한빛아카데미 Hanbit Academy, Inc.

난생처음

# 데이터 분석 with 파이썬

## 회귀분석

# 목차

# 01

# 선형 회귀분석의 개념

# 01. 선형 회귀분석의 개념

## I. 선형 회귀분석의 모형

– 선형 회귀분석(Linear regression analysis) :
두 개 또는 그 이상의 변수 간 인과관계의 패턴을 원래 모습과 가장 가깝게 추정하는 분석 방법.

– 함수 y = 2x + 80의 그래프, 이를 선형 회귀분석의 모형(Model)이라고 부름.

– 선형 회귀분석은 x 변수가 원인, y 변수가 결과로 인과관계여야 한다는 조건 있음.

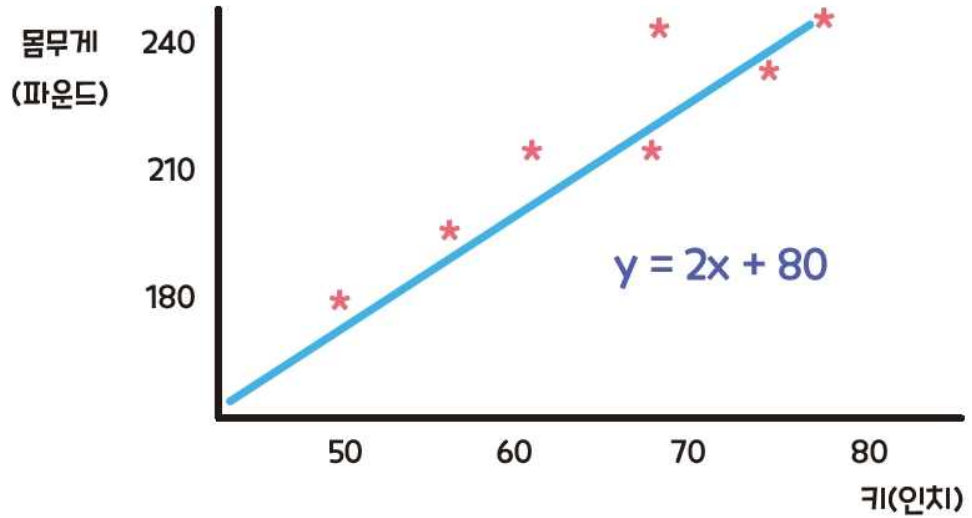



그림 10-1 선형 회귀분석 그래프 ©BU School of Public Health

# 01. 선형 회귀분석의 개념

## I. 선형 회귀분석의 모형

- 회귀분석에서 원인인 x 변수는 독립변수(Independent variable), 결과인 y 변수는 종속변수(Dependent variable)라고 부름.
- 어떠한 결과의 원인이 되는 독립변수가 한 개일 때 단순 선형 회귀분석(Simple linear regression analysis), 두 개 이상이면 다중 선형 회귀분석(Multiple linear regression analysis).
- 공부 시간과 시험 점수의 인과관계를 분석해보면 종속변수는 시험 점수이며 독립변수는 공부 시간만 고려.

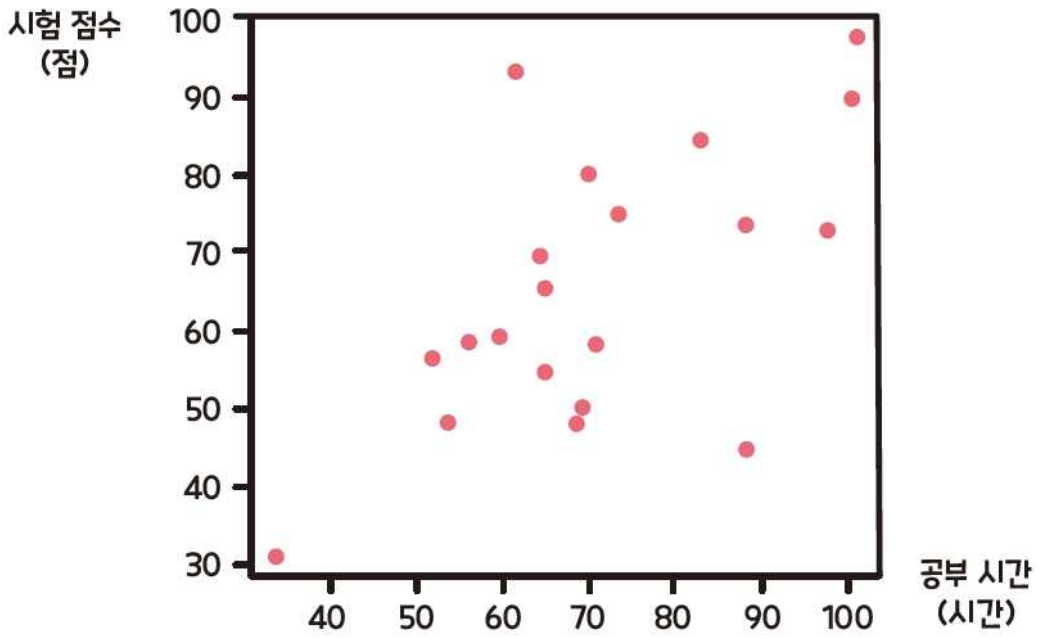



그림 10-2 선형 회귀분석 그래프

# 01. 선형 회귀분석의 개념

## I. 선형 회귀분석의 모형

- [그림 10-2]의 점들을 직선 하나로 표현하고자 함.
- 직선을 일차함수 y = mx + b로 표현한다면 계수 m은 이 직선의 기울기이므로 양수일 것.
- 일반적인 선형 회귀모형은 다음 식과 같이 독립변수 $x_i$앞에 계수 $\beta_i$가 붙음.
  마지막 f은 오차 항.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots\cdots + \varepsilon$$

$\beta_i$ : $x_i$의 계수 　　오차

- 데이터를 표현하는 직선은
  여러 개 존재할 수 있음.

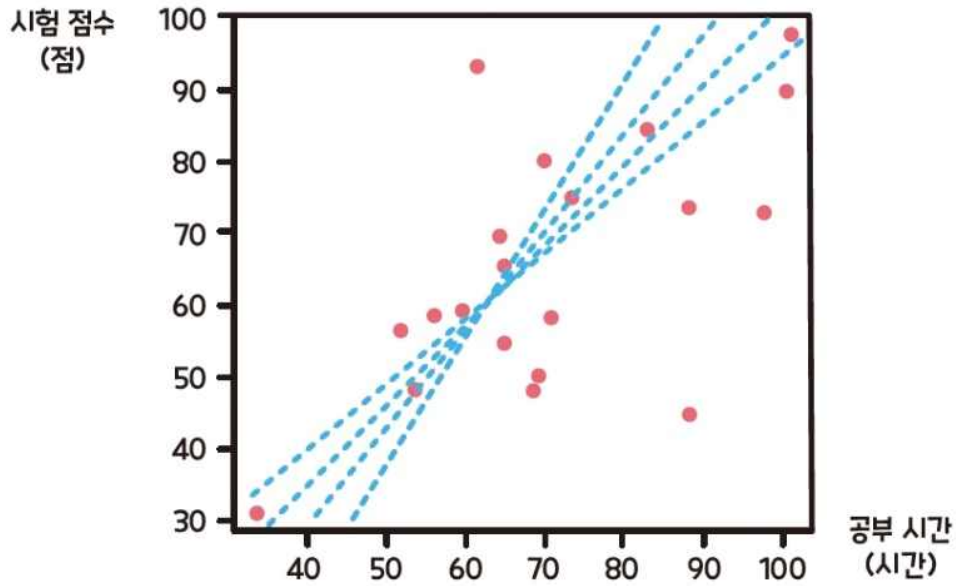



그림 10-3 선형 회귀분석 그래프

# 01. 선형 회귀분석의 개념

## I. 선형 회귀분석의 모형

- 최대한 많은 점과 거리가 가까운 직선이 좋은 직선.
- [그림 10-4]와 같이 점에서 직선까지 y축과 평행한 선분을 그렸을 때 모든 선분 길이의 합을 최소로 하는 직선을 찾는 것.

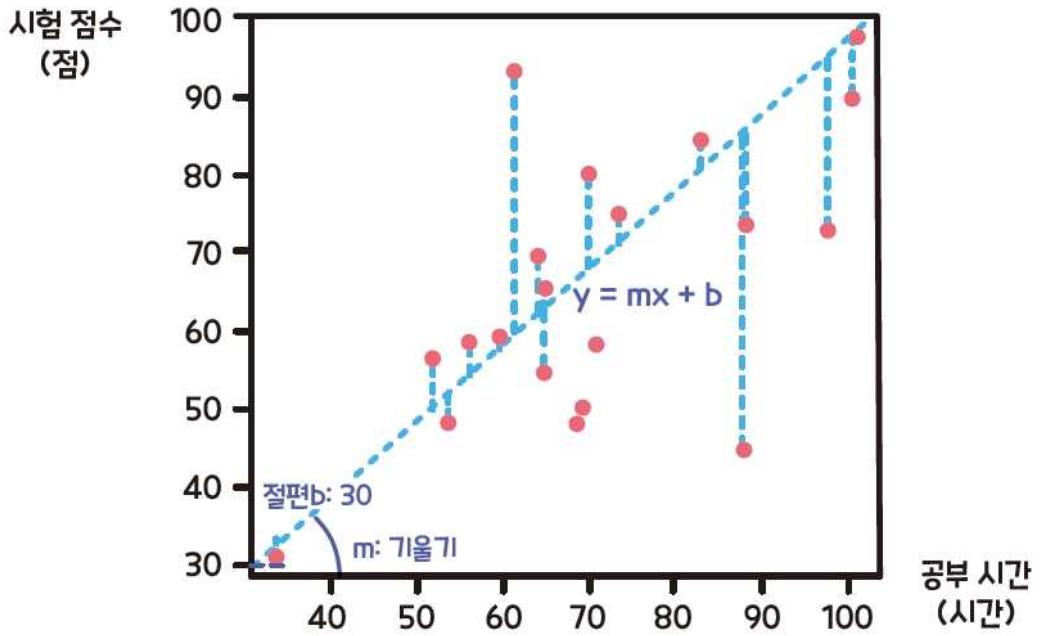



그림 10-4 선형 회귀분석 그래프

- [그림 10-4]에서 점이 가리키는 값과 직선이 예측하는 값의 차를 잔차(Residual)라고 부름.

# 01. 선형 회귀분석의 개념

## I. 선형 회귀분석의 모형

✓**하나 더 알기**: 더미 변수

더미 변수(Dummy variable)는 독립변수를 0과 1로 변환하여 '예'와 '아니오'로 나타낼 수 있는 변수.

더미 변수를 여러 개 두면 '예'와 '아니오'만으로 결과를 세 가지 이상으로 구분할 수 있음.
[표 10-1]과 같이 어린이, 청소년, 성인으로 구분.

① 데이터가 어린이일 때 '어린이' 더미 변수는 1이고 '청소년' 더미 변수는 0.

② 청소년일 때 '어린이' 더미 변수는 0이고 '청소년' 더미 변수는 1.

③ 어린이도 청소년도 아닌 성인은 두 더미 변수에 모두 0을 대입하여 표현할 수 있음.

따라서 구분하고자 하는 데이터의 종류가 N개일 때 더미 변수 N-1개를 선언하면 됨.

표 10-1 더미 변수

| | 어린이 더미 변수 | 청소년 더미 변수 |
|---|---|---|
| 어린이 | 1 | 0 |
| 청소년 | 0 | 1 |
| 성인 | 0 | 0 |

# 01. 선형 회귀분석의 개념

## I. 선형 회귀분석의 모형

- 결정계수
  - 선형 회귀분석에서 모형이 데이터의 패턴을 얼마나 효과적으로 보여주는지 수치화한 값을 결정계수라고 함.
  - 결정계수 R2(R square)를 다음 수식과 같이 정의.

$$R^2 = \frac{(Q - Q_e)}{Q}$$

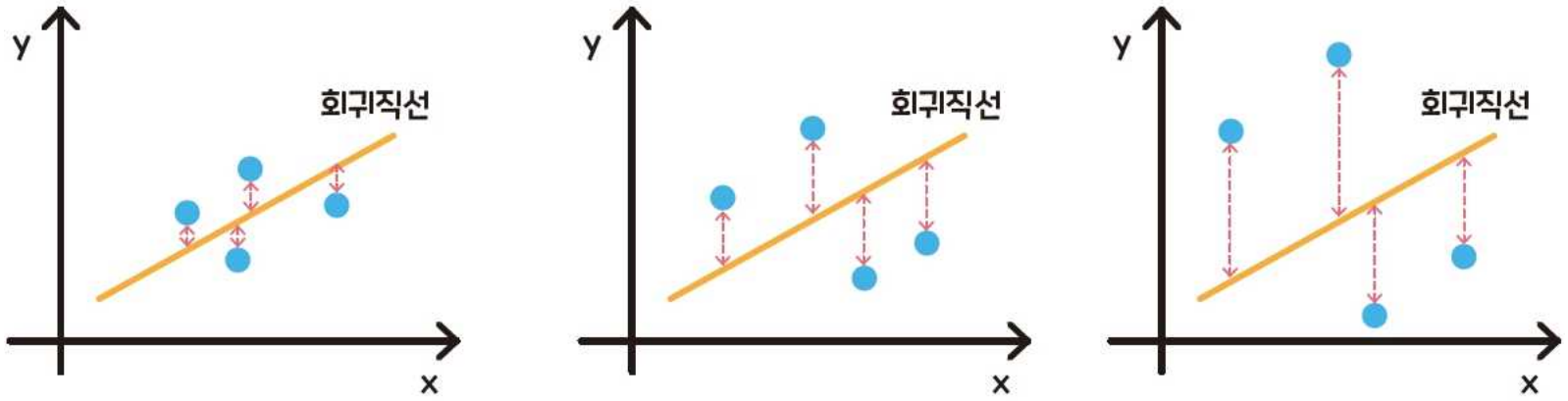
$Q$ = 전체 데이터의 편차 제곱의 합

$Q_e$ = 전체 데이터의 잔차 제곱의 합

# 01. 선형 회귀분석의 개념

## I. 선형 회귀분석의 모형

- 결정계수
  - 편차(Deviation)는 평균과 실제 값의 차이.
  - 결정계수 $R^2$ 는 0 이상 1 이하의 값으로 계산됨.
  - $R^2$값이 1에 가까울 때 잔차가 작고 예측의 정밀도가 높음.
    반면 $R^2$ 값이 0에 가까울 때 잔차가 커 예측의 정밀도가 낮음.




(a) $R^2$가 1에 가까운 경우 (b) $R^2$가 0.5에 가까운 경우 (c) $R^2$가 0에 가까운 경우

그림 10-5 선형 회귀분석 그래프

# 01. 선형 회귀분석의 개념

## I. 선형 회귀분석의 모형

- 수정된 결정계수
  - 결정계수 $R^2$ 는 독립변수의 개수가 많을수록 커지는 경향을 보임.
  - 이러한 문제를 해결하기 위해 다중 선형 회귀분석에서 수정된 결정계수 adj. $R^2$ (adjusted R square)로 설명력을 나타냄.

$$adj.R^2 = \frac{(n-1)}{(n-p-1)(1-R^2)}$$

$R^2 =$ 결정계수 $p =$ 독립변수의 개수 $n =$ 표본 수

# 01. 선형 회귀분석의 개념

## II. 선형 회귀분석의 해석

- 통계적 가설검정
  - 모집단에 대한 추측을 하고 표본의 정보를 기준으로 가설이 타당한지 판정하는 방법.
  - 통계적 가설에는 두 종류가 있음.
    통계학에서 처음부터 거짓일 것으로 기대하는 가설인 귀무가설(Null hypothesis),
    입증하고자 하는 가설인 대립가설(Alternative hypothesis).
  - 실험 결과를 보고 귀무가설을 채택하거나 대립가설을 채택하는 기준을 세워 두어야 하고, 그 기준을 유의수준(Significance level)이라고 함.
  - 귀무가설이 참일 때 실제 결과가 실험 결과와 같을 확률은 p-값(p-value, 유의확률)이라고 부름.

# 01. 선형 회귀분석의 개념

## II. 선형 회귀분석의 해석

- 선형 회귀분석 과정
  - 선형 회귀분석을 수행하고 해석하는 과정은 크게 세 단계
    ① 결과인 종속변수를 y로 두고, 원인이 되는 독립변수를 xi로 둠.
    ② 설명력 $R^2$또는 adj.$R^2$값을 확인. 결정계수가 0.6 또는 0.4 이상이면 해당 회귀모형이 설명력을 갖추었다고 인정.
    ③ 각 독립변수의 p-값이 유의수준보다 작은지 확인. p-값이 유의수준 이상인 변수를 제외하고 남은 독립변수가 결과에 영향을 주는 원인.

# 02

# 선형 회귀분석의 활용

# 02. 선형 회귀분석의 활용

## I. 연봉과 직장 만족도

– 연봉이 직장 만족도와 얼마나 관계 있는지 파악할 수 있음.

표 10-2 직장 만족도(1)

| 직장 만족도(점) | 60 | 75 | 70 | 85 | 90 | 70 | 65 | 95 | 70 | 80 |
|---|---|---|---|---|---|---|---|---|---|---|
| 연봉(만 원) | 3,000 | 4,200 | 4,000 | 5,000 | 6,000 | 3,800 | 3,500 | 6,200 | 3,900 | 4,500 |

– 원인인 독립변수는 연봉이며 결과인 종속변수는 직장 만족도.

**[코드 10-1] 산점도 확인**

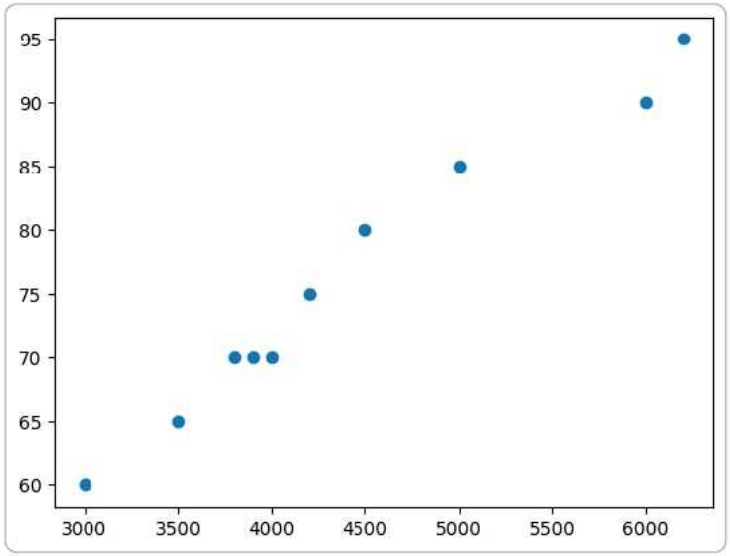
```
import pandas as pd
import matplotlib.pyplot as plt

x = [3000, 4200, 4000, 5000, 6000, 3800, 3500, 6200, 3900, 4500]
y = [60, 75, 70, 85, 90, 70, 65, 95, 70, 80]
data = {'x': x, 'y': y}
df = pd.DataFrame(data)
plt.scatter(df['x'], df['y'])
plt.show( )
```

# 02. 선형 회귀분석의 활용

## I. 연봉과 직장 만족도

**[코드 10-1] 실행결과**




- 리스트 x에 독립변수인 연봉을 할당, 리스트 y에 종속변수인 직장 만족도를 할당.
- 맷플롯립의 scatter( ) 함수로 산점도를 그리면 산점도는 우상향 형태로 나타남.

# 02. 선형 회귀분석의 활용

## I. 연봉과 직장 만족도

- 산점도를 보면 두 변수에 양의 상관관계가 있음을 알 수 있는데, 인과관계가 얼마나 강한지 선형 회귀분석으로 확인 하려함.
- 선형 회귀모형은 다음과 같은 형태로 설정.

```
종속변수 ~ 독립변수1 + 독립변수2 + 독립변수3 + ···
```

**[코드 10-2] 단순 선형 회귀분석**

```
from statsmodels.formula.api import ols
from sklearn.linear_model import LinearRegression

fit = ols('y ~ x', data=df).fit( )
print(fit.summary( ))
```

- ols( ) 함수는 선형 회귀분석을 수행하는 함수. 첫 번째 인자로 회귀모형 'y ~ x', 두 번째 인자로 데이터를 입력.

# 02. 선형 회귀분석의 활용

## I. 연봉과 직장 만족도

**[코드 10-2] 실행결과**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y R-squared:                       0.971
Model:                            OLS Adj. R-squared:                  0.968
Method:                 Least Squares F-statistic:                     271.0
Date:                Mon, 27 Mar 2023 Prob (F-statistic):           1.87e-07
Time:                        07:13:44 Log-Likelihood:                -20.111
No. Observations:                  10 AIC:                             44.22
Df Residuals:                       8 BIC:                             44.83
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     29.0004      2.926      9.913      0.000      22.254      35.747
x              0.0107      0.001     16.463      0.000       0.009       0.012
==============================================================================
Omnibus:                        0.346 Durbin-Watson:                   2.871
Prob(Omnibus):                  0.841 Jarque-Bera (JB):                0.447
Skew:                           0.286 Prob(JB):                        0.800
Kurtosis:                       2.136 Cond. No.                     2.07e+04
==============================================================================
```

# 02. 선형 회귀분석의 활용

## I. 연봉과 직장 만족도

- 첫째, 결정계수 $R^2$값이 0.971이므로 표본 데이터들에 대한 설명력이 97.1%이고
  결정계수가 0.6을 크게 초과하므로 모형의 정밀도가 높음.
- 둘째, 회귀모형의 독립변수 x의 유의수준은 0.000으로 0.05 미만이므로 x는 유의한 독립변수.
- 셋째, 독립변수 x의 계수는 0.0107.

$$y = 29.0004 + (0.0107) \times x + \varepsilon$$

# 02. 선형 회귀분석의 활용

## II. 직장 만족도의 요인 분석

– 연봉 외에 일평균 휴식시간(분)과 일평균 근무시간(시간) 추가하여 각 변수의 영향을 분석.

표 10-3 직장 만족도(2)

| 직장 만족도(점) | 60 | 75 | 70 | 85 | 90 | 70 | 65 | 95 | 70 | 80 |
|---|---|---|---|---|---|---|---|---|---|---|
| 연봉(만 원) | 3,000 | 4,200 | 4,000 | 5,000 | 6,000 | 3,800 | 3,500 | 6,200 | 3,900 | 4,500 |
| 일평균 휴식시간(분) | 120 | 60 | 100 | 100 | 50 | 120 | 90 | 40 | 120 | 120 |
| 일평균 근무시간(시간) | 8 | 6 | 10 | 8 | 10 | 10 | 9 | 7 | 8 | 9 |

– 결과인 직장 만족도가 종속변수이며, 나머지는 독립변수.

– 회귀모형 companySatisfaction ~ salary + breakTime + workingTime을 입력하여 다중 선형 회귀분석을 수행.

# 02. 선형 회귀분석의 활용

## II. 직장 만족도의 요인 분석

**[코드 10-3]** 다중 선형 회귀분석

```
from statsmodels.formula.api import ols
from sklearn.linear_model import LinearRegression

salary = [3000, 4200, 4000, 5000, 6000, 3800, 3500, 6200, 3900, 4500]
breakTime = [120, 60, 100, 100, 50, 120, 90, 40, 120, 120]
workingTime = [8, 6, 10, 8, 10, 10, 9, 7, 8, 9]
companySatisfaction = [60, 75, 70, 85, 90, 70, 65, 95, 70, 80]
data = {'salary': salary, 'breakTime': breakTime, 'workingTime': workingTime,
        'companySatisfaction': companySatisfaction}
df = pd.DataFrame(data)

fit = ols('companySatisfaction ~ salary + breakTime + workingTime',
data=df).fit( )
print(fit.summary( ))
```

# 02. 선형 회귀분석의 활용

## II. 직장 만족도의 요인 분석

### [코드 10-3] 실행결과

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     companySatisfaction  R-squared: 0.988
Model:                             OLS  Adj. R-squared: 0.982
Method:                  Least Squares    F-statistic: 164.0
Date:                 Mon, 27 Mar 2023    Prob (F-statistic): 3.81e-06
Time:                         07:30:46              Log-Likelihood: -15.777
No. Observations:                   10              AIC: 39.55
Df Residuals:                        6              BIC: 40.77
Df Model:                            3
Covariance Type:                nonrobust
==============================================================================
                 coef    std err          t      P>|t|           [0.025 0.975]
------------------------------------------------------------------------------
Intercept     24.9819      5.353       4.667      0.003           11.884 38.080
salary         0.0120      0.001      15.895      0.000            0.010 0.014
breakTime      0.0668      0.027       2.491      0.047            0.001 0.132
workingTime    -0.9718      0.412      -2.356      0.057           -1.981 0.037
==============================================================================
Omnibus:                        0.929 Durbin-Watson:       2.500
Prob(Omnibus):                  0.628 Jarque-Bera (JB):    0.752
Skew:                          -0.441 Prob(JB):           0.686
Kurtosis:                       1.986 Cond. No.           5.06e+04
==============================================================================
```

# 02. 선형 회귀분석의 활용

## II. 직장 만족도의 요인 분석

- 첫째, 수정된 결정계수 $adj.R^2$는 0.988이므로 이 선형 회귀모형의 설명력이 98.8%
- 둘째, 변수 salary는 p-값이 0.000이고 변수 breakTime은 0.047로 0.05 보다 작고 workingTime의 p-값은 0.057로 0.05 보다 큼.
  따라서 연봉과 휴식시간은 유의한 독립변수이고, 근무시간은 유의하지 않은 독립변수.
- 셋째, 유의수준 결과에 따라서 salary와 breakTime을 독립변수로 하는 모형이 구성됨.

$$companySatisfaction = 24.9819 + (0.0120) \times salary + (0.0668) \times breakTime + \varepsilon$$

# 03

# 로지스틱 회귀분석의 개념

# 03. 로지스틱 회귀분석의 개념

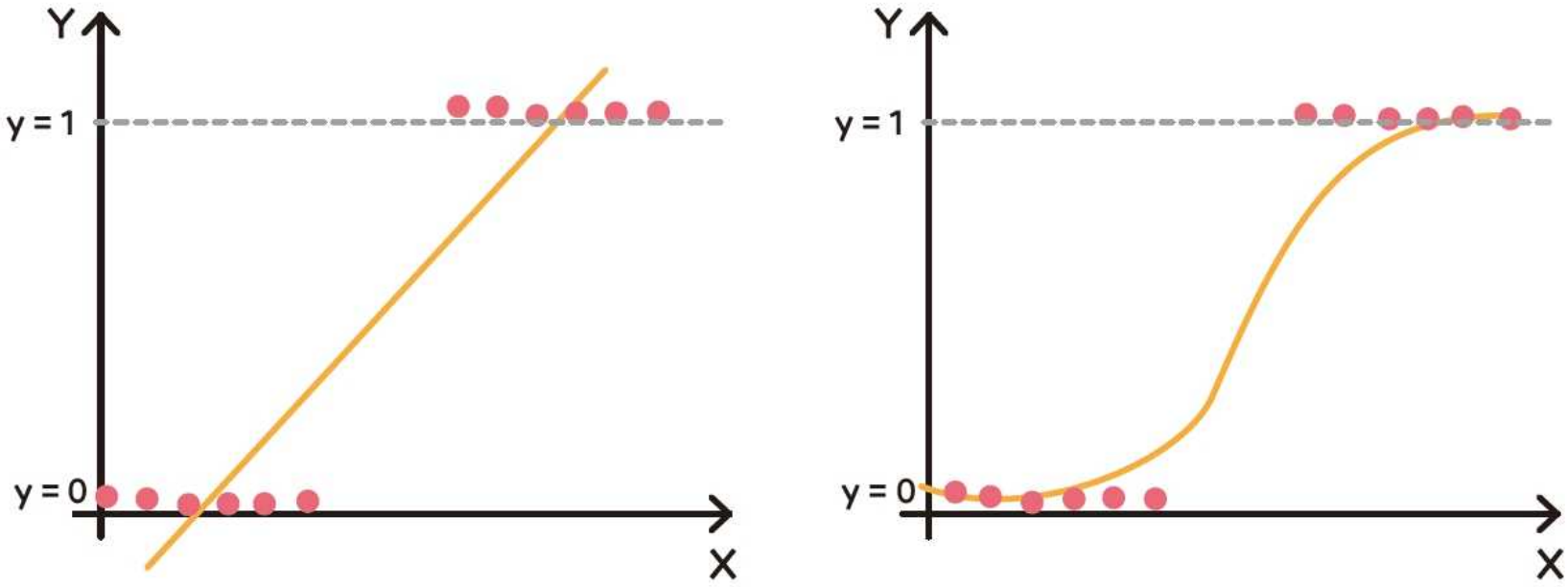
## I. 로지스틱 회귀모형

- 로지스틱 회귀분석(Logistic regression analysis)은 결과인 종속변수에 미치는 요인들을 독립변수로 두고 각 독립변수의 영향을 설명.
- 로지스틱 회귀분석의 종속변수는 범위에 제한이 있음. 로지스틱 회귀분석의 종속변수는 0에서 1사이의 값.

# 03. 로지스틱 회귀분석의 개념

## I. 로지스틱 회귀모형

- [그림 10-6]은 선형 회귀분석과 로지스틱 회귀분석의 모형.
- 왼쪽의 선형 회귀분석은 표본 데이터를 직선으로 그룹화.
- 반면 오른쪽의 로지스틱 회귀분석은 표본 데이터가 0에서 1사이의 값으로 그룹화되어 있음.




(a) 선형 회귀분석 (b) 로지스틱 회귀분석

그림 10-6 회귀모형

# 03. 로지스틱 회귀분석의 개념

## I. 로지스틱 회귀모형

- 로지스틱 회귀분석 결과 오즈비를 얻음. 오즈비(Odds Ratio, OR)는 우리말로 승산비.
- 사건이 발생할 확률을 p라고 할 때, 오즈비를 수식으로 나타내면 다음과 같음.

$$\text{OR} = \frac{\text{사건이 발생할 확률}}{\text{사건이 발생하지 않을 확률}} = \frac{p}{1-p}$$

- 대학 합격이라는 사건에서 합격을 1, 불합격을 0으로 정했을 때 합격할 확률이 0.8이라면 오즈비는 4. 이는 대학에 합격할 확률이 불합격할 확률보다 4배 높다는 뜻.

$$\text{OR} = \frac{p}{1-p} = \frac{0.8}{1-0.8} = 4$$

# 03. 로지스틱 회귀분석의 개념

## II. 로지스틱 회귀분석의 해석

– 로지스틱 회귀분석의 해석 과정은 다음과 같음.

» 첫째, 선형 회귀분석과 마찬가지로 각 독립변수의 p-값을 확인함. p-값이 유의수준보다 작은 독립변수를 통계적으로 유의한 변수라고 판단.

» 둘째, 오즈비를 구해 각 독립변수가 종속변수를 1로 만들 확률을 비교.

– 사과 가격이 사과 판매 여부에 미치는 영향을 분석해보기. 종속변수는 사과 판매 여부로, 사과가 판매되면 1이고 판매되지 않으면 0. 독립변수는 사과 가격.

표 10-4 사과 판매 여부 데이터

| 사과 판매 여부 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|
| 가격(원) | 1,500 | 2,000 | 5,000 | 3,000 | 3,500 | 2,500 | 4,000 | 4,500 | 3,000 |
| 사과 판매 여부 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 가격(원) | 4,500 | 4,000 | 4,500 | 5,500 | 6,500 | 5,000 | 3,500 | 7,000 | |

# 03. 로지스틱 회귀분석의 개념

## II. 로지스틱 회귀분석의 해석

**[코드 10-4] 로지스틱 회귀분석**

```
import statsmodels.api as sm
import pandas as pd
import numpy as np

sales = [1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]
price = [1500, 2000, 5000, 3000, 3500, 2500, 4000, 4500, 3000,\
         4500, 4000, 4500, 5500, 6500, 5000, 3500, 7000]
data = {'sales': sales, 'price': price}
df = pd.DataFrame(data)

logis = sm.Logit.from_formula('sales ~ price', data=df).fit( )
print(logis.summary( ))
print('OR')
print(np.exp(logis.params)))
```

# 03. 로지스틱 회귀분석의 개념

## II. 로지스틱 회귀분석의 해석

### [코드 10-4] 실행결과

```
Optimization terminated successfully.
        Current function value: 0.430873
        Iterations 7
                          Logit Regression Results
==============================================================================
Dep. Variable:                  sales    No. Observations:                   17
Model:                          Logit    Df Residuals:                       15
Method:                           MLE    Df Model:                            1
Date:                Mon, 03 Apr 2023    Pseudo R-squ.:                  0.3768
Time:                        00:48:37    Log-Likelihood:                -7.3248
converged:                       True    LL-Null:                       -11.754
Covariance Type:            nonrobust    LLR p-value:                  0.002917
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.5752      3.300      1.993      0.046       0.108      13.042
price         -0.0016      0.001     -2.008      0.045      -0.003    3.75e-05
==============================================================================
OR
Intercept 717.058841
price 0.998433
dtype: float64
```

# 03. 로지스틱 회귀분석의 개념

## II. 로지스틱 회귀분석의 해석

- 우선 각 독립변수가 유의한지 확인.
- 독립변수 price의 p-값은 0.04로 0.05 미만이므로 유의한 변수.
  y절편인 Intercept의 p-값 0.046도 확인할 수 있음.
- 변수 price의 오즈비가 0.998433이므로 가격을 올렸을 때 판매될 가능성이 판매되지 않을 가능성의 0.998433배.

# 04

# 로지스틱 회귀분석의 활용

# 04. 로지스틱 회귀분석의 활용

## I. 타이타닉 탑승자 생존여부 예측

– 타이타닉호 데이터에서 생존과 사망의 요인을 분석해보기.

**[코드 10-5] 타이타닉 탑승자 데이터**

```
import seaborn as sns

titanic = sns.load_dataset('titanic')
print(titanic)
```

```
     survived pclass     sex    age   sibsp parch     fare  embarked class \
0           0      3    male   22.0       1      0   7.2500         S Third
1           1      1  female   38.0       1      0  71.2833         C First
2           1      3  female   26.0       0      0   7.9250         S Third
3           1      1  female   35.0       1      0  53.1000         S First
4           0      3    male   35.0       0      0   8.0500         S Third
..        ...    ...     ...    ...     ...    ...      ...       ... ...
886         0      2    male   27.0       0      0  13.0000         S Second
887         1      1  female   19.0       0      0  30.0000         S First
888         0      3  female    NaN       1      2  23.4500         S Third
889         1      1    male   26.0       0      0  30.0000         C First
890         0      3    male   32.0       0      0   7.7500         Q Third
```

# 04. 로지스틱 회귀분석의 활용

## I. 타이타닉 탑승자 생존여부 예측

**[코드 10-4] 실행결과(계속)**

```
            who    adult_male   deck   embark_town  alive alone
0           man          True    NaN    Southampton     no False
1         woman         False      C    Cherbourg      yes False
2         woman         False    NaN    Southampton    yes True
3         woman         False      C    Southampton    yes False
4           man          True    NaN    Southampton     no True
..          ...          ...    ...            ...    ... ...
886         man          True    NaN   Southampton     no True
887       woman         False      B   Southampton    yes True
888       woman         False    NaN   Southampton     no False
889         man          True      C   Cherbourg      yes True
890         man          True    NaN   Queenstown      no True
[891 rows x 15 columns]
```

# 04. 로지스틱 회귀분석의 활용

## I. 타이타닉 탑승자 생존여부 예측

**[코드 10-6] 타이타닉 탑승자 데이터의 로지스틱 회귀분석**

```
import statsmodels.api as sm
import numpy as np
from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder( )
encoder.fit(titanic['sex'])
sex = encoder.transform(titanic['sex'])
titanic['sex'] = sex

model = sm.Logit.from_formula('survived ~ pclass + sex + age + fare + parch +
sibsp', data=titanic)
logit = model.fit( )
print(logit.summary( ))

print("OR")
print(np.exp(logit.params))
```

# 04. 로지스틱 회귀분석의 활용

## I. 타이타닉 탑승자 생존여부 예측

### [코드 10-6] 실행 결과

```
Optimization terminated successfully.
             Current function value: 0.445244
             Iterations 6

                           Logit Regression Results
==============================================================================
Dep. Variable:            survived     No. Observations:                  714
Model:                       Logit     Df Residuals:                      707
Method:                       MLE       Df Model:                            6
Date:          Mon, 03 Apr 2023       Pseudo R-squ.:                  0.3408
Time:                  00:59:54       Log-Likelihood:                -317.90
converged:                    True     LL-Null:                       -482.26
Covariance Type:       nonrobust      LLR p-value:                 5.727e-68
==============================================================================
              coef        std err         z        P>|z|         [0.025 0.975]
--------------------- ------------------------------------------------------
Intercept     5.3890         0.604       8.926       0.000          4.206 6.572
pclass       -1.2422         0.163      -7.612       0.000                  -1.562 -0.922
sex          -2.6348         0.220     -11.998       0.000                  -3.065 -2.204
age          -0.0440         0.008      -5.374       0.000                  -0.060 -0.028
fare          0.0022         0.002       0.866       0.386                  -0.003 0.007
parch        -0.0619         0.123      -0.504       0.614                  -0.303 0.179
sibsp        -0.3758       0.127       -2.950       0.003                  -0.625 -0.126
==============================================================================
```

# 04. 로지스틱 회귀분석의 활용

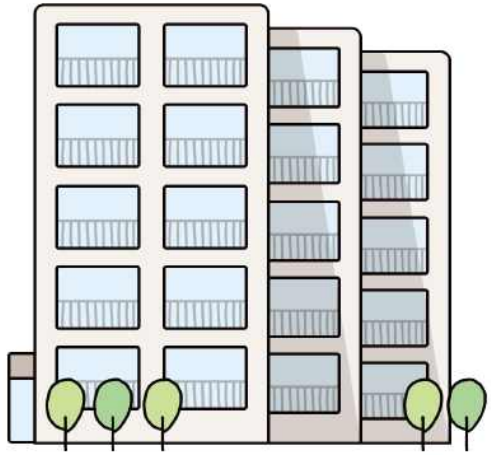
## I. 타이타닉 탑승자 생존여부 예측

- 첫째, 독립변수 중 p-값이 0.05 미만인 변수는 pclass, sex, age, sibsp.
- 이들 변수의 계수가 모두 음수이므로
  독립변수의 값이 증가할 때 생존 가능성이 낮아지는 것으로 판단.
- 둘째, 유의한 독립변수 중 age의 오즈비가 가장 크고, sex의 오즈비가 가장 작음.

# 실전분석 아파트 매매가격의 요인 분석

**[문제]**

아파트 매매가격은 변동하는 값입니다. 어떤 요인이 매매가격에 얼마나 영향을 미치는지 알고 싶습니다. 이럴 때 선형 회귀분석을 수행하여 인과관계를 알아낼 수 있습니다.



**[해결]**

1. 가장 먼저 종속변수와 독립변수를 설정. 아파트 매매가격을 종속변수로 하고, 이에 영향을 미치는 예상 요인들을 독립변수로 함. 아파트 매매가격에 영향을 주는 것은 면적(size), 아파트가 얼마나 오래되었는지(age), 주변 편의시설일 것으로 예상됨.

# 실전분석 아파트 매매가격의 요인 분석

**[해결]**

1. 표는 아파트 매매가격과 앞에서 설정한 독립변수.

| price | size | age | kindergarten | elementarySchool | busStop | hospital | mart |
|---|---|---|---|---|---|---|---|
| 174,000 | 152 | 19 | 22 | 10 | 13 | 19 | 19 |
| 156,500 | 118 | 19 | 22 | 10 | 13 | 19 | 19 |
| 168,000 | 118 | 19 | 22 | 10 | 13 | 19 | 19 |
| 145,000 | 85 | 19 | 22 | 10 | 13 | 19 | 19 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 100,000 | 59 | 11 | 4 | 12 | 29 | 14 | 14 |
| 139,500 | 128 | 11 | 4 | 12 | 29 | 14 | 14 |
| 160,500 | 128 | 11 | 4 | 12 | 29 | 14 | 14 |
| 150,000 | 115 | 11 | 4 | 12 | 29 | 14 | 14 |

# 실전분석 아파트 매매가격의 요인 분석

**[해결]**

2. 선형 회귀모형을 다음과 같이 설정하여 분석을 수행.

```
price ~ size + age + kindergarten + elementarySchool + busStop + hospital + mart
```

```
import pandas as pd
import matplotlib.pylab as plt
from statsmodels.formula.api import ols
from sklearn.linear_model import LinearRegression


price = [174000, 156500, 168000, 145000, (중략), 100000, 139500, 160500, 150000]
size = [152, 118, 118, 85, (중략), 59, 128, 128, 115]
age = [19, 19, 19, 19, (중략), 11, 11, 11, 11]
kindergarten = [22, 22, 22, 22, (중략) , 4, 4, 4, 4]
elementarySchool = [10, 10, 10, 10, (중략) , 12, 12, 12, 12]
busStop = [13, 13, 13, 13, (중략) , 29, 29, 29, 29]
hospital = [19, 19, 19, 19, (중략) , 14, 14, 14, 14]
mart = [19, 19, 19, 19, (중략) , 14, 14, 14, 14]
```

# 실전분석 아파트 매매가격의 요인 분석

## [해결]

3. 각 변수 리스트를 데이터프레임으로 변환하고 선형 회귀분석을 수행.

```
data = {'price': price, 'size': size, 'age': age, 'kindergarten':
kindergarten, 'elementarySchool': elementarySchool, 'busStop': busStop,
'hospital': hospital, 'mart': mart}
df = pd.DataFrame(data)

fit = ols('price ~ size + age + kindergarten + elementarySchool + busStop
+ hospital + mart', data=df).fit( )
print(fit.summary( )))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price    R-squared:                       0.876
Model:                            OLS    Adj. R-squared:                  0.862
Method:                 Least Squares    F-statistic:                     62.45
Date:                Mon, 03 Apr 2023    Prob (F-statistic):           1.11e-25
Time:                        02:26:04    Log-Likelihood:                -734.71
No. Observations:                  70    AIC:                             1485.
Df Residuals:                      62    BIC:                             1503.
Df Model:                           7
Covariance Type:            nonrobust
```

# 실전분석 아파트 매매가격의 요인 분석

## [해결]

3. 각 변수 리스트를 데이터프레임으로 변환하고 선형 회귀분석을 수행.

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.169e+05   1.23e+05      0.948      0.347     -1.3e+05  3.64e+05
size             534.9026     43.081     12.416      0.000      448.785   621.021
age            -1460.9677   1754.535     -0.833      0.408    -4968.233  2046.298
kindergarten    1927.0880    591.638      3.257      0.002      744.421  3109.755
elementarySchool -1599.1185   3858.456     -0.414      0.680    -9312.062  6113.825
busStop          -13.2131    730.790     -0.018      0.986    -1474.042  1447.616
hospital         737.2488    891.948      0.827      0.412    -1045.730  2520.227
mart           -1372.4907   3583.901     -0.383      0.703    -8536.606  5791.625
==============================================================================
Omnibus:                        4.208   Durbin-Watson:     2.150
Prob(Omnibus):                  0.122   Jarque-Bera (JB):  3.332
Skew:                          -0.446   Prob(JB):          0.189
Kurtosis:                       3.589   Cond. No.          1.17e+04
==============================================================================
```

# 실전분석 아파트 매매가격의 요인 분석

**[해결]**

3. 각 변수 리스트를 데이터프레임으로 변환하고 선형 회귀분석을 수행.
   수정된 결정계수 adj.$R^2$가 0.862이므로 이 모형은 86.2%의 설명력을 갖췄음.
   유의한 변수는 p-값이 0.000인 size와 0.002인 kindergarte뿐.
   변수 size의 계수는 534.9026이며, kindergarten의 계수는 1927.0880.

$$price = 0.0000169 + (534.9026) \times size + (1927.0880) \times kindergarten + \varepsilon$$

4. 모형 해석:
   아파트 매매가격에 영향을 미치는 요인은 면적과 유치원까지의 거리. 그러나 특정 지점까지 도보 소요시간이 길어질수록 매매가격이 비싸진다는 해석은 일반적이지 않음.
   따라서 데이터의 시간적, 공간적 범위를 넓히고 독립변수도 추가하여 다시 분석할 필요가 있음.

# Thank You!

한빛아카데미 Hanbit Academy, Inc.