2강. Basic Methods for Regression 1

◈ 담당교수 : 김 동 하

■ 학습개요

이번 강의에서는 회귀모형의 기본이 되는 선형회귀모형에 대해 학습한다. 한 개의 독립변수를 사용하는 단순선형회귀모형과 복수 개의 독립변수를 사용하는 다중선형회귀모형에 대해 배운다. 데이터를 가장 잘 설명하는 모수를 찾기 위한 추정 방법 중 하나인 최소 제곱법에 대해서도 다루도록 한다. 더 나아가, 선형모형이 실제로 데이터를 잘 설명하고 있는지를 확인하기 위한 모형 적합성 검토 방법론에 대해서도 학습한다.

■ 학습목표

1	선형회귀분석에 대해 학습한다.
2	최소 제곱법을 통한 적합 방법에 대해 학습한다.
3	모형의 적합성 검토를 위한 다양한 방법에 대해 학습한다.

■ 주요용어

용어	해설
선형회귀분석	가장 단순한 회귀 모형으로 설명변수와 종속변수 사이의 관계를 설명변수의 선형식으로 규정하는 모형이다. 사용하는 변수의 개수에 따라 단순선형회귀모형, 다중선형회귀모형으로 구분할 수 있다.
최소 제곱법	선형회귀모형의 모수를 추정하는 방식 중 하나. 종속변수와 독립변수의 선형식의 차에 대한 제곱합을 최소로 하는 모수를 찾는다.
선형회귀모형 적합성 검토	선형 모형의 사용이 실제로 데이터에 적합한지를 테스트하기 위한 방법. 선형성, 등분산성, 정규성, 독립성을 체크해야 하며, 각 가정마다 적절한 검정 방법들을 이용하여 검토한다.

■ 학습하기

‘

01. 선형회귀분석

회귀분석이란?

- 종속변수가 독립변수들에 의해 어떻게 설명되는지를 분석하는 통계적 기법.
- 단순선형회귀분석: 한 개의 설명 변수의 선형 함수로 종속변수를 설명.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- 다중회귀분석: 설명 변수가 두 개 이상

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- 다항회귀분석: 설명변수들의 교차 영향이나 다항 영향을 고려.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \cdots + \epsilon$$

질적 설명변수 처리

- 가변수 (Dummy variable) 활용: 범주가 n개 있는 경우 (n-1)개의 가변수를 사용하여 해당 변수를 표현할 수 있다.
- 예:대학교 학년 설명 변수 (1학년~4학년)

1학년	(1,0,0)
2학년	(0,1,0)
3학년	(0,0,1)
4학년 (Reference 변수)	(0,0,0)

선형 회귀식의 추정

- 최소 제곱법: 선형 회귀식에서 절편항과 기울기를 나타내는 숫자를 '모수'라 부른다.
- 주어진 데이터를 잘 설명하는 '모수'를 잘 추정하는 것이 중요.
- 데이터와 모형의 예측값 사이의 오차 제곱합을 최소로 하는 모수를 추정하는 방법.
- 주어진 데이터: $D = (x_1, y_1), \cdots (x_n, y_n)$

모형	단순 선형회귀모형	다중 선형회귀모형
회귀식	$Y = \beta_0 + \beta_1 X + \epsilon$	$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$
오차제곱합	$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$	$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{ip})^2$
예측	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots \hat{\beta}_p x_p$

선형회귀모형을 이용해 예측하기

- 예: 티비 광고 횟수 (X)를 통해 상품 판매량 (Y) 예측하기.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 10 + 20x$$

- 티비 광고 횟수가 5회일 때 $(X = 5)$, 상품 판매량은 $10 + 20*5 = 110$으로 예측할 수 있다.

단순선형회귀모형 적합하기

- Sale 데이터를 이용하여 단순선형회귀 모형을 적합해보자.
- 필요한 패키지 불러오기

```
import os
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf
```

- Sale 데이터 불러오기
  > Adver: 광고량
  > Sales: 상품의 판매량

```
data_file = "./data/Sales.csv"
Sales = pd.read_csv(data_file)
Sales.iloc[0:5]
```

	Company	Adver	Sales
0	1	11	23
1	2	19	32
2	3	23	36
3	4	26	46
4	5	56	93

- 단순선형회귀분석 적합하기

```
## 단순선형회귀분석 적합
SalesFit = smf.ols(formula='Sales~Adver', data=Sales).fit()
print(SalesFit.summary())
```

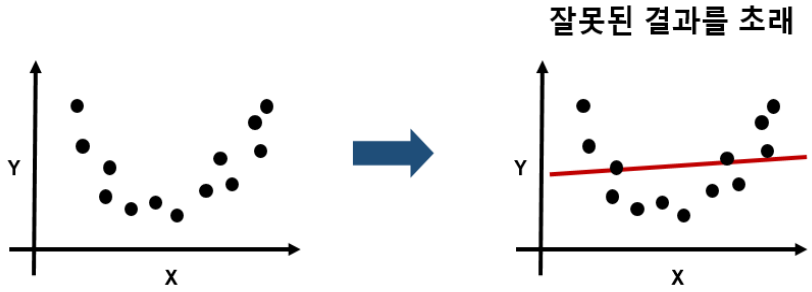
- 결과 확인하기

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.979
Model:                            OLS   Adj. R-squared:                  0.976
Method:                 Least Squares   F-statistic:                     455.5
Date:                Mon, 23 May 2022   Prob (F-statistic):           1.14e-09
Time:                        13:17:56   Log-Likelihood:                -32.059
No. Observations:                  12   AIC:                             68.12
Df Residuals:                      10   BIC:                             69.09
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.2848      2.889      1.137      0.282      -3.153       9.723
Adver          1.5972      0.075     21.343      0.000       1.430       1.764
==============================================================================
Omnibus:                        0.879   Durbin-Watson:                   2.470
Prob(Omnibus):                  0.644   Jarque-Bera (JB):                0.379
Skew:                           0.419   Prob(JB):                        0.828
Kurtosis:                       2.768   Cond. No.                         101.
==============================================================================
```

02. 잔차분석

잔차분석이란?

- 모형 적합성 검토: 데이터가 실제로 선형회귀모형을 따르는지 확인할 필요가 있음.



선형회귀모형의 가정

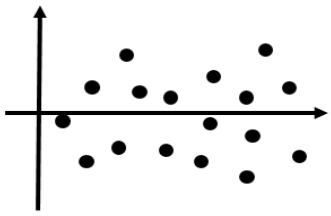
- 선형회귀모형은 다음과 같은 네 가지의 가정을 필요로 함.
  > 선형성
  > 오차항 ϵ의 등분산성
  > 오차항 ϵ의 정규성
  > 오차항 ϵ의 독립성
- 이를 확인하기 위해서 잔차 $\hat{\epsilon} = y - \hat{y}$를 활용하여 검토.

오차항에 대한 검토

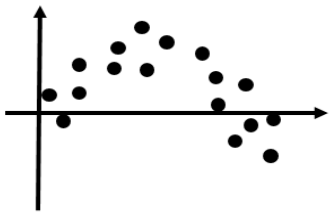
- 선형성 검토 방법의 예
  > 잔차산점도 이용
- 등분산성 검토 방법의 예
  > 잔차산점도 이용, Bruesch-Pagan 검정
- 정규성 검토 방법의 예
  > 정규확률 그림 (Q-Q plot), Jarque-Bera 검정
- 독립성 검토 방법의 예
  > Durbin-Watson 검정

선형성과 등분산성 검토: 잔차 산점도

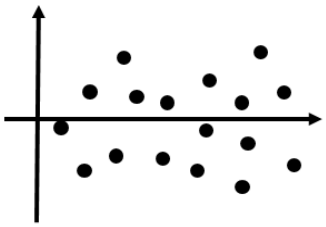
- x축에는 적합값, y축에는 스튜던트화 잔차를 그린 산점도.
- 스튜던트화 잔차가 -2에서 2 사이에서 랜덤하게 흩어져 분포해 있으면 선형성과 등분산성 가정을 만족하는 것으로 간주할 수 있음.



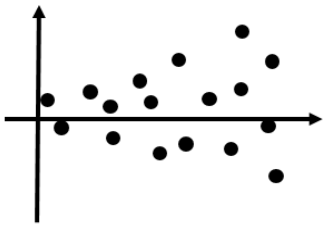
선형성 만족 O



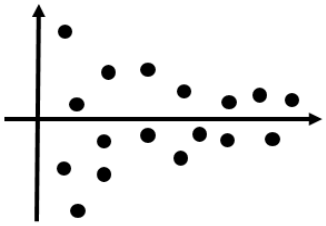
선형성 만족 X



등분산성 만족 O



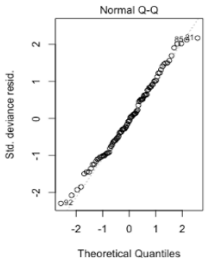
등분산성 만족 X



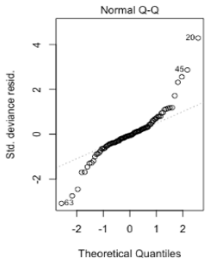
등분산성 만족 X

정규성 검토: 정규확률 그림 (Q-Q plot)

- 정규성을 확인하기 위해 그리는 산점도.
- 점들이 직선 위에 가깝게 분포하고 있으면 정규성을 따르는 것으로 간주할 수 있음.



정규성 만족 O



정규성 만족 X

독립성 검토: Durbin-Watson 검정

- 더빈왓슨 통계량을 사용.
  > 항상 0과 4 사이의 값을 가짐.
  > 2에 가까울수록 독립성을 만족하는 것으로 생각할 수 있음. 반대로, 2에서 멀어질수록 독립성 가정을 만족하지 않는 것으로 판단할 수 있음. d=

■ 연습문제

(객관식)1. 회귀 분석에 대한 설명으로 잘못된 것을 고르시오.

① 종속변수가 독립변수들에 의해 어떻게 설명되는지를 분석하는 통계적 기법이다.
② 종속 변수를 반응 변수라고도 하며 독립변수들에 의해 설명되는 변수를 의미한다..
③ 오차항은 회귀식으로는 설명할 수 없는 랜덤 성분이며, 선형회귀모형의 경우 대체로 라플라스 분포를 가정한다.
④ 단순선형회귀분석의 경우 모수는 절편항과 기울기 두 개이다.

정답 : ③
해설 : 오차항은 주로 정규분포를 가정한다.

(단답형)2. 설명 변수 중 하나가 총 5가지의 수준을 가질 수 있는 질적 변수라 하자. 이를 회귀식에 포함하기 위해서는 몇 개의 가변수가 필요한가?

정답) 4개
해설) 질적 변수가 가질 수 있는 수준의 수보다 하나 작은 가변수가 필요하다.

(단답형)3. 모형 적합성 검토에서 더빈-왓슨 검정을 통해 확인할 수 있는 선형회귀모형의 가정은 무엇인가?

정답 : 독립성
해설 : 해설 없음.

■ 정리하기

1. 회귀모형은 종속변수가 독립변수들에 의해 어떻게 설명되는지를 분석하는 기법으로, 가장 간단한 모형으로는 선형식을 활용한 선형회귀모형이 있다.

2. 선형회귀모형에 필요한 모수들은 제곱합을 최소로 하는 최소 제곱법을 이용하여 추정할 수 있다.

3. 선형모형이 실제로 데이터에 잘 적합하는지를 확인하기 위해 모형 적합성 검토가 필요하다. 선형성, 등분산성, 정규성, 독립성을 확인해야 하며, 이를 위해서 잔차 산점도, Q-Q plot, 더빈-왓승 검정 등이 사용된다.

■ 참고자료 (참고도서, 참고논문, 참고사이트 등)

1. 이성건,강현철, 『파이썬을 활용한 데이터 분석과 응용』, 자유아카데미(2021)
=> 파이썬을 활용한 데이터 분석과 응용은 통계 방법론의 개념과 파이썬을 이용한 구현에 대해 자세히 서술하고 있다.