정 가능.
- ✓ 변수가 k개인 경우, k개 변수들의 일부를 포함하는 총 모델의 수는 2^k개에 이르므로, 모델 모두를 고려하기는 현실적으로 불가능.
- ✓ 단계적 변수 선택(Stepwise Variable Selection) 방법 사용

◆ **단계적 변수 선택 방법의 유형**

구분	표현
전진선택법 (Forward Selection)	모델적합에 가장 큰 영향을 미치는 독립변수를 순서대로 추가하여 성능지표를 비교하면서 변수를 선택하는 방법. 한 번 선택한 변수는 제거하지 않음
후진선택법 (Backward Elimination)	전체 변수로 시작해서 모델적합에 가장 약하게 영향을 미치는 독립변수를 순서대로 제거해가며 성능지표를 비교하면서 변수를 선택하는 방법
단계적방법 (Stepwise Method)	전진선택법과 후진제거법을 병행하는 방법 추가된 변수에 의해 기존 변수의 중요도가 작아지면 추가된 변수를 제거