

AI 기초

2019-2020

강봉주

다중 회귀 분석

회귀 분석

[필요한 패키지]

```
In      # 필요한 패키지
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

import scipy
from scipy import linalg as la
import scipy.stats as ss
import scipy.special

# 한글출력
plt.rcParams['font.family'] = 'Malgun Gothic'
plt.rcParams['axes.unicode_minus'] = False

# 필요한 패키지 2
import statsmodels.formula.api as smf
import statsmodels.api as sm

sm.__version__
Out      '0.10.1'
```

회귀 분석

[개요]

- 독립변수가 2개 이상인 다중 선형회귀(multiple linear regression)
- 모형식: $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, i = 1, \dots, n$
- 오차항에 대한 가정: $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$

회귀 분석

[모수의 추정]

- $\{(x_i, y_i) | i = 1, \dots, n\}$: 표본

$$y_1 = \beta_0 + \beta_1 x_{11} + \dots + \beta_k x_{1k} + \epsilon_1$$

▪

...

$$y_n = \beta_0 + \beta_1 x_{n1} + \dots + \beta_k x_{nk} + \epsilon_n$$

- $y_i = x_i^T \beta + \epsilon_i, h(x_i) = x_i^T \beta$

- 행렬 표현식: $y = X\beta + \epsilon$

- $y = (y_1, \dots, y_n)^T, X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \beta = (\beta_0, \beta_1, \dots, \beta_k)^T, \epsilon = (\epsilon_1, \dots, \epsilon_n)^T$

회귀 분석

[모수의 추정]

- $\underset{\beta}{\text{minimize}}((y - X\beta)^T(y - X\beta))$
- β 에 대하여 미분하여 정리
- $X^T X \beta = X^T y$: 정규 방정식
- $\hat{\beta} = (X^T X)^{-1} X^T y$: 역행렬이 존재하면.
- 역행렬을 이용하지 않은 경우: QR 분해 기법 이용

회귀 분석

[모수의 추정]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 모수의 추정값을 구해보자.

회귀 분석

[모수의 추정]

예제 [HOUSING] 자료에서 크기가 100 인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 모수의 추정값을 구해보자.

```
In      # 표본 크기
        size = 100

        # 씨앗값 정의
        np.random.seed(123)
        index = np.random.choice(np.arange(len(df)), size=size, replace=False)

        # 표본 구성
        xvars = ['rm', 'age', 'lstat']
        target = 'medv'
        sdf = df.loc[index, xvars+[target]]
        sdf.shape

Out      (100, 4)
```


회귀 분석

[모수의 추정]

예제 [HOUSING] 자료에서 크기가 100 인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 모수의 추정값을 구해보자.

```
In    ## QR 분해를 통한 모수의 추정
      X = sdf[xvars]
      X = sm.add_constant(X).values
      y = sdf[target].values
      Q, R = la.qr(X.T@X)

      beta = la.inv(R) @ Q.T @ X.T @ y
      beta.round(3)

Out   array([[ 1.3063e+01],
             [ 2.9310e+00],
             [ 3.0000e-03],
             [-7.1100e-01]])
```

회귀 분석

[모수의 추정]

예제 [HOUSING] 자료에서 크기가 100 인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 모수의 추정값을 구해보자.

```
In      ## statsmodels 에 의한 모수의 추정
        # 기호식 구성
        fmla = target + '~' + '+'.join(xvars)
        fmla
Out      'medv~rm+age+lstat'
```

회귀 분석

[모수의 추정]

예제 [HOUSING] 자료에서 크기가 100 인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 모수의 추정값을 구해보자.

```
In      # 적합
        fit = smf.ols(fmla, data=sdf).fit()
        fit.params.round(3)
Out      Intercept    13.063
         rm           2.931
         age           0.003
         lstat        -0.711
         dtype: float64dtype: float64
```

회귀 분석

[추정된 예측값]

- $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$
- y 를 \hat{y} 로 보내주는 하나의 변환: 모자(hat) 행렬
- 모자 행렬의 성질

1. $H^T = H$: 대칭(symmetric)

2. $HX = X$

3. $HH = H^2 = H$: 멱등행렬(idempotent)

4. $(I - H)(I - H) = (I - H)^2 = I - H$

회귀 분석

[오차 분산의 추정]

- $\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-p} = \frac{SSE}{n-p}$
- $$\begin{aligned} \frac{1}{n-p} (y - X\hat{\beta})^T (y - X\hat{\beta}) &= \frac{1}{n-p} (y - Hy)^T (y - Hy) \\ &= \frac{1}{n-p} y^T (I - H)^T (I - H) y \\ &= \frac{1}{n-p} y^T (I - H) y \end{aligned}$$

회귀 분석

[오차 분산의 추정]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 오차분산의 추정값을 구해보자.

회귀 분석

[오차 분산의 추정]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 오차분산의 추정값을 구해보자.

```
In      # 직접 계산
        H = X @ la.inv(X.T @ X) @ X.T
        n = H.shape[0]
        p = 1 + 3
        I = np.identity(n)
        MSE = 1/(n - p) * y.T @ (I-H) @ y
        MSE.round(3)
Out      array([[41.496]])
```

회귀 분석

[오차 분산의 추정]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 오차분산의 추정값을 구해보자.

In	# 결과 확인 fit.mse_resid.round(3)
Out	41.496

회귀 분석

[모수의 추론]

- $\epsilon \sim N(0, I\sigma^2)$
- $E(Y) = E(X\beta + \epsilon) = X\beta$
- $Var(Y) = Var(X\beta + \epsilon) = Var(\epsilon) = I\sigma^2$
- $$\begin{aligned} Cov(AY) &= E\left((AY - E(AY))(AY - E(AY))^T\right) \\ &= E\left(A(Y - E(Y))(Y - E(Y))^T A^T\right) \\ &= A Cov(Y) A^T \end{aligned}$$

회귀 분석

[모수의 추론]

- $Y \sim N(X\beta, I\sigma^2)$
- $E(\hat{\beta}) = E\left((X^T X)^{-1} X^T Y\right) = (X^T X)^{-1} X^T E(Y) = \beta$
- $$\begin{aligned} \text{Cov}(\hat{\beta}) &= \text{Cov}\left((X^T X)^{-1} X^T Y\right) = (X^T X)^{-1} X^T \text{Cov}(y) \left((X^T X)^{-1} X^T\right)^T = \\ &= (X^T X)^{-1} X^T \{I\sigma^2\} X (X^T X)^{-1} = (X^T X)^{-1} \sigma^2 \end{aligned}$$
- $\hat{\beta} \sim N\left(\beta, (X^T X)^{-1} \sigma^2\right), \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p), \hat{\beta}, \hat{\sigma}^2 \text{ 이 서로 독립}$

회귀 분석

[모수의 추론]

- $\sqrt{(X^T X)^{-1}} \stackrel{let}{=} D$
- $t_k = \frac{\hat{\beta}_k - \beta_k}{D_{kk} \hat{\sigma}} \sim t(n - p)$
- $H_0: \beta_k = 0$
- $|t_{k0}| \geq t_{\frac{\alpha}{2}}(n - p)$, 영가설 기각, α 는 유의 수준

회귀 분석

[모수의 추론]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 회귀계수에 대한 유의성을 검증해보자.

회귀 분석

[모수의 추론]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 회귀계수에 대한 유의성을 검증해보자.

```
In      # 검정 통계량 계산
        beta_hat = la.inv(X.T @ X) @ X.T @ y
        D = np.sqrt(la.inv(X.T @ X))

        t_zero = beta_hat / (np.diag(D).reshape(-1, 1) * np.sqrt(MSE))
        t_zero
Out      array([[ 1.40846797],
                [ 2.20709057],
                [ 0.08976109],
                [-4.61866789]])
```

회귀 분석

[모수의 추론]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 회귀계수에 대한 유의성을 검증해보자.

```
In      # p-values 계산
        pvalue = 2*(1 - ss.t.cdf(np.abs(t_zero), df=n-p))
        pvalue
Out      array([[1.62223028e-01],
                [2.96907987e-02],
                [9.28664075e-01],
                [1.19892794e-05]])
```

회귀 분석

[모수의 추론]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 회귀계수에 대한 유의성을 검증해보자.

In	# 결과 확인
	fit.tvalues
Out	Intercept 1.408468
	rm 2.207091
	age 0.089761
	lstat -4.618668
	dtype: float64

회귀 분석

[모수의 추론]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 회귀계수에 대한 유의성을 검증해보자.

In	# 결과 확인
	fit.pvalues
Out	Intercept 0.162223
	rm 0.029691
	age 0.928664
	lstat 0.000012
	dtype: float64

회귀 분석

[회귀직선의 유의성 검증: 분산 분석표]

- $F = \frac{SSR/k\sigma^2}{SSE/(n-k-1)\sigma^2} \sim F(k, n-k-1)$, k 는 절편을 제외한 모수의 개수

요인	제곱합	자유도	평균제곱	F값	P값
회귀	SSR	k	$MSR = \frac{SSR}{k}$	$f_0 = \frac{MSR}{MSE}$	$\Pr(F \geq f_0)$
잔차	SSE	$n - k - 1$	$MSE = \frac{SSE}{n - k - 1}$		
계	SST	$n - 1$			

회귀 분석

[회귀직선의 유의성 검증: 분산 분석표]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 분산분석 방법에 의한 회귀직선의 유의미성을 검증해보자.

회귀 분석

[회귀직선의 유의성 검증: 분산 분석표]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 분산분석 방법에 의한 회귀직선의 유의미성을 검증해보자.

```
In      # 변수 정의
        xvars = ['rm', 'age', 'lstat']
        target = 'medv'

        # 자유도 정의
        n = len(sdf)
        p = len(xvars) + 1
        [n, p]
Out      [100, 4]
```

회귀 분석

[회귀직선의 유의성 검증: 분산 분석표]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 분산분석 방법에 의한 회귀직선의 유의미성을 검증해보자.

```
In      # 제곱합 계산
        y = sdf[target]
        SSE = np.sum((y - fit.fittedvalues)**2)
        SST = np.sum((y - np.mean(y))**2)
        SSR = SST - SSE

        MSE = SSE / (n-p)
        MSR = SSR / (p-1)
        np.round([MSE, MSR], 3)
Out      array([ 41.496, 1468.649])
```

회귀 분석

[회귀직선의 유의성 검증: 분산 분석표]

예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 분산분석 방법에 의한 회귀직선의 유의미성을 검증해보자.

```
In      # F 값 계산
        f_zero = MSR/MSE

        # p-value 계산
        pvalue = 1 - ss.f.cdf(f_zero, dfn=p-1, dfd=n-p)

        np.round([f_zero, pvalue], 3)
Out      array([35.392,  0.   ])
```

회귀 분석

[회귀직선의 유의성 검증: 분산 분석표]

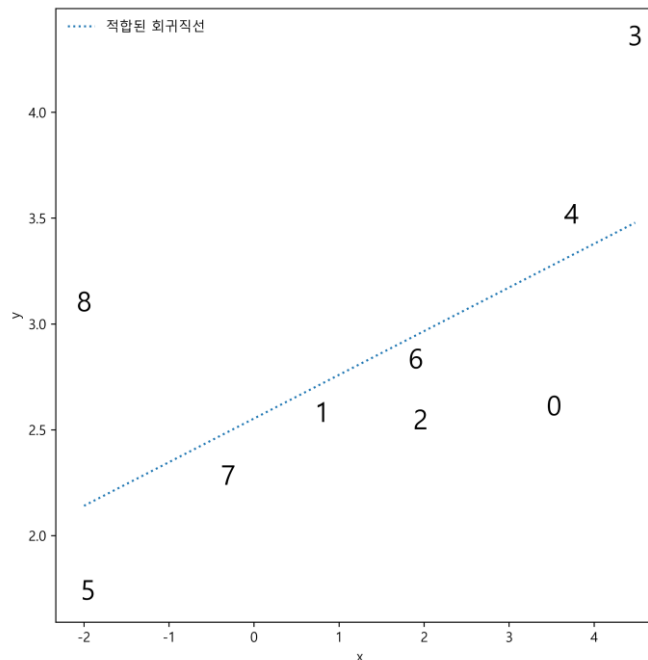
예제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 분산분석 방법에 의한 회귀직선의 유의미성을 검증해보자.

```
In      # 결과 확인
        np.round([fit.fvalue, fit.f_pvalue], 3)
Out      array([35.392,  0.   ])
```

회귀 분석

[이상점 판정]

- 이상값(outlier)은 평균으로부터 멀리 떨어진(outlying) 값을 의미한다.
선형회귀 분석에서는 오차 값이 매우 큰 값을 의미한다.



회귀 분석

[이상점 판정]

- $\text{var}(r_i) = \sigma^2(1 - H_{ii})$: 잔차 분산
- $t_i = \frac{r_i}{\hat{\sigma}\sqrt{1-H_{ii}}}$: 표준화 잔차
- $t_{(i)} = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-H_{ii}}}$: 외부 스튜던트화 잔차, $\hat{\sigma}_{(i)}$ 는 (y_i, x_i) 관측값을 제거한 후 계산한 표준오차
- $\hat{\sigma}_{(i)}^2 = \frac{SSE_{(i)}}{n-k-2} = \frac{SSE - \frac{r_i^2}{1-H_{ii}}}{n-k-2}$: 계산식

회귀 분석

[이상점 판정]

- 표준화 잔차와 외부 스튜던트화 잔차

In	# 이상점 파악 print(sfit.get_influence().summary_frame()[['standard_resid', 'student_resid']])		
Out	standard_resid	student_resid	
	0	-1.256752	-1.322216
	1	-0.258594	-0.240563
	2	-0.737492	-0.710964
	3	1.739347	2.137032
	4	0.344098	0.321302
	5	-0.851857	-0.833026
	6	-0.212793	-0.197648
	7	-0.392486	-0.367437
	8	1.945842	2.658769

회귀 분석

[이상점 판정]

- 표준화 잔차와 외부 스튜던트화 잔차

In	<pre># p-value pvalue = 2*(1 - ss.t.cdf(np.abs(sfit.get_influence().summary_frame()['standard_resid']), df=len(x)-2)) pvalue.round(3)</pre>
Out	<pre>array([0.249, 0.803, 0.485, 0.126, 0.741, 0.422, 0.838, 0.706, 0.093])</pre>

회귀 분석

[이상점 판정]

과제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 표준화 잔차 값이 5% 유의수준에서 유의미한 이상점을 찾아보자. 단, 표본추출시 난수 씨앗값은 123으로 하며, 중복을 허용하지 않는다.

회귀 분석

[영향점 판정]

- 영향점은 회귀계수의 추정값($\hat{\beta}$)과 예측값($X\hat{\beta}$)에 영향을 주는 관측값
- $\hat{y}_i = \sum_j H_{ij}y_j = H_{ii}y_i + \sum_{j \neq i} H_{ij}y_j$
- $H^2 = H$ 에서 $h_{ii} = h_{ii}^2 + h_{ij}^2 + \sum_{k \neq i,j} h_{ik}^2$
- 모자 행렬의 성질
 - 1) $\frac{1}{n} \leq H_{ii} \leq 1, \forall i$
 - 2) $-0.5 \leq H_{ij} \leq 0.5, \forall j \neq i$
 - 3) $\text{tr}(H) = \sum_i H_{ii} = k + 1$

회귀 분석

[영향점 판정]

- 영향점은 회귀계수의 추정값($\hat{\beta}$)과 예측값($X\hat{\beta}$)에 영향을 주는 관측값
- $\hat{y}_i = \sum_j H_{ij}y_j = H_{ii}y_i + \sum_{j \neq i} H_{ij}y_j$
- H_{ii} 가 매우 크면 가령 1에 가까우면 다른 H_{ij} 는 모두 매우 작아진다. 즉, \hat{y}_i 의 값에 y_i 가 지배적으로 영향을 끼친다. 여기서 H_{ii} 를 y_i 의 지렛대(leverage) 이라고 부른다.

회귀 분석

[영향점 판정]

- 영향점은 회귀계수의 추정값($\hat{\beta}$)과 예측값($X\hat{\beta}$)에 영향을 주는 관측값
- $\hat{y}_i = \sum_j H_{ij}y_j = H_{ii}y_i + \sum_{j \neq i} H_{ij}y_j$
- $H_{ii} > \frac{2(k+1)}{n}$ 인 점을 고 지렛대점

회귀 분석

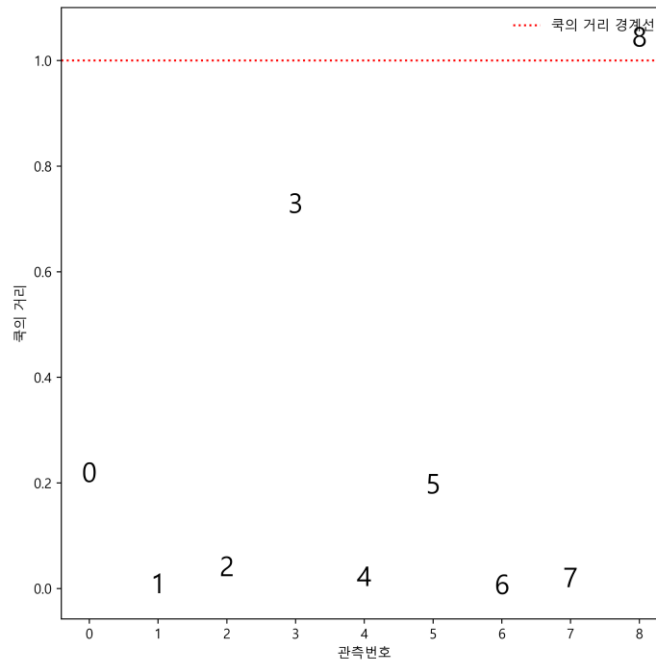
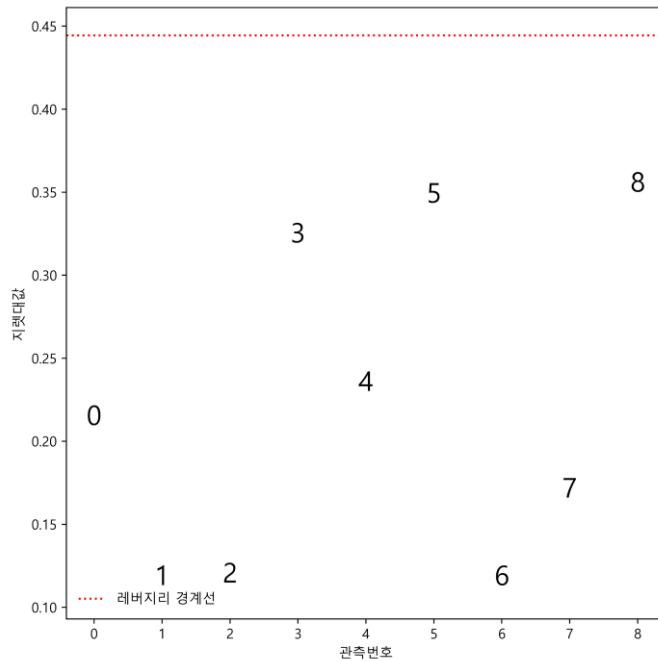
[영향점 판정]

- 영향점은 회귀계수의 추정값($\hat{\beta}$)과 예측값($X\hat{\beta}$)에 영향을 주는 관측값
- 쿡의 거리(Cook's distance)
- i 관측값을 제거한 후 추정된 값을 $\hat{\beta}_{(i)}$, $\hat{y}_{(i)}$
- $$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{(k+1)\hat{\sigma}^2}$$
- $$D_i = \frac{(\hat{y}_{(i)} - \hat{y})^T (\hat{y}_{(i)} - \hat{y})}{(k+1)\hat{\sigma}^2}$$
- $$D_i = \frac{r_i^2}{k+1} \left(\frac{H_{ii}}{1-H_{ii}} \right)$$
- $D_i > 1$: 영향점

회귀 분석

[영향점 판정]

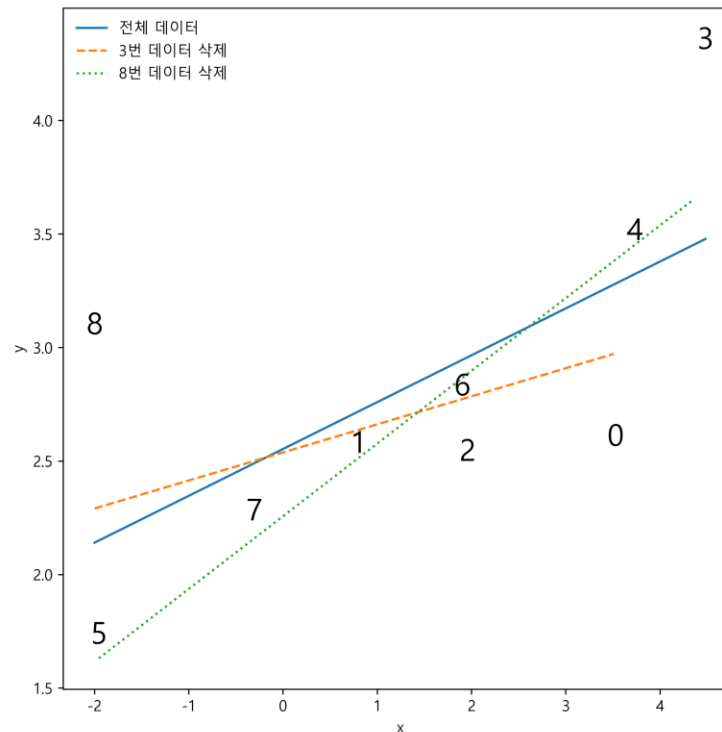
- 영향점은 회귀계수의 추정값($\hat{\beta}$)과 예측값($x\hat{\beta}$)에 영향을 주는 관측값



회귀 분석

[영향점 판정]

- 영향점은 회귀계수의 추정값($\hat{\beta}$)과 예측값($x\hat{\beta}$)에 영향을 주는 관측값



회귀 분석

[영향점 판정]

과제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat'] 으로 하여 선형회귀 모형을 적합 시킬 때 쿡의 거리 기준으로 하여 이상점을 찾아보자. 단, 표본추출시 난수 씨앗값은 123으로 하며, 중복을 허용하지 않는다.

회귀 분석

[다중공선성 판정]

- 공선성(collinearity)은 2개의 독립변수 간의 선형적인 관계
- 다중공선성(multicollinearity)은 2개 이상의 독립 변수들 간의 선형적인 관계
- 데이터 행렬의 행렬 계수(rank)가 완전 계수(full rank)가 되지 않으므로 $X^T X$ 의 역행렬이 존재하지 않음. 의사 역행렬로 계산은 가능하나 모수 추정값에 대하여 매우 큰 값의 표준오차가 발생하여 그 결과를 믿을 수 없음

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 3 \\ 1 & 4 & 4 \end{bmatrix} \quad X^T X = \begin{bmatrix} 4 & 10 & 10 \\ 10 & 30 & 30 \\ 10 & 30 & 30 \end{bmatrix} \quad R = \begin{bmatrix} -14.7 & -43.55 & -43.55 \\ 0 & -1.92 & -1.92 \\ 0 & 0 & 0 \end{bmatrix}$$

회귀 분석

[다중공선성 판정]

- 공선성(collinearity)은 2개의 독립변수 간의 선형적인 관계
- 다중공선성(multicollinearity)은 2개 이상의 독립 변수들 간의 선형적인 관계
- 다중공선성을 탐지하는 방법: 분산팽창계수(VIF: variance inflation factor)

$$VIF_j = \frac{1}{1 - R_j^2}$$

- R_j^2 는 설명 변수 j 를 종속 변수로 하고 나머지 설명 변수들을 설명 변수로 하여 회귀분석을 실시한 경우의 결정계수
- $R_j^2 = 0.9$ 보다 크게 되면 VIF_j 는 10보다 크게 됨
- $VIF_j > 10$: 다중공선성이 있으므로 해당 변수 제거

회귀 분석

[다중공선성 판정]

예제

데이터 행렬 $\begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 4 \end{bmatrix}$ 에 대하여 VIF를 구해보자.

```
In      # 필요한 패키지
        from statsmodels.stats.outliers_influence import variance_inflation_factor

        # 다중 공선성 계산
        data_matrix = np.array([1, 2, 3, 4, 1, 2, 3, 4]).reshape(2, -1).T
        vif = [(i, variance_inflation_factor(data_matrix, i)) for i in np.arange(2)]
        vif

Out      [(0, inf), (1, inf)]
```

회귀 분석

[다중공선성 판정]

과제 [HOUSING] 자료에서 크기가 100인 표본을 추출하여 종속변수를 주택중위가격(medv), 독립변수를 ['rm', 'age', 'lstat']으로 하여 선형회귀 모델을 적합 시킬 때 독립변수 간의 다중공선성을 확인해보자.