파이썬 데이터 분석 실무과정

이 선 순

범주형 자료분석

적합도 검정(Goodness of fit test)

- 범주형 자료(categorical data)
  - 관측값이 몇 가지 속성(범주, category) 분류하고 각 속성별 도수(빈도)로 나타낸 자료
  - 예 : 선거에서의 정당별 득표수, 요일별 교통사고 건수, 범죄형태별 범죄발생횟수 등

▸ **적합도 검정 : 도수분포표의 형태로 나타낸 자료들이 이론적으로 가정된 모형과 일치하는지에 대한 검정**

- 1차원 빈도표 : 크기 n의 표본을 상호배반인 r개의 범주로 분류하고 범주별 관측빈도($O_1, O_2, \cdots, O_r$)를 얻었다.(단, $\sum_{i=1}^{r} O_i = n$)

범주	1	2	$\cdots$	r	계
관측도수	O_1	O_2	$\cdots$	O_r	n
기대도수	E_1	E_2	$\cdots$	E_r	n

- 위 분할표의 자료들이 귀무가설하에서 이론으로 가정된 확률모형과 일치하는지를 검정
  - 귀무가설 $H_0 : p_1 = p_{10}, \, p_2 = p_{20}, \, \cdots, \, p_r = p_{r0}$ 이라고 한다면 H_0하에서 기대도수(Expected Frequency) E_i를 계산

    $E_i = (\text{표본크기}) \times (H_0\text{하에서의 추정확률}) = np_{i0}$

- 범주의 관측도수가 귀무가설 H_0하에서 기대도수와 차이가 많이나면 H_0를 기각

▸ **적합도 검정 : r개의 다항모집단에 대한 비교(표본크기가 큰 경우 : $E_i \ge 5$)**

- 가설 $H_0 : p_1 = p_{10}, p_2 = p_{20}, \cdots, p_r = p_{r0}$ $H_1 : H_0$가 사실이 아니다.

- 검정통계량 : $\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i} (\sim \chi^2(r-1)\ under\ H_0)$

  검정통계량의 관측값 계산 : χ_0^2

- 유의확률 : $P(\chi^2 \ge \chi_0^2)$: 유의수준 α와 비교

▸ **예. 완두콩의 교배실험을 통해서 얻은 것으로, 총 556개 중 둥글고 노란 완두콩(1), 둥글고 녹색인 완두콩(2), 모나고 노란 완두콩(3), 모나고 녹색인 완두콩(4)이 다음과 같이 관측되었다.**
이 실험의 결과가 멘델의 유전법칙 9 : 3 : 3 : 1 을 따르는지 유의수준 $\alpha = 0.05$에서 알아보자.

완두콩의 형태	둥노(1)	동녹(2)	모노(3)	모녹(4)	계
관측도수	315	108	101	32	556
관측비율	56.7%	19.4%	18.2%	5.8%	100%

- 가설 $H_0 :\ p_1 = \frac{9}{16},\ p_2 = \frac{3}{16},\ p_3, = \frac{3}{16},\ p_4 = \frac{1}{16}$

  $E_1 = n \times p_{10} = 556 \times \frac{9}{16} = 312.75,\ E_2 = n \times p_{20} = 556 \times \frac{3}{16} = 104.25,\ E_3 = n \times p_{30} = 556 \times \frac{3}{16} = 104.25,\ E_4 = n \times p_{40} = 556 \times \frac{1}{16} = 34.75$

- chisquare함수(scipy.stats 모듈) : 일차원 빈도표에 대하여 적합도 검정을 위한 카이제곱 검정 수행

<table>
<tr>
<td>

```
import numpy as np
Pea = np.array([315,108,101,32])
P0 = np.array([9/16,3/16,3/16,1/16])*np.sum(Pea)
print(P0)


from scipy.stats import chisquare
chisquare(Pea,P0)
```

</td>
<td>

▶ 기대도수 계산

$E_i = (\text{표본크기}) \times (H_0\text{하에서의 추정확률}) = np_{i0}$

▶ chisquare(f_obs=, f_exp=, ddof=) : 완두콩의 관측도수와 기대도수를 구하여 카이제곱 검정 수행

- f_obs= : 각 범주별 관측도수 지정
- f_exp= : 각 범주별 기대도수 지정, 이 값이 지정되지 않으면 모든 범주의 비율이 동일하게 가정
- ddof(delta degrees of freedom) : p-값 계산시 카이제곱분포의 자유도 보정에 사용되는 값, 기본값은 ddof=0

</td>
</tr>
</table>

```
In [8]: import numpy as np

In [9]: Pea = np.array([315,108,101,32])

In [10]: P0 = np.array([9/16,3/16,3/16,1/16])*np.sum(Pea)

In [11]: print(P0)
[312.75 104.25 104.25  34.75]

In [12]: from scipy.stats import chisquare

In [13]: chisquare(Pea,P0)
Out[13]: Power_divergenceResult(statistic=0.4700239808153477, pvalue=0.925425895103616)
```

▶ 가설 $H_0: \ p_1 = \frac{9}{16}, \ p_2 = \frac{3}{16}, \ p_3, = \frac{3}{16}, \ p_4 = \frac{1}{16}$ $\quad H_1 : H_0$ 가 사실이 아니다.

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} (\approx \chi^2(3) \ \ under \ \ H_0)$$

카이제곱 검정 결과, 검정통계량의 관측값은 $\chi_0^2 = 0.4700$, 유의확률($p-value$)은 0.9254로 유의수준 5%하에서 귀무가설을 기각할 수 없다. 즉, 관측된 완두콩들의 비율은 멘델의 유전법칙을 따른다고 할 수 있다.

▸ **예. 어느 공항 주차장에는 5개의 출입구가 있다. 일정시간 동안에 각 입구를 주차장에 들어온 자동차의 수는 아래 표와 같다. 주차장에 들어온 자동차의 수는 주차장 출입구의 위치와 관계없이 동일한지를 유의수준 $\alpha = 0.05$하에서 검정하시오.**

적합도 검정통계량 계산표

주차장 입구	입구A	입구B	입구C	입구D	입구E
자동차수	238	216	302	294	199

```
import numpy as np
x = np.array([238,216,302,294,199])
np.sum(x)
E = np.array([1/5,1/5,1/5,1/5,1/5])*np.sum(x)

from scipy.stats import chisquare
chisquare(x,E)
```

독립성 검정

▸ **독립성 검정 : $r \times c$ 분할표(2차원 도수분포표)의 형태로 나타낸 두 변수간에 연관성이 있는지에 대한 검정**

$A \backslash B$	B_1	B_2	…	B_c	계
A_1	O_{11}	O_{12}	…	O_{1c}	$O_{1.}$
A_2	O_{21}	O_{22}	…	O_{2c}	$O_{2.}$
… … …	…	…	…	…	…
A_r	O_{r1}	O_{r2}	…	O_{rc}	$O_{r.}$
합계	$O_{.1}$	$O_{.2}$		$O_{.c}$	n

- 가설 H_0 : 행변수와 열변수는 서로 독립이다. vs H_1 : 행변수와 열변수는 서로 독립이 아니다.
- 검정통계량 : $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} (\approx \chi^2(r-1)(c-1) \ \ under \ \ H_0)$

  검정통계량의 관측값 계산 : χ_0^2
- 유의확률 $P = P(\chi^2 \geq \chi_0^2)$: 유의수준 α와 비교

▸ **예. (cars93.csv) cars93의 자동차제조국(Origin)과 자동차의 타입(Type) 변수사이에는 연관성이 있는지 유의수준 $\alpha = 0.05$에서 알아보자.**

- 가설
  - 귀무가설 H_0 : 자동차 제조국과 자동차 타입은 서로 독립이다
  - 대립가설 H_1 : 자동차 제조국과 자동차 타입은 서로 독립이 아니다.

- chi2_contingency함수(scipy.stats 모듈) : 분할표에 대하여 독립성 검정을 위한 카이제곱 검정 수행

```
import pandas as pd
cars93 = pd.read_csv("d:/data/cars93.csv")

table = pd.crosstab(index = cars93["Origin"],
            columns = cars93["Type"])

print(table)

from scipy.stats import chi2_contingency
chi2_contingency(table)
```

▶ pd.crosstab함수 : 분할표 작성

▶ chi2_contingency(observed) : Cars93의 "Origin"변수와 "Type"변수의 분할표에 대하여 카이제곱 검정 수행
- observed = : 관측빈도가 포함된 분할표 지정

```
In [6]: import pandas as pd
   ...: cars93 = pd.read_csv("d:/data/cars93.csv")

In [7]: table = pd.crosstab(index = cars93["Origin"],
   ...:                     columns = cars93["Type"])

In [8]: print(table)
Type      Compact  Large  Midsize  Small  Sporty  Van
Origin
USA             7     11       10      7       8    5
non-USA         9      0       12     14       6    4

In [9]: from scipy.stats import chi2_contingency
   ...: chi2_contingency(table)
Out[9]:
Chi2ContingencyResult(statistic=14.079853971260219, pvalue=0.015110051037674505, dof=5,
expected_freq=array([[ 8.25806452,  5.67741935, 11.35483871, 10.83870968,  7.22580645,
         4.64516129],
       [ 7.74193548,  5.32258065, 10.64516129, 10.16129032,  6.77419355,
         4.35483871]]))
```

▶ 가설 H_0 : 자동차 제조국과 자동차 타입은 서로 독립이다. H_1 : 자동차 제조국과 자동차 타입은 서로 독립이 아니다.

검정통계량 $\chi^2 = \sum_{i=1}^{6}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} (\approx \chi^2(5)\ \ under\ \ H_0)$

카이제곱 검정통계량의 관측값은 14.0799, 유의확률($p-value$)은 0.01511로 유의수준 $\alpha=0.05$에서 귀무가설을 기각할 수 있다. 독립성 검정 결과, 자동차 제조국과 자동차 타입은 서로 독립이 아니라고 할 수 있다.

검정통계량의 자유도와 각 cell별 기대빈도가 출력되었다.

▸ **예. 학업성적과 학생의 성별이 어떤 관계가 있는지 알아보기 위하여 대학생 200명을 무작위 추출하여 조사한 결과 다음과 같은 데이터를 얻었다. 학업성적과 학생의 성별이 어떤 관계가 있는지 유의수준 5%에서 검정하시오.**

성별 \ 학업성적	A	B	C	D	합계
여학생	19	44	13	3	79
남학생	19	58	31	13	121
합계	38	102	44	16	200

```
import numpy as np
import pandas as pd
x = np.array([[19,44,13,3],[19,58,3,13]])
table2 = pd.DataFrame(x,['Male', 'Female'],['A','B','C','D'])
print(table2)

from scipy.stats import chi2_contingency
chi2_contingency(table2)
```

상관분석

상관분석(Correlation Analysis)

▸ **두 변수 사이의 선형적 연관성에 관한 추론**

▸ **X와 Y의 모상관계수 :** $\rho = Corr(X, Y) = \dfrac{Cov(X, Y)}{\sqrt{Var(X)\, Var(Y)}} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

▸ **X와 Y의 표본상관계수**

- 두 변수 x, y의 선형관계를 나타내는 측도 = Pearson의 표본상관계수

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- 성질

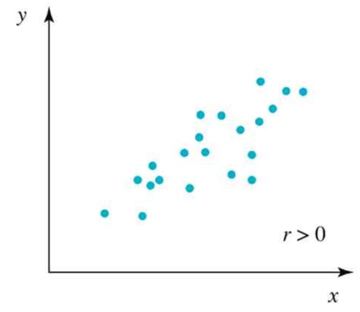
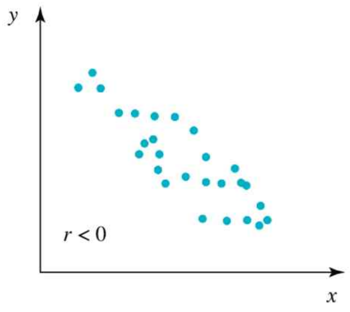
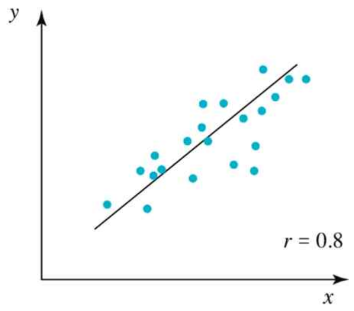
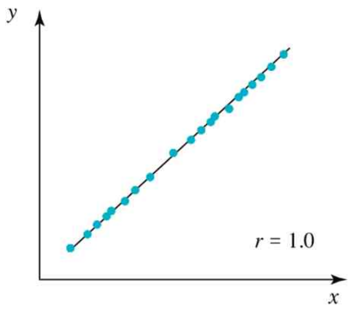
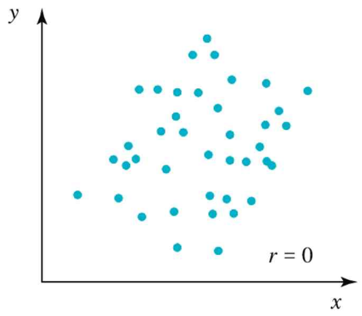
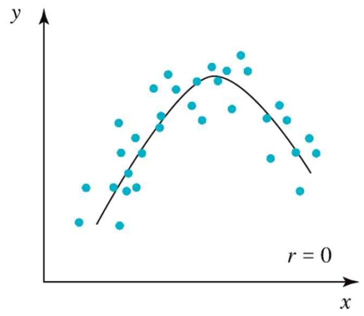
① $-1 \le r \le 1$

② $r = 0$: 두 변수 사이의 관계가 없음이 아니라 선형관계가 성립하지 않음을 뜻하는 것임

- 상관계수>0 : 양의 상관관계, 상관계수<0 : 음의 상관관계

③ r의 값은 관측값이 직선 관계에 가까울수록, -1(기울기가 음의 방향) 또는 1(기울기가 양의 방향)에 가까움

▸ **산점도와 상관계수**

• 상관계수와 부호 - 상관계수의 부호는 증감의 방향성을 나타냄	 
• 상관계수의 크기 - 상관계수의 절대값의 크기는 직선의 주변에 자료가 어느 정도 집중되어 있는지 나타냄	 
• 상관계수가 0인 경우의 예 - 상관계수는 두 변수의 ‘선형집중성’만을 재는 측도로서 비선형 연관관계를 반영하지 못함.	 

▸ **상관관계의 유무에 관한 검정**

- 모집단에서 두 변수 간의 선형적인 연관성을 분석
- 가설
  - 귀무가설 H_0 : 두 변수간의 선형적 연관성이 존재하지 않는다
  - 대립가설 H_1 : 두 변수간의 선형적 연관성이 존재한다(두 변수간의 상관관계가 존재한다)

▸ **예. (cars93.csv) cars93의 자동차 중량(Weight)과 자동차 연비(MPG.city) 사이에는 선형적 연관성(상관관계)가 있는지 유의수준 $\alpha = 0.05$에서 알아보자.**

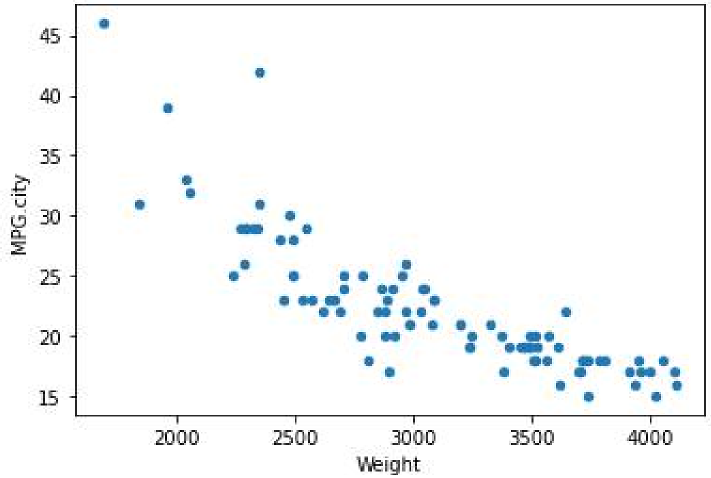
- 가설 - 귀무가설 H_0 : 자동차 중량(Weight)과 자동차 연비(MPG.city) 사이에 선형적 연관성이 존재하지 않는다

  - 대립가설 H_1 : 자동차 중량(Weight)과 자동차 연비(MPG.city) 사이에 선형적 연관성이 존재한다

- cars93의 Weight와 MPG.city 사이의 산점도를 그리고 상관계수를 구하여 상관관계 여부를 확인

```
import pandas as pd
import numpy as np
cars93 = pd.read_csv("d:/data/cars93.csv")

cars93.plot(kind='scatter', x='Weight',y='MPG.city')
np.corrcoef(cars93["Weight"],cars93["MPG.city"])
```




```
In [14]: np.corrcoef(cars93["Weight"],cars93["MPG.city"])
Out[14]:
array([[ 1.        , -0.84313855],
       [-0.84313855,  1.        ]])
```

- cars93의 Weight와 MPG.city 상관계수와 산점도로 두 변수간에 양의 상관관계가 있다고 판단되어 상관분석 실시
- pearsonr(scipy.stats 모듈) : 두 양적변수에 대하여 상관분석 실시

`from scipy.stats import pearsonr` `pearsonr(cars93["Weight"],cars93["MPG.city"])`	▶pearsonr(x= , y=) : 두 변수간의 상관관계의 유무에 관한 검정 실시 - x= , y= : 두 변수를 지정 - <실행결과> 주어진 두 변수 x, y에 대한 피어슨의 상관계수와 유의확률을 출력.

```
In [15]: from scipy.stats import pearsonr
    ...: pearsonr(cars93["Weight"],cars93["MPG.city"])
Out[15]: PearsonRResult(statistic=-0.8431385479968199, pvalue=2.967048334373354e-26)
```

▶ 가설 - 귀무가설 H_0 : 자동차 중량(Weight)과 자동차 연비(MPG.city) 사이에 선형적 연관성이 존재하지 않는다

대립가설 H_1 : 자동차 중량(Weight)과 자동차 연비(MPG.city) 사이에 선형적 연관성이 존재한다

두 변수간의 상관관계의 유무에 관한 검정결과, 상관계수 $r=-0.8431$, 유의확률($p-value$)은 2.967×10^{-26}로 유의수준 $\alpha=0.05$에서 귀무가설을 기각한다. 따라서 유의수준 $\alpha=0.05$에서 자동차 중량(Weight)과 자동차 연비(MPG.city) 사이에 선형적 연관성이 존재한다고 할 수 있다.

▸ **예. 시본(seaborn)패키지에 있는 tips 데이터프레임에서 레스토랑의 총매출액(total_bill) 과 팁(tip) 사이의 산점도를 그리고 상관계수를 먼저 구해보자. 두 변수간의 상관관계가 있는지 유의수준 $\alpha = 0.05$에서 확인하시오.**

변수	설명
total_bill	총매출액(달러)
tip	팁(달러)
sex	성별(Female, Male)
smoker	손님의 흡연여부 (No, Yes)_
day	요일(Thur, Fri, Sat, Sun)
time	식사시간(Lunch, Dinner)
size	식사인원

```
import seaborn as sns
tips = sns.load_dataset("tips")

import pandas as pd
import numpy as np

tips.plot(kind='scatter', x='tip',y='total_bill')
np.corrcoef(tips["tip"],tips["total_bill"])

from scipy.stats import pearsonr
pearsonr(tips["tip"],tips["total_bill"])
```

회귀분석

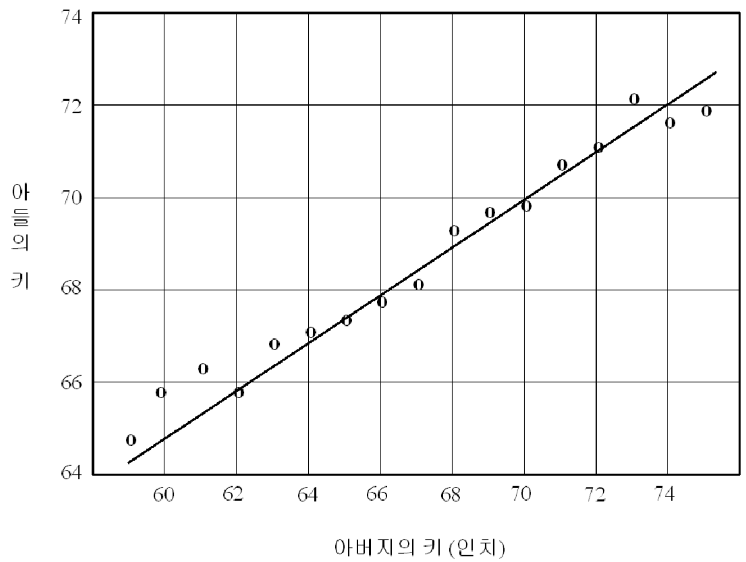
회귀분석

■ 회귀분석(regression analysis)

▸ 반응변수와 설명변수 사이의 함수관계를 규명

- $y = f(x)$의 함수관계가 있을 때,
  - x : 설명변수(explanatory variable) 또는 독립변수(independent variable)
  - y : 반응변수(response variable) 또는 종속변수(dependent variable)
  - 설명변수와 반응변수 모두 양적자료

- 예. 아버지의 키와 아들의 키를 함수관계로 규명
  - 아버지의 키 그룹별로 대응되는 아들 그룹의 평균 키를 표시, 이러한 관계를 잘 나타낼 수 있는 직선 표현



▸ **단순회귀분석(simple regression analysis) : 설명변수가 1개인 회귀모형**

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, ..., n$$

$\alpha, \quad \beta$: 회귀모수(모회귀계수) (α: 상수항, β: 기울기)

$x_1, \cdots, x_n$: 설명변수(독립변수), $Y_1, \cdots, Y_n$: 반응변수(종속변수)

$e_1, \cdots, e_n$: 오차항으로 서로 독립인 $N(0, \sigma^2)$ 확률변수-선형성, 독립성, 등분산성 가정

▸ **중회귀분석(multiple regression analysis) : 설명변수가 2개 이상인 회귀모형**

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + e_i, \quad i = 1, \cdots, n$$

$\beta_0, \beta_1, \cdots . \beta_k$: 회귀모수(모회귀계수)

$x_{1i}, \cdots, x_{ki}$: 설명변수(독립변수)

$Y_1, \cdots, Y_n$: 반응변수(종속변수)

$e_1, \cdots, e_n$: 서로 독립인 $N(0, \sigma^2)$ 확률변수로 선형성, 등분산성, 독립성 가정

■ 단순회귀분석

- $y = \alpha + \beta x$ **: 모회귀직선 (population regression line)**

- $\hat{y} = \hat{\alpha} + \hat{\beta} x$ **: 적합된 회귀직선(fitted regression line)**

  - 최소제곱법(least squares method)에 의하여 α, β를 추정
  - α와 β의 최소제곱 추정량 : $\hat{\beta} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$, $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$

- **결정계수(coefficient of determination)**
  - 단순회귀모형에서 회귀직선에 의해 반응변수가 설명되어지는 정도
  - $0 \le R^2 \le 1$: 결정계수가 1에 가까운 값을 가질수록 반응변수가 회귀직선에 의해 설명이 잘됨을 의미함

▸ 회귀직선의 유의성 검정

- 설명변수 x를 고려한 회귀직선이 의미가 있는가를 검정 : 회귀직선이 유의미한가?
- 회귀직선의 유의성을 나타내는 가설

  - 귀무가설 H_0 : 회귀직선은 유의하지 않다.

    대입가설 H_1 : 회귀직선은 유의하다.

▸ 잔차분석(residual analysis)

- 회귀모형($Y_i = \alpha + \beta x_i + e_i$)과 오차항($e_i$)에 대하여 다음과 같은 가정이 내포됨

  ① 선형성 : $E(e_i) = 0$, 즉, $E(Y_i | x_i) = \alpha + \beta x_i$

  ② 등분산성 : $Var(e_1) = \ldots = Var(e_n) = \sigma^2 (> 0)$

  ③ 독립성 : $e_1, e_2, \ldots, e_n$은 서로 독립, $Cov(e_i, e_j) = 0, \quad i \neq j$

  ④ 정규성 : $e_i \sim N(0, \sigma^2)$

▸ **잔차도(residual plot) : 설명변수와 스튜던트화 잔차 $(x_1,\widehat{e_{st,1}}),(x_2,\widehat{e_{st,2}}),...,(x_n,\widehat{e_{st,n}})$를 산점도로 나타낸 것**

- 잔차도가 다음과 같은 성질을 만족하면 단순선형회귀모형 적용이 타당하다고 할 수 있음

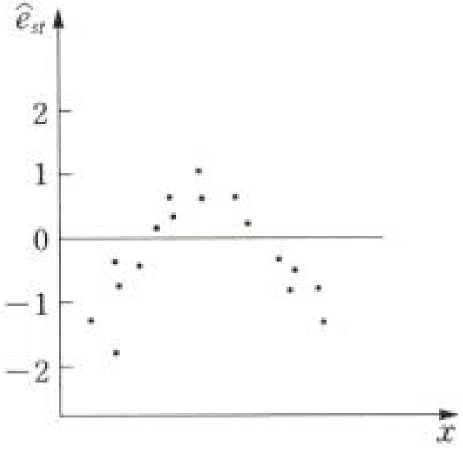
  ① 선형성 : 대략 0에 관하여 대칭.

  ② 등분산성 : 설명변수 값에 따른 잔차의 산포가 크게 다르지 않음.

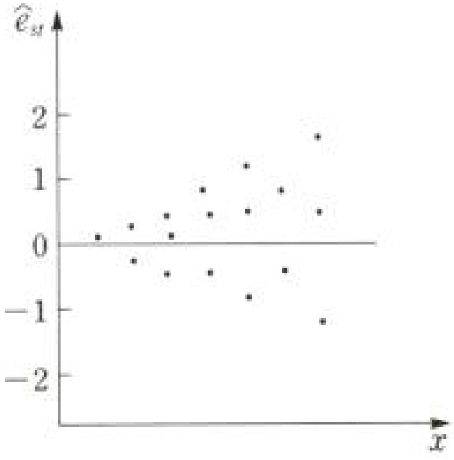
  ③ 독립성 : 점들이 산재된 모양에 특별한 경향이 없음.

  ④ 정규성 : 대부분의 점들이 (-2, +2)의 범위 내에 있어야 함

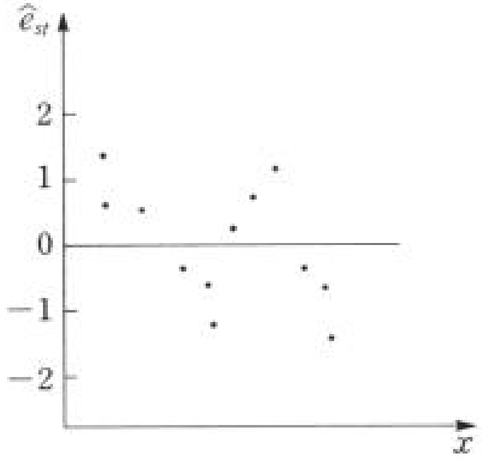
- 예. 단순선형회귀모형의 가정에 어긋나는 경우



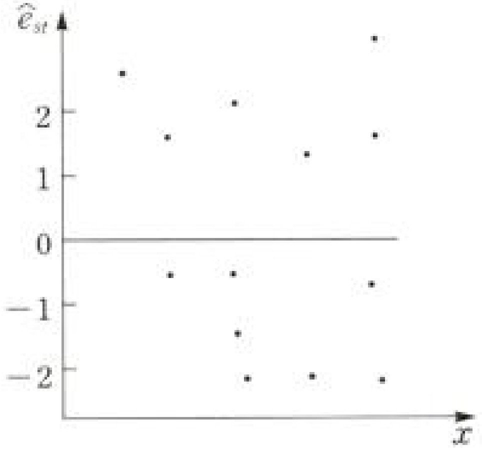
(a) 선형성에 벗어나는 경우



(b) 등분산성에 벗어나는 경우



(c) 독립성에 벗어나는 경우



(d) 정규성에 벗어나는 경우

■ 회귀분석의 절차

(1) 산점도 작성

(2) 회귀모형 적합

(3) 잔차분석

▸ 예. 사용년수에 따른 중고차 가격자료는 다음과 같다.

사용년수에 따른 중고차 가격의 변화 (단위 : 백만원)

사용년수	1.0	1.5	2.0	2.0	3.0	3.0	3.2	4.0	4.5	5.0	5.0	5.5
중고차가격	4.5	4.0	3.2	3.4	2.5	2.3	2.3	1.6	1.5	1.0	0.8	0.4

(1) 단순회귀직선을 추정하시오.

(2) 결정계수를 구하시오.

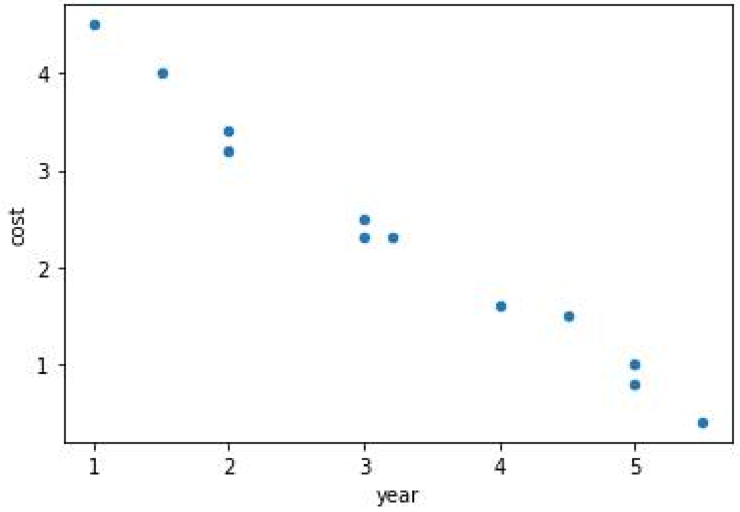
(3) 회귀직선의 유의성을 검정하시오.(유의수준 $\alpha = 0.05$)

- ols.fit함수(statsmodels.formula.api 모듈) : 독립변수(x)와 종속변수(y)을 이용하여 단순선형회귀분석을 수행 적합 결과를 확인하기 위해서는 summary()함수를 사용한다.

import pandas as pd
import numpy as np
 	
rawdata= {'year' :[1.0,1.5,2.0,2.0,3.0,3.0,
                    3.2,4.0,4.5,5.0,5.0,5.5],
          'cost':[4.5,4.0,3.2,3.4,2.5,2.3,
                 2.3,1.6,1.5,1.0,0.8,0.4]
          }
 	
from pandas import DataFrame
usedcar = DataFrame(rawdata)
 	
usedcar.plot(kind='scatter', x='year',y='cost')
	▶데이터프레임 작성
	
	▶ 산점도 그리기

from statsmodels.formula.api import ols model = ols("cost ~ year", usedcar).fit() print(model.summary())	▶단순선형회귀모형 적합
fitted = model.fittedvalues residual = model.resid rstandard = model.resid_pearson import matplotlib.pyplot as plt plt.scatter(x=fitted, y=rstandard) plt.axhline(y=0) plt.title("studentized residuals vs fitted values") plt.ylabel("studentized residuals") plt.xlabel("fitted values") plt.show()	▶ 잔차분석

```
In [25]: import pandas as pd
    ...: import numpy as np
    ...:
    ...: rawdata= {'year' :[1.0,1.5,2.0,2.0,3.0,3.0,
    ...:                    3.2,4.0,4.5,5.0,5.0,5.5],
    ...:           'cost':[4.5,4.0,3.2,3.4,2.5,2.3,
    ...:                   2.3,1.6,1.5,1.0,0.8,0.4]
    ...:           }
    ...:
    ...: from pandas import DataFrame
    ...: usedcar = DataFrame(rawdata)
    ...:
    ...: usedcar.plot(kind='scatter', x='year',y='cost')
```



▶ 중고차 사용년수와 중고차 가격의 산점도 작성 : 두 변수간의 선형관계가 보임

```
In [26]: from statsmodels.formula.api import ols
    ...: model = ols("cost ~ year", usedcar).fit()
    ...: print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   cost   R-squared:                       0.984
Model:                            OLS   Adj. R-squared:                  0.983
Method:                 Least Squares   F-statistic:                     629.8
Date:                Tue, 06 Aug 2024   Prob (F-statistic):           2.31e-10
Time:                        05:53:52   Log-Likelihood:                 5.3366
No. Observations:                  12   AIC:                            -6.673
Df Residuals:                      10   BIC:                            -5.703
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.1329      0.123     41.600      0.000       4.858       5.408
year          -0.8588      0.034    -25.095      0.000      -0.935      -0.783
==============================================================================
Omnibus:                        0.703   Durbin-Watson:                   1.413
Prob(Omnibus):                  0.704   Jarque-Bera (JB):                0.588
Skew:                           0.073   Prob(JB):                        0.745
Kurtosis:                       1.926   Cond. No.                         9.66
==============================================================================
```

▶① 단순선형회귀모형 적합

적합된 회귀모형은 $\widehat{cost} = 45.1329 - 0.8588 year$

- coef : 추정된 회귀계수
- 사용년수가 1년 증가하면 중고차가격은 858,000 원씩 감소함을 의미함

② 결정계수(R-squared) : 0.984

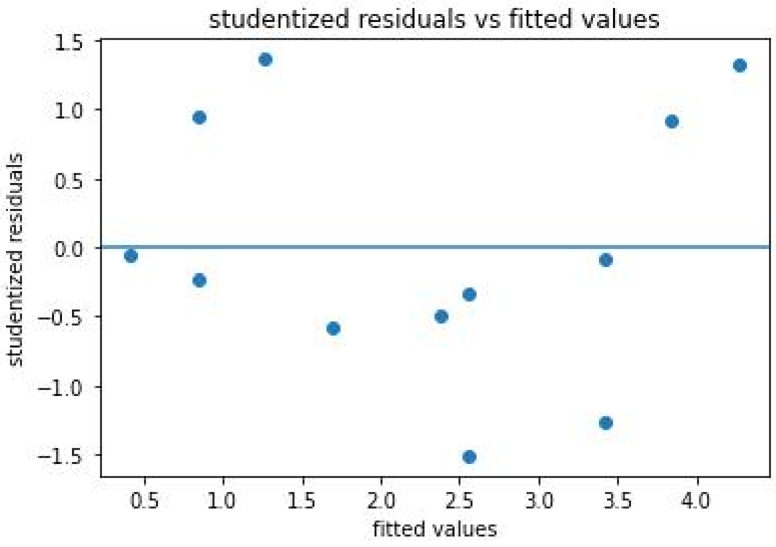
즉 회귀직선이 전체 반응변수값의 산포 중 98.4%를 설명한다는 것을 의미함

③ 회귀직선의 유의성 검정

- 귀무가설(H_0) : 회귀직선은 유의하지 않다.
  대립가설(H_1) : 회귀직선은 유의하다.
- 검정통계량의 관측값(F-statistic) : 629.8
- 유의확률(Prob) : 2.31×10^{-10}
  ⇒ 유의수준 $\alpha = 0.05$보다 작으므로 귀무가설을 기각한다. 따라서 회귀직선은 유의하다.

```python
fitted = model.fittedvalues
residual = model.resid
rstandard = model.resid_pearson

import matplotlib.pyplot as plt
plt.scatter(x=fitted, y=rstandard)
plt.axhline(y=0)
plt.title("studentized residuals vs fitted values")
plt.ylabel("studentized residuals")
plt.xlabel("fitted values")
plt.show()
```



▶잔차분석

① 선형성 : 대략 0에 관하여 대칭.

② 등분산성 : 설명변수 값에 따른 잔차의 산포가 크게 다르지 않음.

③ 독립성 : 점들이 산재된 모양에 특별한 경향이 없음.

④ 정규성 : 대부분의 점들이 (-2, +2)의 범위 내에 있어야 함

⇒ 잔차분석 결과 4가지 가정이 만족함을 알 수 있다.

▸ **예. 시본(seaborn)패키지에 있는 tips 데이터프레임에서 레스토랑의 총매출액(total_bill) 과 팁(tip) 사이의 회귀모형을 적합해보자. 또한 결정계수를 구하고 회귀모형이 유의한지 유의수준 $\alpha = 0.05$에서 확인하시오.**

변수	설명
total_bill	총매출액(달러)
tip	팁(달러)
sex	성별(Female, Male)
smoker	손님의 흡연여부 (No, Yes)_
day	요일(Thur, Fri, Sat, Sun)
time	식사시간(Lunch, Dinner)
size	식사인원

```
from statsmodels.formula.api import ols
model2 = ols("total_bill ~ tip+size", tips).fit()
print(model2.summary())

fitted = model2.fittedvalues
residual = model2.resid
rstandard = model2.resid_pearson

import matplotlib.pyplot as plt
plt.scatter(x=fitted, y=rstandard)
plt.axhline(y=0)
plt.title("studentized residuals vs fitted values")
plt.ylabel("studentized residuals")
plt.xlabel("fitted values")
plt.show()
```

감사합니다