머신러닝응용 제02강

Basic Methods for Regression 1

첨단공학부 김동하교수

제02강 Basic Methods for Regression 1

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

핵심 단어

- 선형회귀분석
- 최소 제곱법
- 모형 적합성 검토

02강. Basic Methods for Regression 1

01. 선형회귀분석

1) 회귀분석이란

회귀분석

- 종속변수가 독립변수들에 의해 어떻게 설명되는지를 분석하는 통계적 기법

종속변수 (Y)

반응(Response) 변수

독립변수들에 의해 설명되는 변수

독립변수 (X)

설명(Explanatory) 변수

설명에 이용되는 변수

2) 선형회귀분석

단순선형회귀분석

- 한 개의 설명변수의 선형 함수로 종속변수를 설명.
- 예시

$$\underline{\text{음료수판매량}(Y)} = \beta_0 + \beta_1 * \underline{\text{기온}(X)} + \epsilon$$

종속변수 (Y)

독립변수 (X)

2) 선형회귀분석

단순선형회귀분석

- 예시

$$\text{음료수판매량}(Y) = \beta_0 + \beta_1 * \text{기온}(X) + \epsilon$$

절편항 (β_0)

설명변수에 영향을 받지 않는 값

입력값이 0일 때 종속 변수의

기대값

기울기 (β_1)

설명변수 X가 한 단위 증가할

때마다 증가하는 종속 변수의 양

오차항 (ϵ)

회귀식으로는 설명할 수 없는

랜덤 성분

정규 분포를 가정

$(\epsilon \sim N(0, \sigma^2))$

2) 선형회귀분석

다양한 선형회귀분석

- **다중**회귀분석: 설명 변수가 두 개 이상

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- **다항**회귀분석: 설명변수들의 교차 영향이나 다항 영향 고려

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \cdots + \epsilon$$

2) 선형회귀분석

설명 변수가 질적 변수일 때의 처리 방법

- 가변수 활용 (Dummy variables)
  - 범주가 n개 있는 경우 (n-1)개의 가변수를 사용하여 해당 변수를 표현할 수 있다.

2) 선형회귀분석

◆ 설명 변수가 질적 변수일 때의 처리 방법

- 예: 대학교 학년 설명 변수 (1~4)

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

3) 선형회귀식의 추정

최소 제곱법

- 선형 회귀식에서 절편항과 기울기를 '모수'라 부른다.
- 주어진 데이터를 잘 설명하는 '모수'를 잘 추정하는 것이 중요.
- 최소 제곱법을 활용.
  - 데이터와 모형의 예측값 사이의 오차 제곱합을 최소로 하는 모수를 추정하는 방법.

3) 선형회귀식의 추정

최소 제곱법

- 주어진 데이터
  - $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$

모형	단순 선형회귀모형	다중 선형회귀 모형
회귀식	$Y = \beta_0 + \beta_1 X + \epsilon$	$Y = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + \epsilon$
오차제곱합	$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$	$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$
예측	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$

4) 예측하기

선형 회귀모형을 이용하여 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 상품 판매량은 10+20*5=110으로 예측할 수 있다.

5) 단순선형회귀모형 적합

Sale 데이터를 이용한 단순선형회귀분석

- 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
```

5) 단순선형회귀모형 적합

Sale 데이터를 이용한 단순선형회귀분석

- 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

5) 단순선형회귀모형 적합

Sale 데이터를 이용한 단순선형회귀분석

- 적합하기

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

5) 단순선형회귀모형 적합

◆ Sale 데이터를 이용한 단순선형회귀분석

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.

5) Python을 이용한 실습

Sale 데이터를 이용한 단순선형회귀분석

- 적합된 모형을 이용하여 예측값 살펴보기

```
## 적합된 모형을 이용한 적합값 및 신뢰구간
predictions = SalesFit.get_prediction()
predictions.summary_frame(alpha=0.05).round(3).iloc[0:3]
```

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	20.854	2.152	16.058	25.649	11.057	30.650
1	33.631	1.667	29.916	37.346	24.316	42.946
2	40.019	1.457	36.773	43.266	30.881	49.158

5) Python을 이용한 실습

- Satisfaction 데이터를 이용한 다중선형회귀분석
  - Satisfaction 데이터
    - 700명의 고객을 대상으로 특정 제품에 대해 조사.
      - X1: 디자인 만족도
      - X2: 사용 편리성 만족도
      - X3: 성능 만족도
      - X4: 고장 및 경고성 만족도
      - Gender: 성별 (1: 남자, 2: 여자)
      - Age: 나이 (1: 10대, 2: 20대, …)
      - Y 구입 의향 점수

5) Python을 이용한 실습

- Satisfaction 데이터를 이용한 다중선형회귀분석
  - 데이터 불러오기
    - 500개, 200개로 나누어 각각을 훈련, 시험 자료로 사용.

```
data_file = "./data/Satisfaction.csv"
Satisfaction = pd.read_csv(data_file)

Tr_Sat = Satisfaction.iloc[1:500,:]
Ts_Sat = Satisfaction.iloc[500:700,:]
```

5) Python을 이용한 실습

Satisfaction 데이터를 이용한 다중선형회귀분석

- 모형 적합하기
  - Age, Gender는 범주형 변수로 취급

```
SatFit = smf.ols(formula='Y~X1+X2+X3+X4+C(Age)+ \
                          C(Gender)', data=Tr_Sat).fit()
```

5) Python을 이용한 실습

Satisfaction 데이터를 이용한 다중선형회귀분석

- 적합된 모형의 추정 계수 정보만 따로 추출해보자.

```
## 적합 모형의 추정 계수 정보만 따로 추출
print(SatFit.params.round(5))
```

```
Intercept          1.68115
C(Age)[T.2]       -0.27592
C(Age)[T.3]       -0.33035
C(Age)[T.4]       -0.13041
C(Age)[T.5]       -0.06705
C(Gender)[T.2]     0.19021
X1                 0.12487
X2                 0.05236
X3                 0.38463
X4                 0.06871
```

5) Python을 이용한 실습

Satisfaction 데이터를 이용한 다중선형회귀분석

- 모형 적합하기
  - Treatment(reference='변수 이름') 으로 reference 변수를 설정할 수 있음.
  - 따로 입력하지 않은 경우 첫번째 값이 reference 변수가 됨.

```
SatFit2 = smf.ols(formula='Y~X1+X2+X3+X4+\
                  C(Age, Treatment(reference=5))+ \
                  C(Gender, Treatment(reference=2))', \
                   data=Tr_Sat).fit()
```

5) Python을 이용한 실습

Satisfaction 데이터를 이용한 다중선형회귀분석

- 새롭게 적합된 모형의 추정 계수 정보만 따로 추출해보자.

```
print(SatFit2.params.round(5))
```

```
Intercept                                        1.80431
C(Age, Treatment(reference=5))[T.1]              0.06705
C(Age, Treatment(reference=5))[T.2]             -0.20886
C(Age, Treatment(reference=5))[T.3]             -0.26329
C(Age, Treatment(reference=5))[T.4]             -0.06335
C(Gender, Treatment(reference=2))[T.1]          -0.19021
X1                                               0.12487
X2                                               0.05236
X3                                               0.38463
X4                                               0.06871
```

5) Python을 이용한 실습

Satisfaction 데이터를 이용한 다중선형회귀분석

- Test set에 적용하여 예측값을 알아보자.

```
## 예측해보기
predictions = SatFit.get_prediction(Ts_Sat)
predictions.summary_frame(alpha=0.05).round(3).iloc[0:3]
```

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	4.817	0.136	4.549	5.085	2.494	7.140
1	4.585	0.182	4.228	4.943	2.250	6.921
2	5.324	0.146	5.037	5.612	2.999	7.650

01강. Data handling with Python

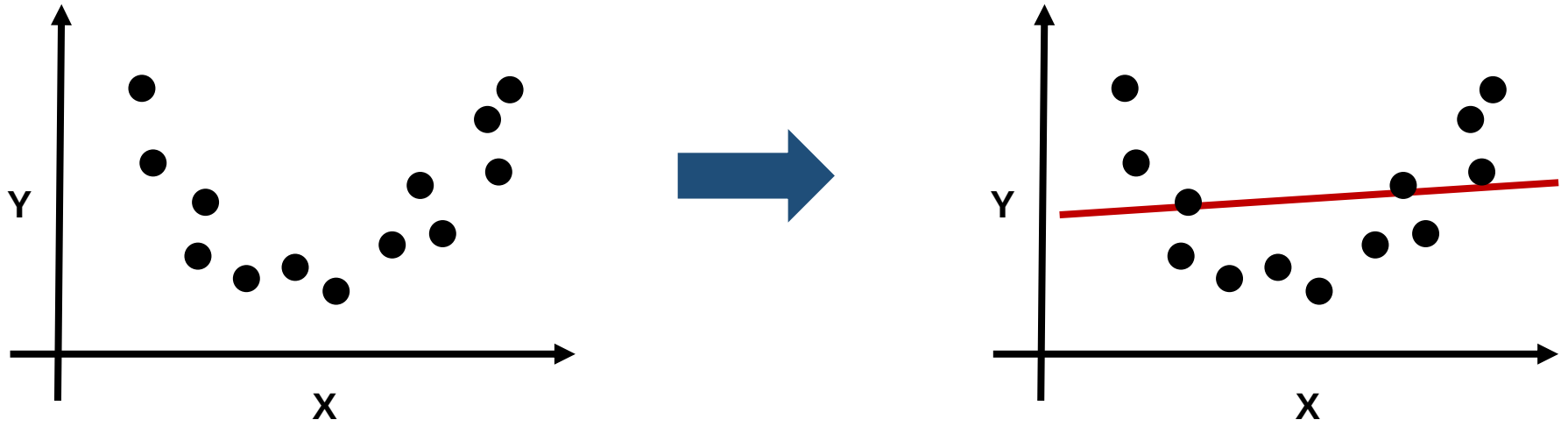
02. 잔차분석

1) 잔차분석이란

모형 적합성 검토

- 데이터가 실제로 선형회귀모형을 따르는지를 확인할 필요가 있음.
  - 잔차분석 활용

잘못된 결과를 초래



2) 오차항에 대한 검토

- **선형회귀모형에서의 가정**
  - 선형성
  - 오차항 ϵ 의 등분산성
  - 오차항 ϵ 의 정규성
  - 오차항 ϵ 의 독립성

- **이를 확인하기 위해 잔차($\hat{\epsilon}$)를 활용하여 검토.**
  - $\hat{\epsilon} = y - \hat{y}$

2) 오차항에 대한 검토

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

3) 선형성과 등분산성 검토

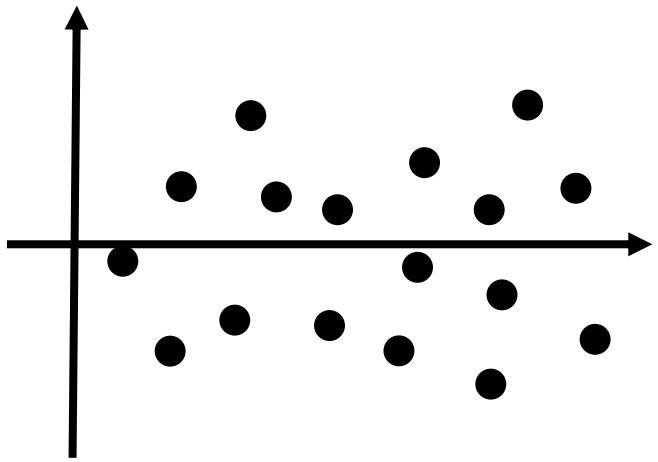
잔차 산점도

- x축에는 적합값, y축에는 스튜던트화 잔차를 그린 산점도.

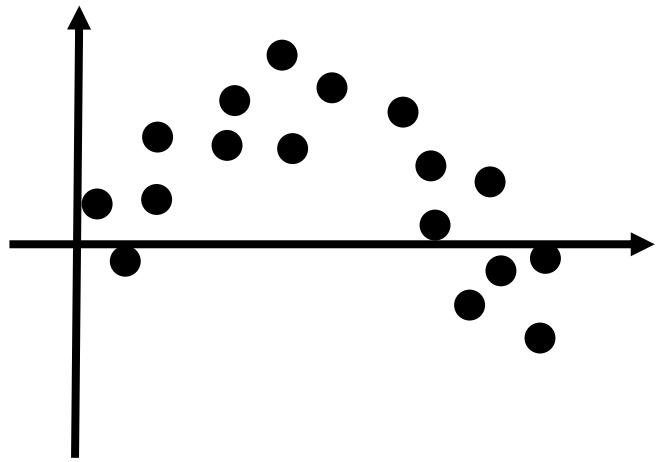
- 스튜던트화 잔차가 -2~2 사이에서 랜덤하게 흩어져 있으면 선형성과 등분산성 가정을 만족하는 것으로 생각.

3) 선형성과 등분산성 검토

◆ 잔차 산점도



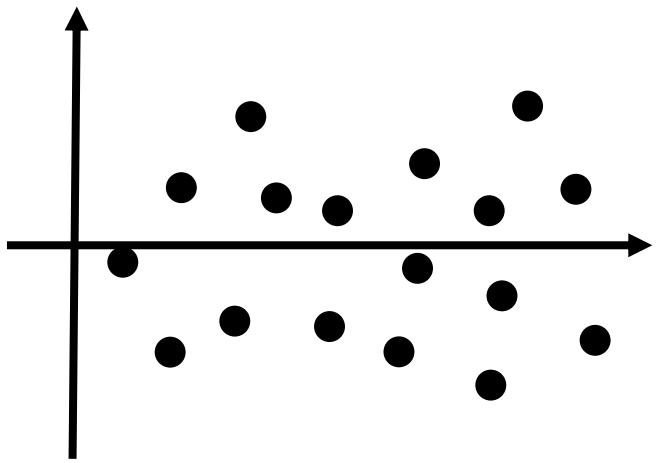
선형성 만족 O



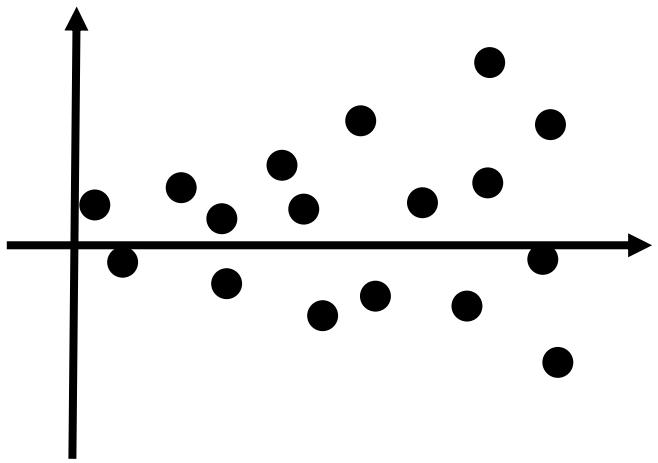
선형성 만족 X

3) 선형성과 등분산성 검토

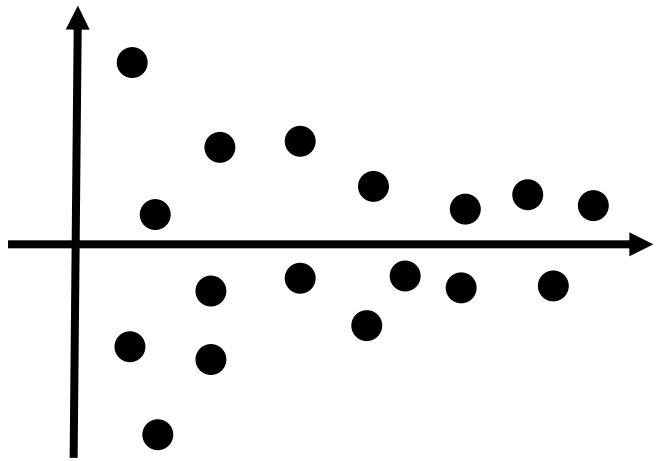
◆ 잔차 산점도



등분산성 만족 O



등분산성 만족 X



등분산성 만족 X

4) 정규성 검토

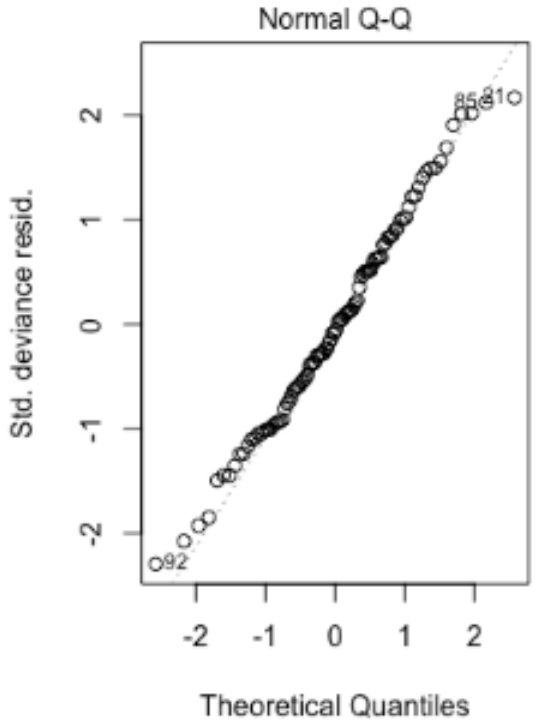
정규확률 그림 (Q-Q plot)

- 정규성을 확인하기 위해 그리는 산점도.

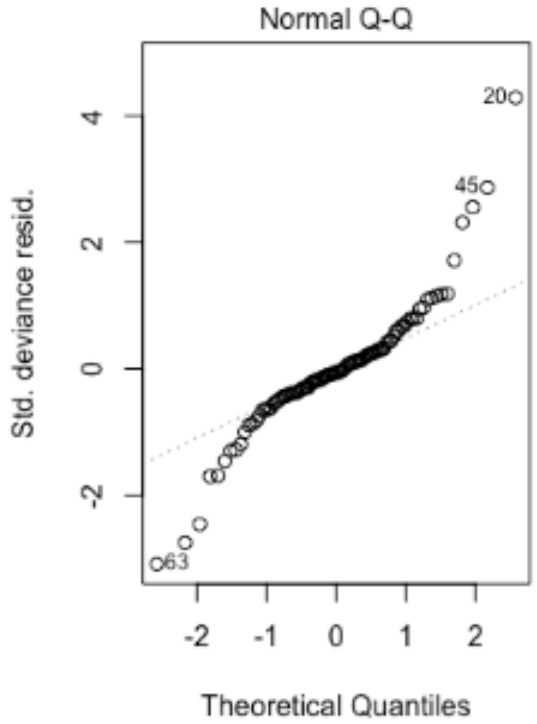
- 점들이 직선 위에 가깝게 분포하고 있으면 정규성을 따르는 것으로 생각.

4) 정규성 검토

◆정규확률 그림 (Q-Q plot)



정규성 만족 O



정규성 만족 X

출처: https://www.seascapemodels.org/rstats/2017/10/06/qqplot-non-normal-glm.html

5) 독립성 검토

- Durbin-Watson 검정
  - 더빈 왓슨 통계량을 사용.
    - 항상 0~4 사이의 값을 가짐
    - 2에 가까울 수록 독립성을 만족
    - 2에서 멀어질수록 독립성 가정을 만족하지 않는 것으로 판단

6) Python을 이용한 실습

잔차 계산하기

- SalesFit 이용 (단순선형회귀 결과)

```
Fitted = SalesFit.predict()
Residual = SalesFit.resid
RStandard = SalesFit.resid_pearson

## 잔차, 표준화잔차의 탐색
print(pd.DataFrame({'Fitted': Fitted, 'Residual': Residual, \
                    'RStandard': RStandard}).iloc[0:5])
```

6) Python을 이용한 실습

◆ 잔차 계산하기

- SalesFit 이용 (단순선형회귀 결과)

	Fitted	Residual	RStandard
0	20.853505	2.146495	0.559896
1	33.630747	-1.630747	-0.425367
2	40.019368	-4.019368	-1.048420
3	44.810833	1.189167	0.310185
4	92.725489	0.274511	0.071604

6) Python을 이용한 실습

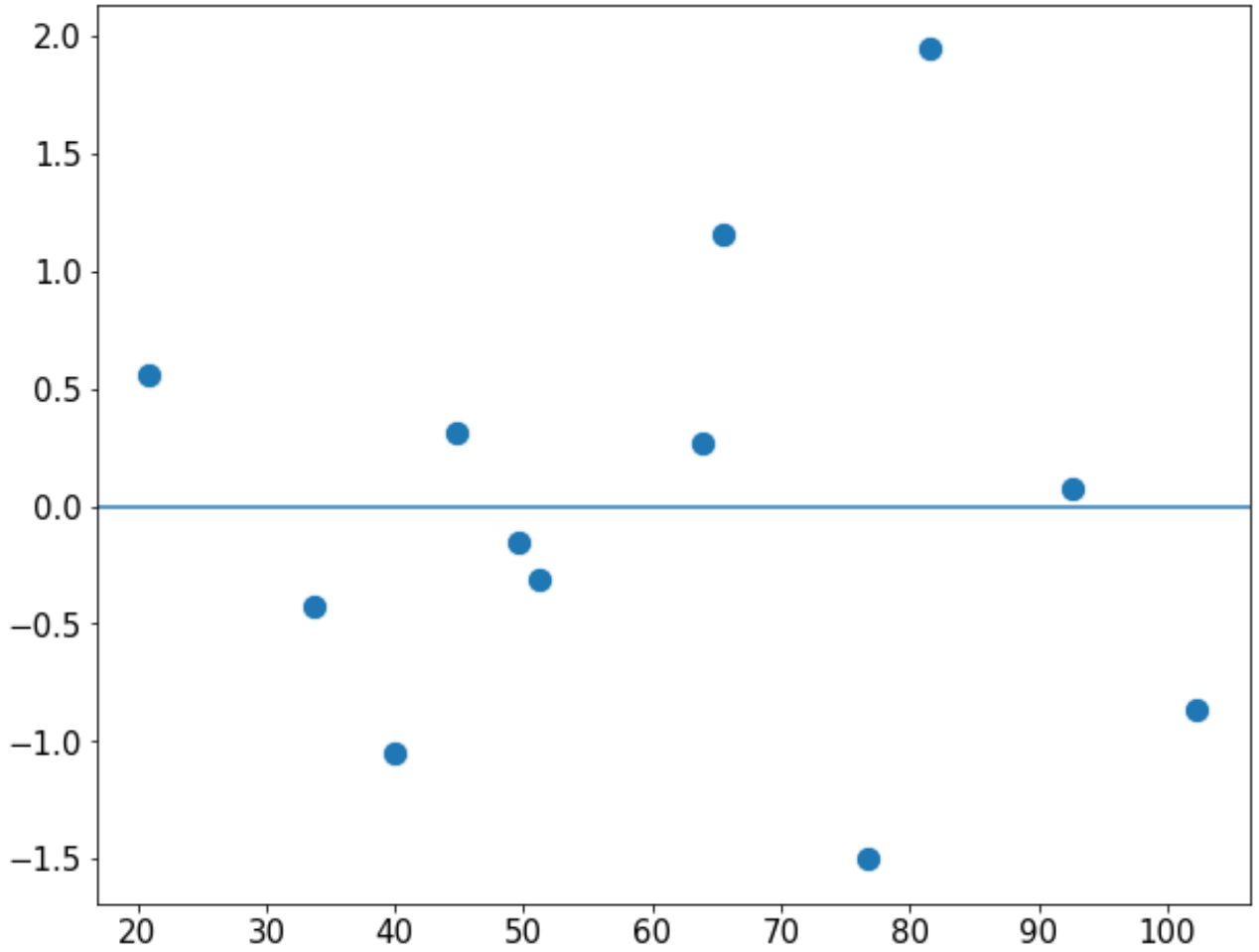
잔차 산점도 그리기 (선형성, 등분산성 검토)

- x축: 적합값
- y축: 표준화 잔차

```
## 잔차 도표
fig, ax = plt.subplots(figsize=(10,8))
sns.scatterplot(x=Fitted,y=RStandard, s=150)
ax.xaxis.set_tick_params(labelsize=15)
ax.yaxis.set_tick_params(labelsize=15)
ax.axhline(y=0)
```

6) Python을 이용한 실습

- 잔차 산점도 그리기 (선형성, 등분산성 검토)
  - -2~2 사이에 랜덤하게 분포하고 있음을 확인할 수 있음.
  - 선형성, 등분산성 O



6) Python을 이용한 실습

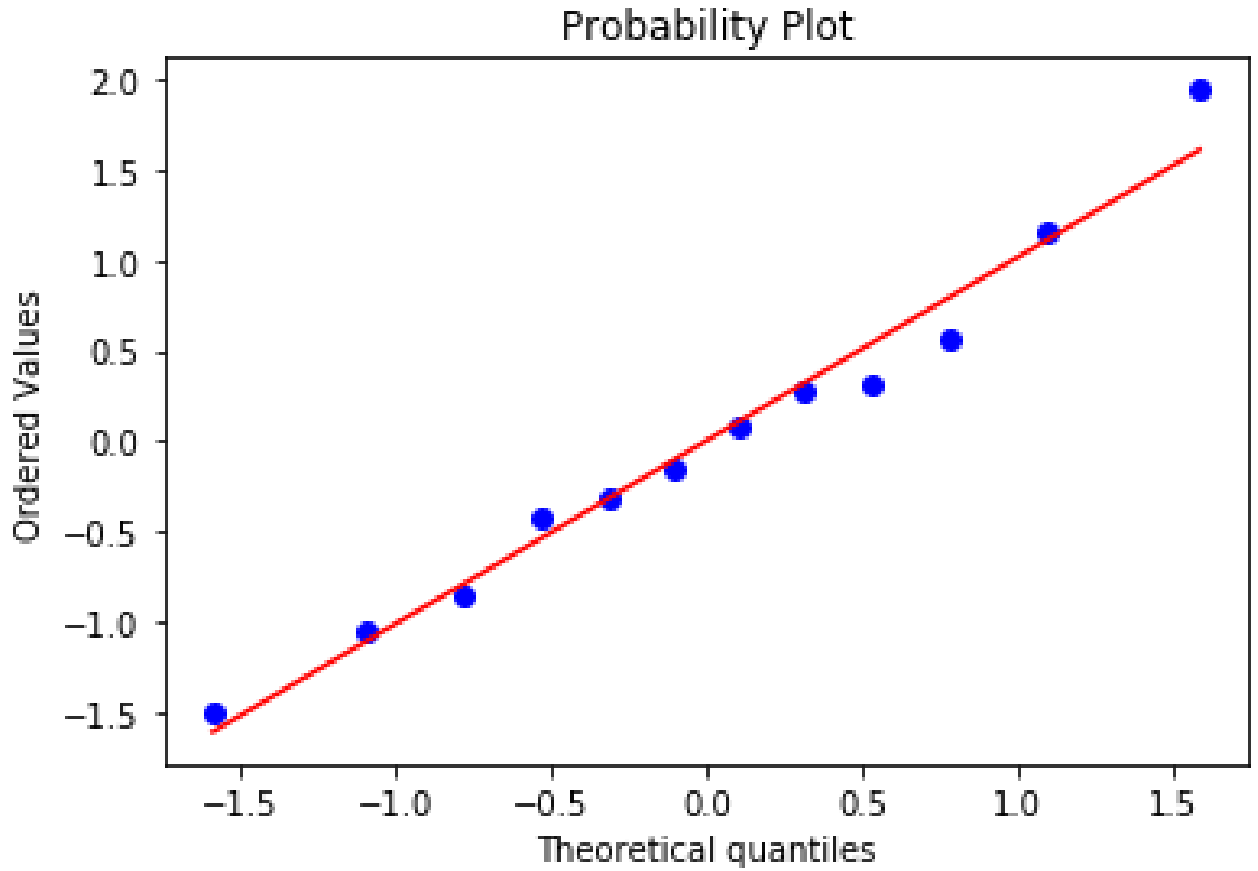
Q-Q plot 그리기 (정규성 검토)

```
from scipy.stats import probplot
## Q-Q plot
probplot(RStandard, plot=plt)
```

6) Python을 이용한 실습

Q-Q plot 그리기 (정규성 검토)

- 점들이 직선 위에 분포
- 정규성 만족함을 알 수 있음.



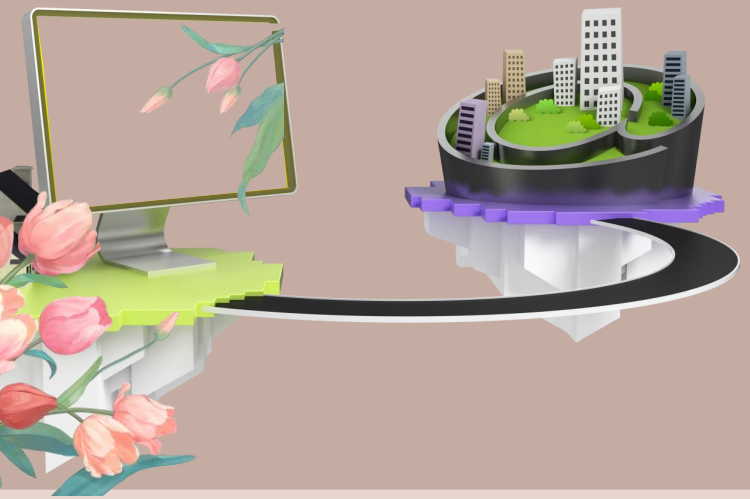
6) Python을 이용한 실습

더빈-왓슨 통계량 (독립성 검토)

- 2에 가까운 수치를 갖는 것으로 보아 독립성을 만족하는 것을 알 수 있음.

```
from statsmodels.stats.stattools import durbin_watson
durbin_watson(RStandard).round(5)
```

```
2.47031
```



다음시간안내

제03강

Basic Methods for Regression 2.