

START

머신러닝과 딥러닝

Machine Learning & deep Learning

02

Chapter 2. 선형회귀분석 I

Machine Learning & Deep Learning

손영두

e-mail: youngdoo@dongguk.edu



회귀분석

회귀분석

$$y_i = a + b_1x_{1i} + b_2x_{2i} + \dots + b_kx_{ki} + e_i$$

✓ 독립변수와 종속변수 사이의 관계를 선형의 관계로 가정

✓ 독립변수의 종류

- 양적 입력
- 양적 입력의 변환 (log, 루트 등)
- 입력 변수의 다항식 (2차, 3차 등)
- 두 변수 사이의 교호작용 ($X_3 = X_1 * X_2$)
- 질적인 입력을 위한 dummy variable (one-hot encoding)

COLOR
Red
Red
Yellow
Green
Yellow



RED	YELLOW	GREEN
1	0	0
1	0	0
0	1	0
0	0	1



단순회귀분석

두 변수 X, Y 의 n 개의 확률표본 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 에 대한 관계를 다음과 같이 직선으로 가정한다.

$$y_i = a + bx_i + e_i, \quad i = 1, 2, \dots, n$$

이때 $e_i \sim N(0, \sigma^2)$ 이고 서로 독립이며 a, b 는 미지의 모수이다. 그리고 y_i 는 종속변수로 독립변수 x_i 에 따라 결정되는 값이다.



단순회귀분석

■ 단순회귀분석의 적합 (최소제곱 계수 추정)

식 (7.2)의 모회귀계수인 a, b 는 최소제곱법(method of least square)에 의해 다음의 오차의 제곱합 D 를 최소화하는 추정량으로 추정한다.

$$D = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

이를 위해서 우선 오차의 제곱합 D 를 다음과 같이 a, b 로 각각 편미분 한 후, 0으로 놓고 풀면

$$\frac{\partial D}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$\frac{\partial D}{\partial b} = -2 \sum_{i=1}^n x_i (y_i - a - bx_i) = 0$$

a, b 의 추정량을 \hat{a}, \hat{b} 로 하는 다음의 정규방정식(normal equation)을 구할 수 있다.

$$n\hat{a} + \hat{b} \sum x_i = \sum y_i$$

$$\hat{a} \sum x_i + \hat{b} \sum x_i^2 = \sum x_i y_i$$



단순회귀분석

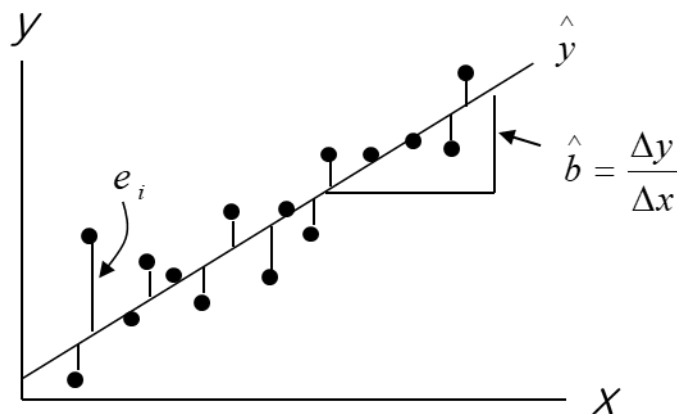
단순회귀분석: 단순회귀분석의 적합 (최소제곱 계수 추정)

상기의 식을 풀면 오차의 제곱합 D 를 최소로 하는 최소제곱 추정량 \hat{a}, \hat{b} 은 다음과 같다.

$$D = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

이들 값은 [그림 7-4] 에서 각각의 관측값에서 부터 추정된 모델상의 예측 값 까지의 거리의 합을 최소로 하는 값이다.



[그림 7-4] 최소제곱법에 의한 모델의 계산



단순회귀분석

■ 단순회귀분석: 단순회귀분석의 적합 (최소제곱 계수 추정)

따라서 회귀직선의 추정식을 다음과 같이 나타낼 수 있는데 이 직선이 x, y 의 산점도에 상에서 x, y 관계를 가장 잘 대표한다고 할 수 있다.

$$\hat{y} = \hat{a} + \hat{b}x$$

여기서 \hat{a} 은 $x=0$ 에서 \hat{y} 의 절편이고, \hat{b} 은 기울기 이다.



단순회귀분석의 가정

단순회귀분석의 가정

단순회귀모형 대한 기본가정을 세분하여 다시 정리하면 다음과 같다.

단순 회귀모형의 기본가정

- ① 두 변수 X 와 Y 간에는 직선관계가 성립되어야 한다.
- ② 오차들은 평균이 0이고 분산이 σ^2 인 정규분포를 따라야 한다.
- ③ 오차들의 분산은 σ^2 로 같아야 한다.
- ④ 오차들은 서로 독립이어야 한다.

적합된 회귀식이 상기의 기본가정을 만족하는지의 여부를, 기울기 b 에 대한 유의성 검정이나 결정계수 R^2 을 구함으로써, 어느 정도 파악할 수 있다.

그러나 경우에 따라서는 결정계수가 크에도 불구하고 상기의 단순회귀모형의 기본가정이 위배되는 경우가 있다.

따라서 회귀분석을 통해 모델을 구한 후에는 모델이 충분히 잘 적합되었는지의 여부를 판정하기 위해서 반드시 잔차분석을 실시해야 한다.



단순회귀분석

결정계수 (coefficient of determination)

데이터들의 변동 중 회귀 분석에 의해 설명되는 부분을 결정계수 라고 하며, 따라서 결정계수가 클수록 회귀식이 관측 데이터를 잘 설명하는 것으로 적합성이 보증된다고 할 수 있다. 그러나 이 경우에도 잔차분석을 통하여 적합성을 검토할 필요가 있다.

$$✓ \text{ 총변동 : } SST = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2$$

$$✓ \text{ 회귀변동 : } SSR = \sum (\hat{y}_i - \bar{y})^2 = \sum (\hat{a} + \hat{b}x_i - \bar{y})^2 = \sum (\bar{y} - \hat{b}\bar{x} + \hat{b}x_i - \bar{y})^2 \\ \sum [\hat{b}(x_i - \bar{x})]^2 = \hat{b}^2 \sum (x_i - \bar{x})^2 = \hat{b}^2 S_{xx}$$

$$✓ \text{ 잔차변동 : } SSE = SST - SSR = \left(\frac{S_{xy}}{S_{xx}} \right)^2 \cdot S_{xx} = \frac{S_{xy}^2}{S_{xx}}$$

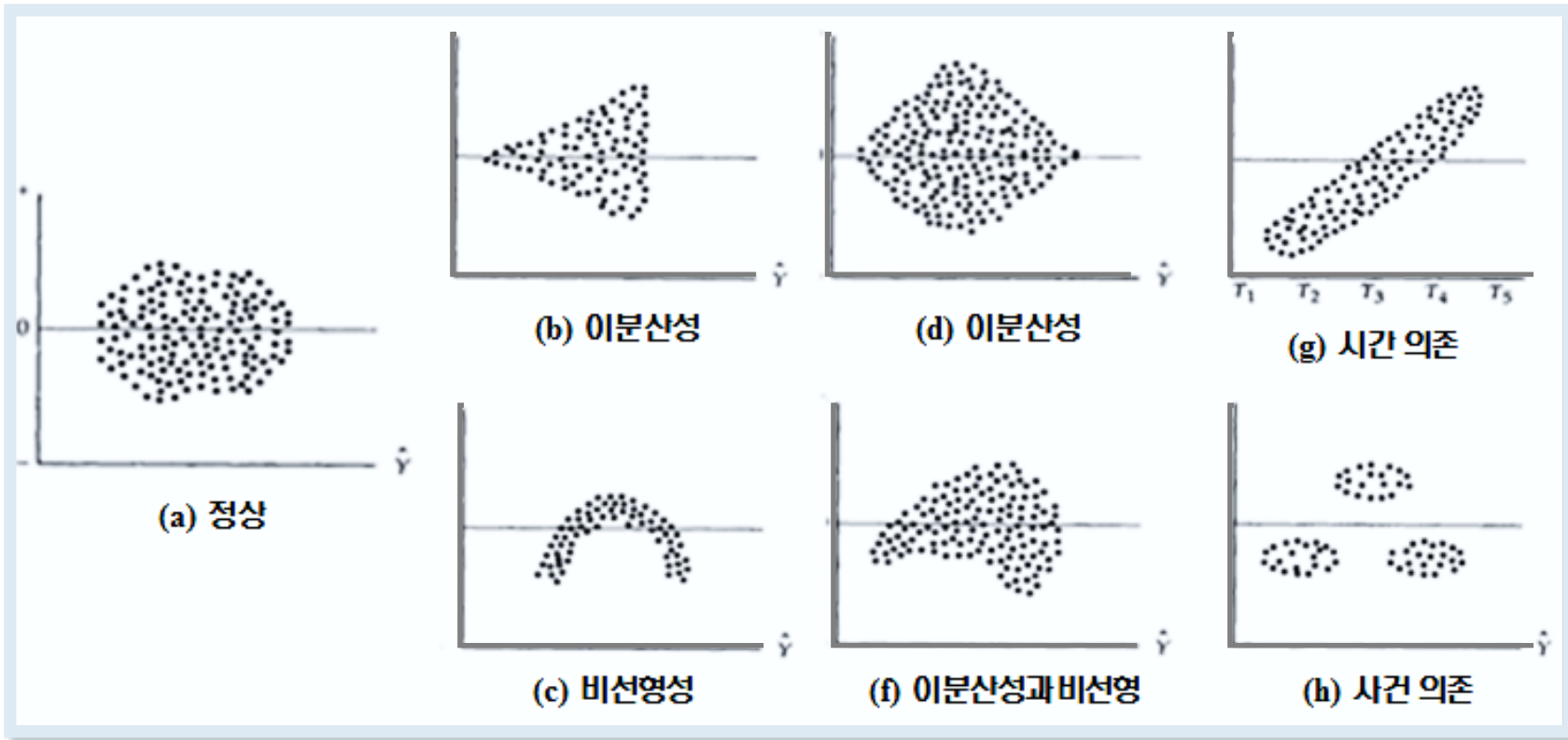
$$✓ \text{ 결정계수 : } R^2 = \frac{SSR}{SST} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$



단순회귀분석

잔차분석

다음의 그림에서와 같이 가로축을 추정값 \hat{y}_i 세로축을 잔차 e_i 로 하여 잔차 산점도를 작성 할 경우 만일 모델이 잘 적합 되었다면 (a)와 같이 잔차는 어떤 경향도 없이 랜덤하게 분포 할 것이다.





단순회귀분석

예제

✓ 다음의 데이터를 활용하여 단순회귀분석을 해보자.

(단위: 백만원)

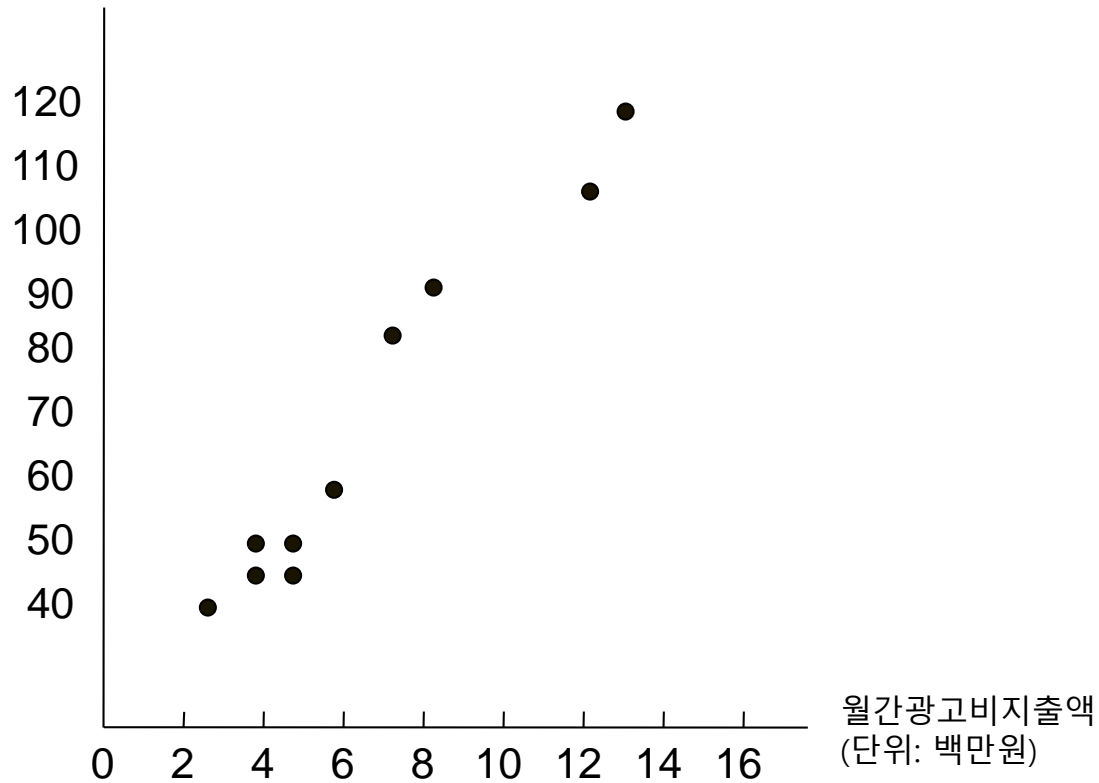
월 (i)	월간광고비지출액 (x_i)	월간매출액 (y_i)
1	3	40
2	4	50
3	5	45
4	4	45
5	5	50
6	6	55
7	7	70
8	8	85
9	12	100
10	13	115



단순회귀분석

■ **예제(계속)** ✓ 데이터의 대략적인 산포도 및 분석을 위한 계산 값은 다음과 같다.

월간매출액(단위: 백만원)



$x_i y_i$	x_i^2	y_i^2
120	9	1,600
200	16	2,500
225	25	2,025
180	16	2,025
250	25	2,500
330	36	3,025
490	49	4,900
680	64	7,225
1,200	144	10,000
1,495	169	13,225



단순회귀분석

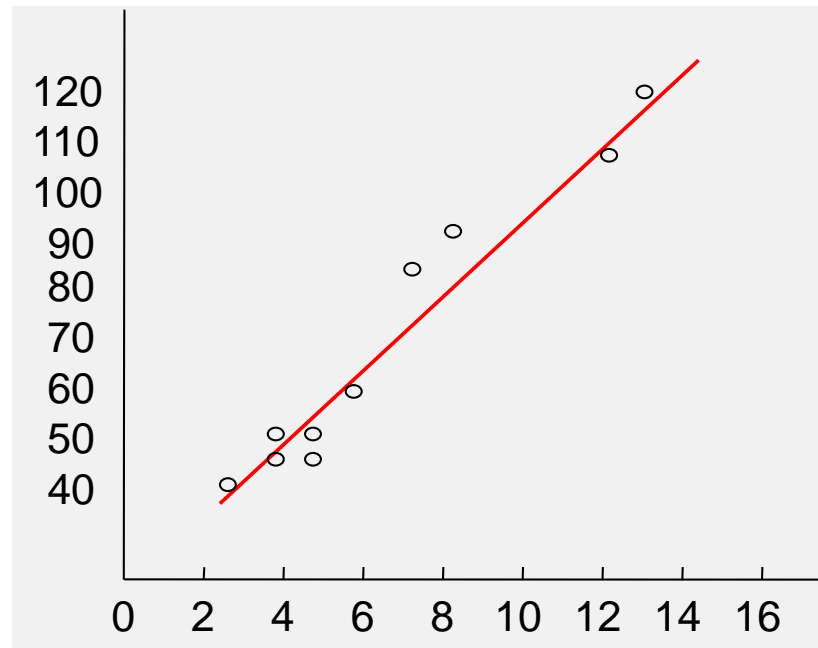
■ **예제계속** ✓ 회귀방정식은 다음과 같이 구할 수 있다.

$$\hat{y} = \hat{a} + \hat{b}x$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{5170 - 10(6.7)(65.5)}{553 - 10(6.7)^2} = 7.5072$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 65.5 - (7.5072)(6.7) = 15.2017$$

$$\therefore \hat{y} = 15.2017 + 7.5072x$$





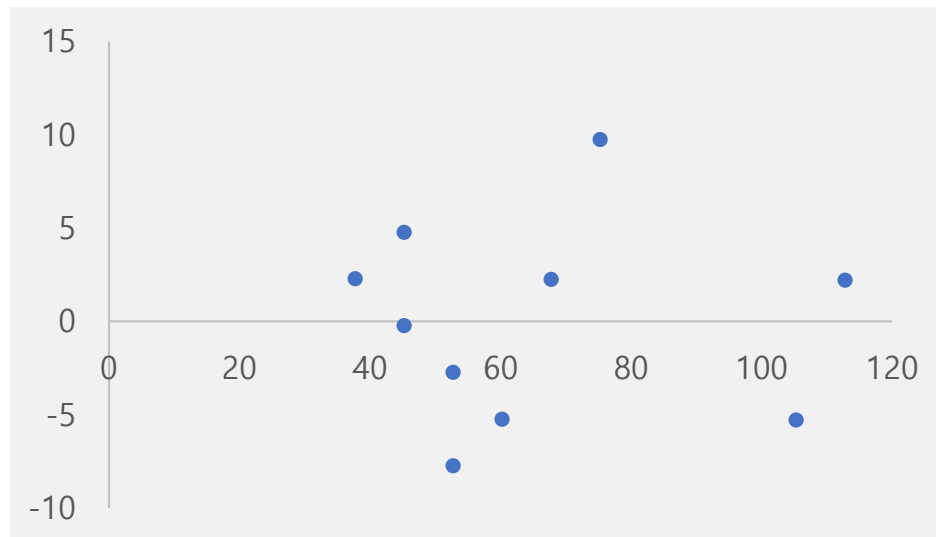
단순회귀분석

- **예제(계속)** ✓ 결정계수와 잔차분석표는 다음과 같다. 결정계수가 크고, 잔차분포에서 특별한 경향이 나타나지 않은 것으로 보아 적합한 선형회귀방식으로 볼 수 있다.

i	\hat{y}	잔차
1	37.72334	2.276657
2	45.23055	4.769452
3	52.73775	-7.73775
4	45.23055	-0.23055
5	52.73775	-2.73775
6	60.24496	-5.24496
7	67.75216	2.247839
8	75.25937	9.740634
9	105.2882	-5.28818
10	112.7954	2.204611



$$R^2 = \frac{SSR}{SST} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} \approx 0.958$$





다중회귀분석

다중회귀분석

- ✓ 다수의 중요 인자를 독립변수가 중요할 경우, 이들을 다중회귀 분석을 통해 분석을 한다. 다음과 같이 2개 이상의 독립변수와 종속변수와의 관계를 선형으로 가정하는 회귀모형을 다중회귀모형(multiple linear regression model)이라고 한다.

$$y_i = a + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki} + e_i$$

$$e_i \sim N(0, \sigma^2) \text{ 이고 서로 독립}$$

$$i = 1, 2, \cdots, n$$



다중회귀분석

다중회귀분석의 적합: 최소제곱법

- ✓ 다중회귀의 경우 교호작용이나 제곱항이 포함된다는 것 이외에는 모델을 구하는 원리는 단순회귀의 경우와 동일하다.
- ✓ 식 (7.1)을 행렬로 표현하면 다음과 같다.

여기서

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} a \\ b_1 \\ \vdots \\ b_k \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

최소제곱법을 적용하기 위해 오차의 제곱합을 행렬로 표현하면 다음과 같다.

$$\begin{aligned} D &= \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \end{aligned}$$

상기의 식을 \mathbf{b} 로 미분하여 0으로 놓으면 다음의 최소제곱추정치를 얻을 수 있다.

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{y}$$



참고: 행렬 및 벡터의 미분

- The Matrix Cookbook

<https://www.ics.uci.edu/~welling/teaching/KernelsICS273B/MatrixCookBook.pdf>

$$\left[\frac{\partial \mathbf{x}}{\partial y} \right]_i = \frac{\partial x_i}{\partial y} \quad \left[\frac{\partial x}{\partial \mathbf{y}} \right]_i = \frac{\partial x}{\partial y_i} \quad \left[\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right]_{ij} = \frac{\partial x_i}{\partial y_j}$$

$$f = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}$$

$$\nabla_{\mathbf{x}} f = \frac{\partial f}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T) \mathbf{x} + \mathbf{b}$$

$$\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T} = \mathbf{A} + \mathbf{A}^T$$



다중회귀분석

다중회귀분석의 적합: 최소제곱법

✓ 여기서 $\mathbf{X}\mathbf{X}$ 와 $\mathbf{X}\mathbf{y}$ 는 각각 다음과 같다.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum x_{1i} & \cdots & \sum x_{ki} \\ \sum x_{1i} & \sum x_{1i}^2 & \cdots & \sum x_{1i}x_{ki} \\ \vdots & \vdots & & \vdots \\ \sum x_{ki} & \sum x_{1i}x_{ki} & \cdots & \sum x_{ki}^2 \end{bmatrix} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \vdots \\ \sum x_{ki} y_i \end{bmatrix}$$

✓ 식(7-2)의 양변에 $(\mathbf{X}'\mathbf{X})^{-1}$ 을 곱하면 다음과 같이 오차의 제곱합을 최소로 하는 b 의 추정값을 얻을 수 있다.

$$\hat{\mathbf{b}} = \begin{bmatrix} \hat{a} \\ \hat{b}_1 \\ \hat{b}_2 \\ \vdots \\ \hat{b}_k \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

✓ 따라서 추정식은 다음이 된다.

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$$



다중회귀분석

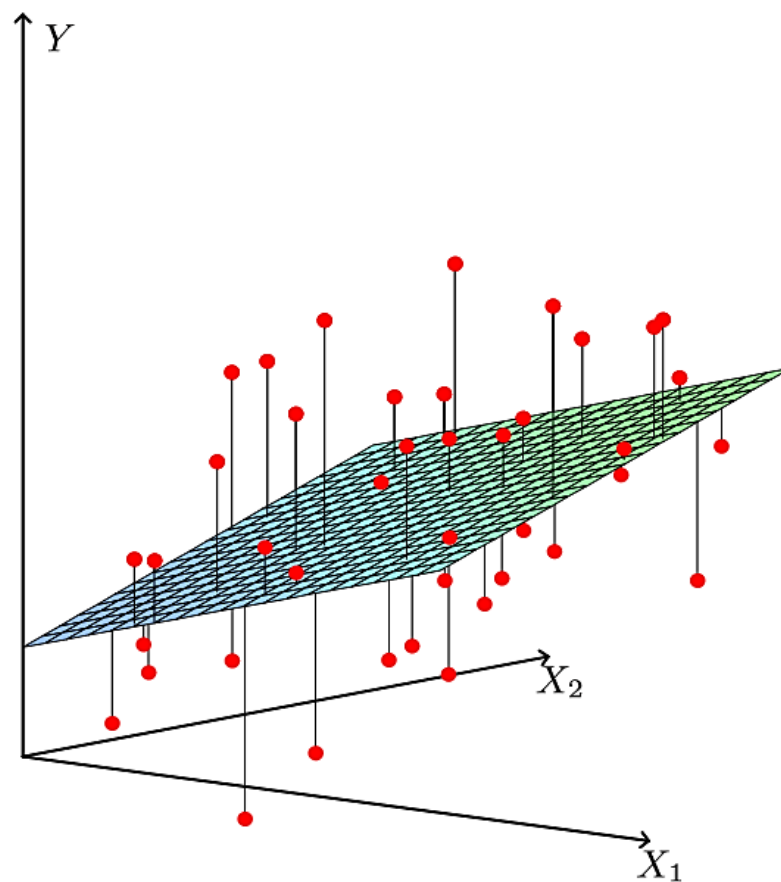


FIGURE 3.1. Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y .



다중회귀분석

예제

✓ 다음의 데이터를 활용하여 다중회귀 추정식을 구해보자.

월 (i)	관할 인구 (단위: 천명) (x_{1i})	평균월수입 (단위: 백원) (x_{2i})	총판매액 (단위: 십만원) (y_i)
1	274	2450	162
2	180	3254	120
3	375	3802	223
4	205	2838	131
5	86	2347	67
6	265	3782	169
7	98	3008	81
8	330	2450	192
9	195	2137	116
10	53	2560	55
11	430	4020	252
12	372	4427	232
13	236	2660	144
14	157	2088	103
15	370	2605	212



다중회귀분석

■ 예제

✓ 오차의 제곱합을 최소화 하는 b의 추정 값은 다음과 같다.

$$\hat{y} = X\hat{b}$$

$$\hat{b} = \begin{bmatrix} \hat{a} \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = (X'X)^{-1}X'y = \begin{bmatrix} 3.452613 \\ 0.496005 \\ 0.009199 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 274 & 2450 \\ 1 & 180 & 3254 \\ \dots & \ddots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ 1 & 370 & 2605 \end{bmatrix} \quad y = \begin{bmatrix} 162 \\ 120 \\ \dots \\ \dots \\ \dots \\ 212 \end{bmatrix}$$

따라서 추정식은 다음이 된다.

$$\hat{y} = 3.452613 + 0.496005x_1 + 0.009199x_2$$



회귀분석의 의미

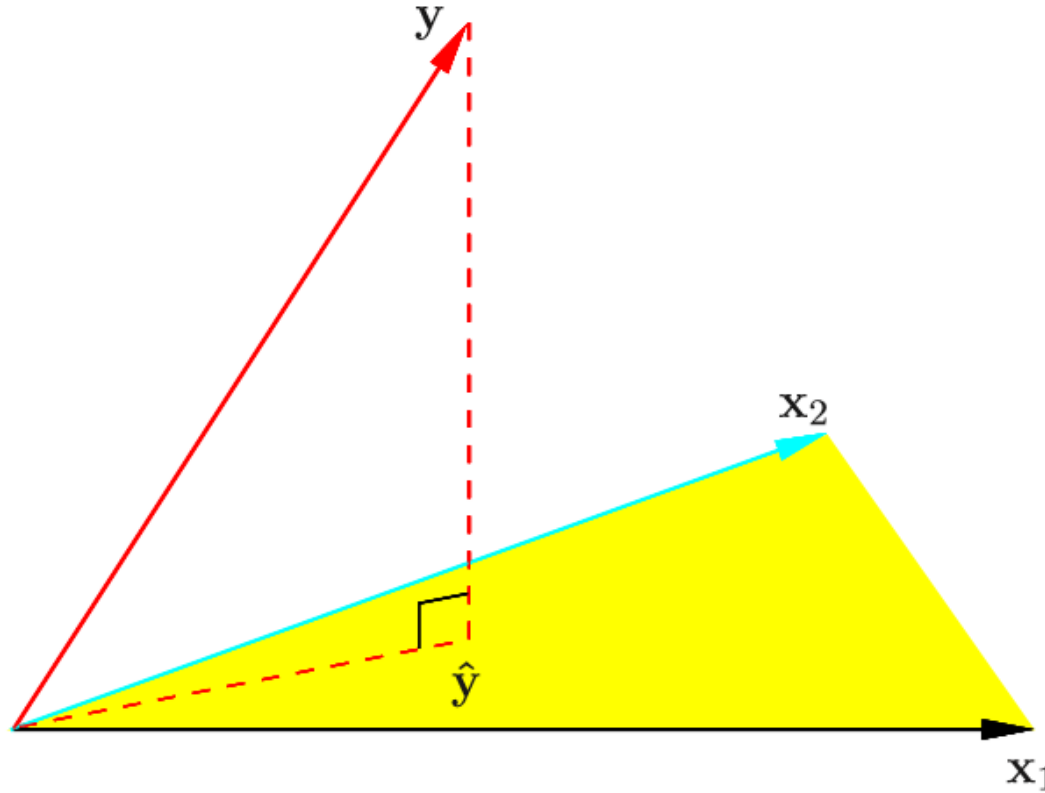


FIGURE 3.2. The N -dimensional geometry of least squares regression with two predictors. The outcome vector \mathbf{y} is orthogonally projected onto the hyperplane spanned by the input vectors \mathbf{x}_1 and \mathbf{x}_2 . The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions



곡선회귀분석

곡선회귀분석

- ✓ 독립변수가 1개일 경우 x, y 의 관계가 직선보다 곡선이 더 적절하다고 판단될 경우 $k(\geq 2)$ 차 곡선회귀 모형(curvilinear regression model)을 적합시키는 것이 바람직하다.

$$y = a + b_1x + b_2x^2 + \cdots + b_kx^k + e$$

여기서 $e_i \sim N(0, \sigma^2)$ 이고 서로 독립

- ✓ 곡선회귀 분석을 다음과 같이 중회귀 모형으로 바꾸면 n 개의 측정값에 대해서 식 (7.9)는 다음과 같이 변환되는 데, 이는 앞에서 설명한 중회귀모형과 동일하다.

$$y_i = a + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki} + e_i$$

$e_i \sim N(0, \sigma^2)$ 이고, 서로 독립이며

$i=1, 2, \dots, n$ 이다.

$$\begin{bmatrix} x = x_{1i} \\ x^2 = x_{2i} \\ \vdots \\ x^k = x_{ki} \end{bmatrix}$$



다중공선성 문제

다중공선성의 정의

- ✓ 설명변수들간의 상관계수가 높을 경우, 회귀계수의 값이 매우 커진다.
- ✓ 특정 설명변수가 다른 변수들의 선형결합(linear combination)으로 표현되는 경우
- ✓ 회귀계수의 변동성이 커져서 통계량과 모수가 서로 반대 부호를 가질 수 있다.
- ✓ F-통계량이 크나, t-통계량들이 작으면 의심해 볼 수 있다.

분산팽창계수 : Variance Inflation Factor

- ✓ For $X_j, X_j = \sum_{i=1}^m \alpha_i X_i + \varepsilon, i \neq j :$

나머지 설명변수들로 새로운 회귀식 추정 후

- ✓ $VIF_j = \frac{1}{1-R_j^2}$ for predictor X_j , where R_j^2 is the coefficient of determinant
- ✓ $VIF_j > 10$ 이면 의심의 여지가 많음.



다중공선성 문제

다중공선성을 없애기는 어렵다

- ✓ 사용되는 독립변수들 사이의 상관관계가 어느 정도 존재하는 것이 일반적이다.
- ✓ 따라서, 다중공선성을 최소화하는 것이 낫다: Forward or Backward Selection 활용
- ✓ 예측이 목적인 경우에는 어느 정도 허용될 수 있다: Prediction, Extrapolation

다중공선성 문제의 해결 방안

- ✓ 신중한 변수 선택:
 - ✓ Step-wise regression: Forward or Backward Selection 활용
- ✓ OLS 추정치의 대안:
 - ✓ 능동회귀 (Ridge regression)
 - ✓ 주성분 회귀 (PCA)



다중회귀분석 모형의 선택

■ 모형 선정 척도(Model Selection Measures) = 적합 결핍(Lack of Fit) + 복잡도(complexity)

- ✓ 적합 결핍, 복잡도 모두 적을수록 좋으나 양자는 상충관계(Trade-off)에 있다.
- ✓ Occam's Razor: The Principle of Parsimony, MDL Principle

■ AIC: Akaike Information Criterion

$$✓ AIC = n \ln(SSE) - n \ln(n) + 2p$$

■ BIC: Bayesian Information Criterion

$$✓ BIC = n \ln(SSE) - n \ln(n) + p \ln(n)$$

■ Mallows' C_p :

$$✓ C_p = p + \frac{(MSE_p - MSE_{All})(n-p)}{MSE_{All}} = \frac{MSE_p}{MSE_{All}} - (n - 2p)$$



서브셋 선택

■ 최소제곱법의 결정적인 약점 두 가지

- ✓ 예측 정확도: 최소제곱법은 종종 편향(bias)은 낮지만 분산(variance)을 높게 추정하는 경우가 있다.
- ✓ 설명력: 가장 좋은 효과를 보이는 매우 작은 서브셋은 설명력을 저하시킨다

■ 최적 서브셋 회귀 (Best subset regression)

- ✓ RSS 값이 가장 낮은 예측 모형의 최적 서브셋을 구하는 과정이다.

■ 순방향, 역방향, 혼합 단계적 선택

- ✓ 순방향 단계적 선택 (Forward stepwise selection: 공헌도가 높은 변수 추가)
- ✓ 역방향 단계적 선택 (Backward stepwise selection: 공헌도가 낮은 변수 삭제)
- ✓ 혼합 단계적 선택 (Hybrid stepwise selection: 추가 또는 삭제)
- ✓ AIC, F-test 등에 기반하여 결정한다.
- ✓ c.f. Forward stagewise regression: 잔차 상관관계가 가장 높은 예측 모형을 선택한다.



회귀분석에서의 변수선택

Extra Sum of Squares

- ✓ 회귀분석 모형에 변수가 추가된 경우 SSR의 증가분(SSE의 감소분)을 의미

$$SSR(X_1 | X_2) = SSR(X_1, X_2) - SSR(X_2) = SSE(X_2) - SSE(X_1, X_2)$$

$$\begin{aligned} SSR(X_1 | X_2, X_3) &= SSR(X_1, X_2, X_3) - SSR(X_2, X_3) \\ &= SSE(X_2, X_3) - SSE(X_1, X_2, X_3) \end{aligned}$$

$$SSR(X_1, X_2 | X_3) = SSR(X_1, X_2, X_3) - SSR(X_3) = SSE(X_3) - SSE(X_1, X_2, X_3)$$

$$\begin{aligned} SST &= SSE(X_1, X_2, \dots, X_p) + SSR(X_1, X_2, \dots, X_p) \\ &= SSE(X_1, X_2, \dots, X_p) + SSR(X_1) + SSR(X_2 | X_1) \\ &\quad + SSR(X_3 | X_1, X_2) + \dots + SSR(X_p | X_1, X_2, \dots, X_{p-1}) \end{aligned}$$



회귀분석에서의 변수선택

Extra Sum of Squares

분산분석에서 SSR 분해

변동	SS	df	MS
회귀	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3) = SSR(X_1, X_2, X_3)/3$
	$SSR(X_1 \mu)$	1	$MSR(X_1 \mu) = SSR(X_1 \mu)$
	$SSR(X_2 X_1)$	1	$MSR(X_2 X_1) = SSR(X_2 X_1)$
	$SSR(X_3 X_1, X_2)$	1	$MSR(X_3 X_1, X_2) = SSR(X_3 X_1, X_2)$
오차	$SSE(X_1, X_2, X_3)$	$n - 4$	$MSE(X_1, X_2, X_3) = SSE(X_1, X_2, X_3)/(n - 4)$
수정 총변동	SST	$n - 1$	

모형 간의 통계적 검정

$$\frac{(SSE_R - SSE_F) / m}{SSE_F / (n - p - 1)} = \frac{(SSE_R - SSE_F) / m}{MSE_F} \sim F(m, n - p - 1)$$



회귀분석에서의 변수선택

Extra Sum of Squares

✓ 통계적 검정 예시

▣ 귀무가설 $\beta_k = 0$ (설명 변수 X_k 는 유의하지 않다)에 대한 검정

Full model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + e_i$

Reduced model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + (\beta_k x_{ik} \text{ drop}) + \dots + \beta_p x_{ip} + e_i$

줄어든 모수의 개수 $m=1$

검정통계량 : $T = \frac{(SSE_R - SSE_F) / 1}{SSE_F / (n - p - 1)} \sim F(1, n - p - 1)$



회귀분석에서의 변수선택

Extra Sum of Squares

✓ 통계적 검정 예시

▣ 귀무가설 $\beta_k = \beta_m = 0$ (설명 변수 X_k 는 유의하지 않다)에 대한 검정

Full model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + e_i$

Reduced model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + (\beta_k x_{ik} \text{ drop}) + (\beta_m x_{im} \text{ drop}) + \dots + \beta_p x_{ip} + e_i$

줄어든 모수의 개수 $m=2$

검정통계량 : $T = \frac{(SSE_R - SSE_F) / 2}{SSE_F / (n - p - 1)} \sim F(2, n - p - 1)$



회귀분석에서의 변수선택

Extra Sum of Squares

✓ 통계적 검정 예시

■ 귀무가설 $\beta_k = 0.5$ (설명 변수 X_k 는 유의하지 않다)에 대한 검정

Full model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + e_i$

Reduced model: $(y_i - 0.5x_k) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + (\beta_k x_{ik} \text{ drop}) + \dots + \beta_p x_{ip} + e_i$

줄어든 모수의 개수 $m = 1$

검정통계량 : $T = \frac{(SSE_R - SSE_F) / 1}{SSE_F / (n - p - 1)} \sim F(1, n - p - 1)$



회귀분석에서의 변수선택

Backward Elimination

- ✓ 모든 독립변수를 이용한 후, 유의하지 않은 독립변수를 순차적으로 제외
- ✓ 다음의 F 값이 유의하지 않으며 가장 작은 독립변수를 제외

$$F = \frac{MSR(X_k | X_1, X_2, X_{k-1}, X_{k+1}, \dots, X_p)}{MSE(X_1, X_2, \dots, X_p)} = \frac{SSR(X_1, X_2, \dots, X_p) - SSR(X_1, X_2, \dots, X_{k-1}, X_{k+1}, \dots, X_p)}{MSE(X_1, X_2, \dots, X_p)}$$

- ✓ 모든 독립변수가 유의할 때까지 반복

Forward Selection

- ✓ 설명력이 높은 독립변수부터 순차적으로 모형에 추가
- ✓ 다음의 F 값이 유의하며 가장 큰 독립변수를 추가

$$F = \frac{MSR(X_l | X_1, \dots, X_k, X_l)}{MSE(X_1, X_2, \dots, X_k, X_l)} = \frac{SSR(X_1, \dots, X_k, X_l) - SSR(X_1, \dots, X_k)}{MSE(X_1, X_2, \dots, X_k, X_l)}$$

- ✓ 유의한 독립변수가 없을 때까지 반복

Stepwise Selection

- ✓ Forward Selection과 유사하지만 변수 선택 이후, 모형에 포함된 나머지 변수에 대하여 유의성 검정을 하여 유의하지 않은 변수를 제거



Best Subset Selection

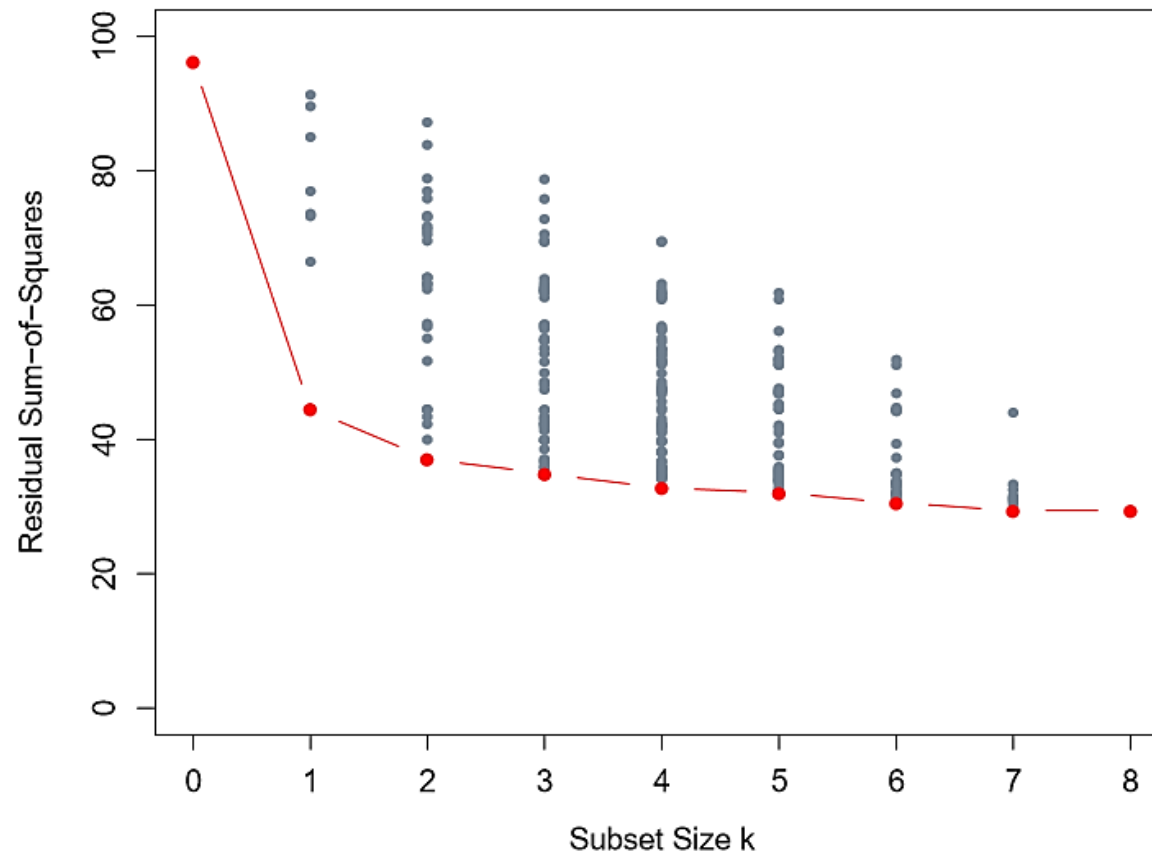


FIGURE 3.5. All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.



Other Subset Selection Methods

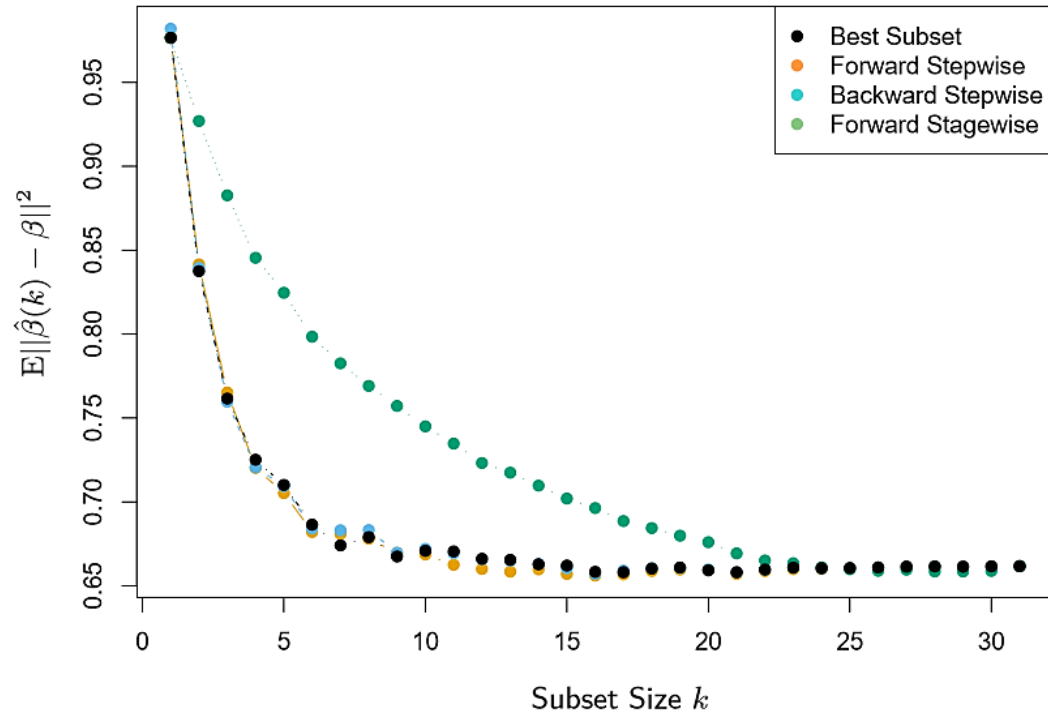


FIGURE 3.6. Comparison of four subset-selection techniques on a simulated linear regression problem $Y = X^T \beta + \varepsilon$. There are $N = 300$ observations on $p = 31$ standard Gaussian variables, with pairwise correlations all equal to 0.85. For 10 of the variables, the coefficients are drawn at random from a $N(0, 0.4)$ distribution; the rest are zero. The noise $\varepsilon \sim N(0, 6.25)$, resulting in a signal-to-noise ratio of 0.64. Results are averaged over 50 simulations. Shown is the mean-squared error of the estimated coefficient $\hat{\beta}(k)$ at each step from the true β .



회귀 모형 만들기 - 단순 회귀모형

```
from sklearn.linear_model import LinearRegression
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

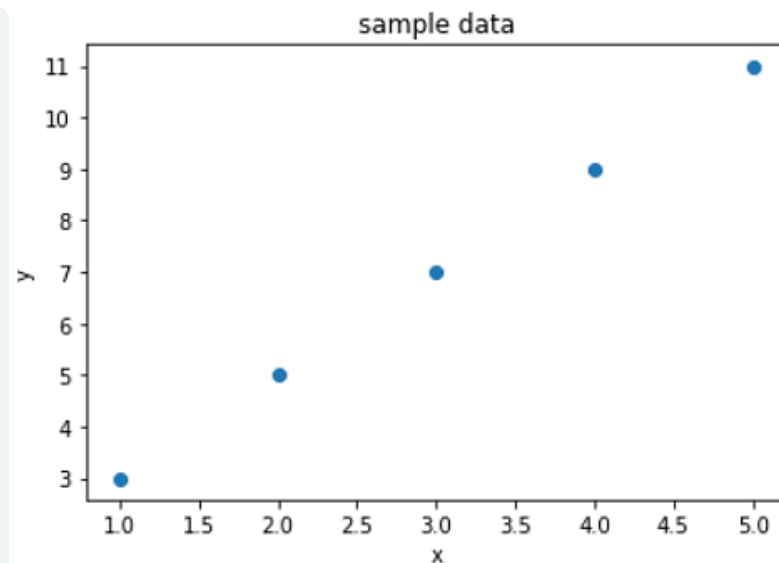
기본적인 모듈 import

```
reg=LinearRegression()
```

reg라는 변수명에 선형회귀를 할당

```
Xsample=[[1],[2],[3],[4],[5]]
Ysample=[[3],[5],[7],[9],[11]]
plt.title('단순 회귀 샘플 데이터')
plt.xlabel('x')
plt.ylabel('y')
plt.plot(Xsample,Ysample)
```

임의의 샘플 데이터 생성





회귀 모형 만들기 - 단순 회귀모형 결과 확인

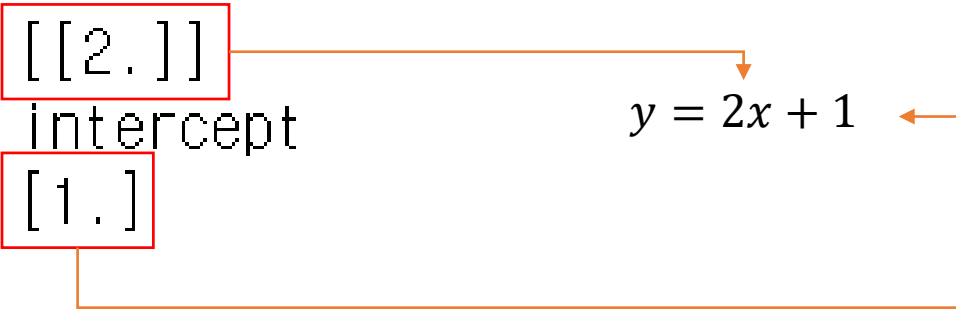
```
Model = reg.fit(Xsample, Ysample)
print("coef")
print(Model.coef_)
print("intercept")
print(Model.intercept_)
```

coef

[[2.]]

intercept

[1.]

$$y = 2x + 1$$




회귀 모형 만들기 - 단순 회귀모형 prediction

```
Model = reg.fit(Xsample, Ysample)
print("coef")
print(Model.coef_)
print("intercept")
print(Model.intercept_)
```

coef

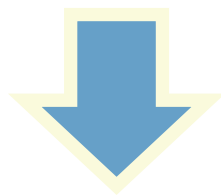
[[2.]]

intercept

[1.]

$$y = 2x + 1$$

X가 15라면? Y=31



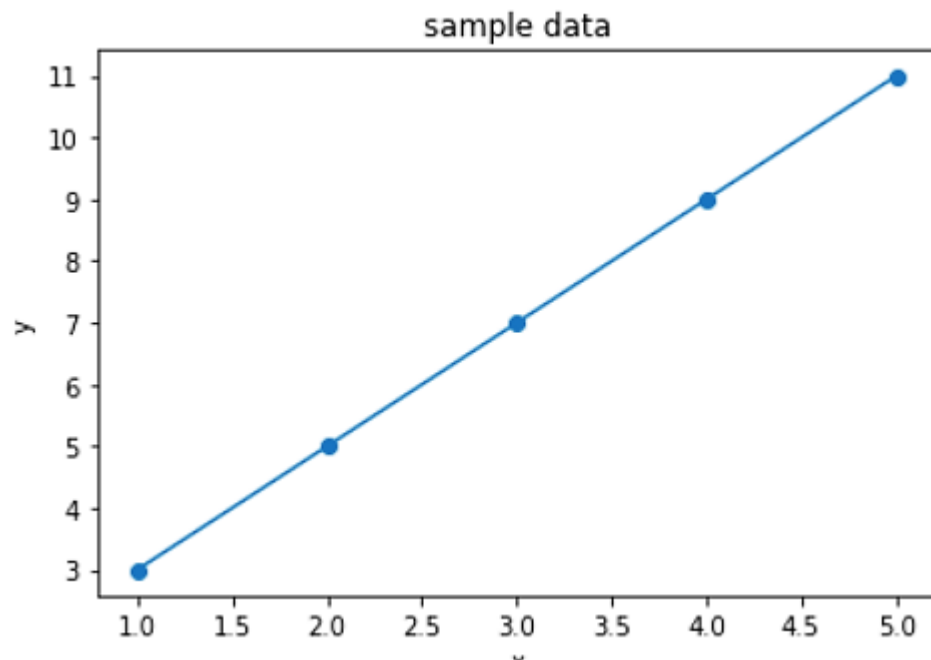
```
Model.predict([[15]])
```

```
array([[31.]])
```



회귀 모형 만들기 - 단순 회귀모형 visualiazation

```
plt.title('sample data')  
plt.xlabel('x')  
plt.ylabel('y')  
plt.scatter(Xsample, Ysample)  
plt.plot(Xsample, Model.coef_*Xsample + Model.intercept_)
```





회귀 모형 만들기 - 데이터 로드

Scikit-learn에서 제공하는 예제 데이터(fetch_california_housing dataset) load

```
from sklearn.datasets import fetch_california_housing  
  
california = fetch_california_housing()  
X=california.data  
DF=pd.DataFrame(X,columns=california.feature_names)  
Y=california.target  
print(DF)
```




회귀 모형 만들기 - 데이터 설명

Scikit-learn에서 제공하는 예제 데이터(fetch_california_housing dataset)

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85
...
20635	1.5603	25.0	5.045455	1.133333	845.0	2.560606	39.48
20636	2.5568	18.0	6.114035	1.315789	356.0	3.122807	39.49
20637	1.7000	17.0	5.205543	1.120092	1007.0	2.325635	39.43
20638	1.8672	18.0	5.329513	1.171920	741.0	2.123209	39.43
20639	2.3886	16.0	5.254717	1.162264	1387.0	2.616981	39.37

	Longitude
0	-122.23
1	-122.22
2	-122.24
3	-122.25
4	-122.25
...	...
20635	-121.09
20636	-121.21
20637	-121.22
20638	-121.32
20639	-121.24

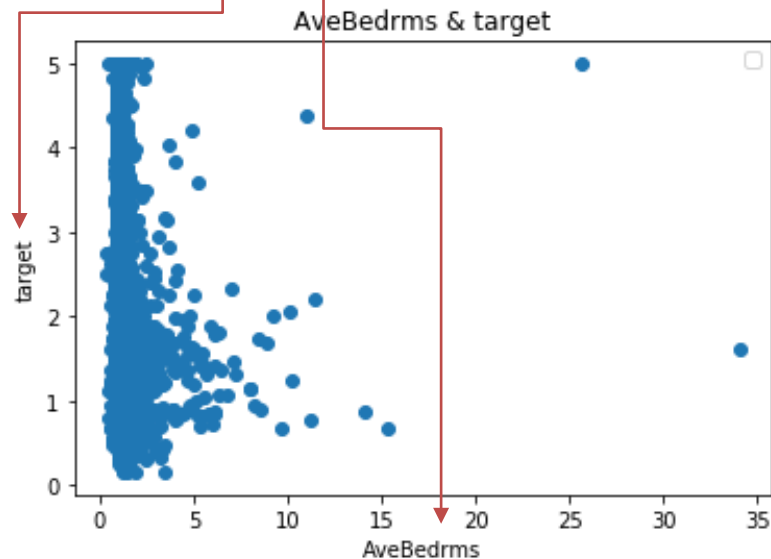
- 타겟 데이터
 - 1990년 캘리포니아의 각 행정 구역 내 주택 가격
- 특징 데이터
 - MedInc : 행정 구역 내 소득의 중앙값
 - HouseAge : 행정 구역 내 주택 연식의 중앙값
 - AveRooms : 평균 방 개수
 - AveBedrooms : 평균 침실 개수
 - Population : 행정 구역 내 인구 수
 - AveOccup : 평균 자가 비율
 - Latitude : 해당 행정 구역의 위도
 - Longitude : 해당 행정 구역의 경도



회귀 모형 만들기 - 데이터 시각화 예시

```
#i번째 feature와 타겟 값 사이의 관계 시각화
i=3
plt.title(california.feature_names[i]+' & '+ 'target')
plt.xlabel(california.feature_names[i])
plt.ylabel('target')
plt.scatter(DF[california.feature_names[i]],Y)
plt.legend()
plt.show()
```

No handles with labels found to put in legend.



- 특징 데이터
 - MedInc : 행정 구역 내 소득의 중앙값
 - HouseAge : 행정 구역 내 주택 연식의 중앙값
 - AveRooms : 평균 방 개수
 - **AveBedrooms : 평균 침실 개수**
 - Population : 행정 구역 내 인구 수
 - AveOccup : 평균 자가 비율
 - Latitude : 해당 행정 구역의 위도
 - Longitude : 해당 행정 구역의 경도



회귀 모형 만들기 - 학습 및 결과 확인

```
Model=reg.fit(X,Y)
print("coef")
print(Model.coef_)
print("intercept")
print(Model.intercept_)
```

coef

[4.36693293e-01 9.43577803e-03 -1.07322041e-01 6.45065694e-01
-3.97638942e-06 -3.78654265e-03 -4.21314378e-01 -4.34513755e-01]

회귀 계수

intercept

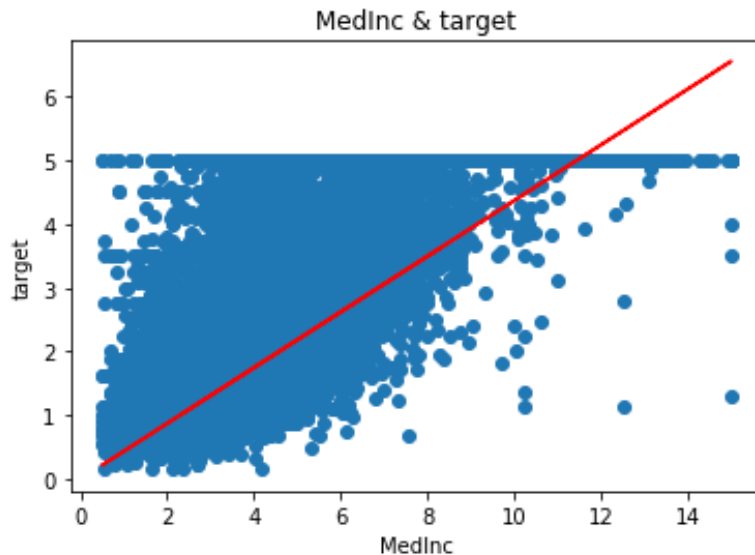
-36.94192020718434

Bias



회귀 모형 만들기 - 학습 결과 시각화

```
i=0  
plt.title(california.feature_names[i]+' & '+'target')  
plt.xlabel(california.feature_names[i])  
plt.ylabel('target')  
plt.scatter(DF[california.feature_names[i]],Y)  
plt.plot(DF[california.feature_names[i]], Model.coef_[i]*DF[california.feature_names[i]], 'r-')  
plt.show()
```



데이터 포인트 ●

모델에 의해 학습된 추세선 —



회귀 모형 만들기 - 학습 결과를 이용한 prediction

DF.mean() = 각 변수의 평균 값을 나타냄

```
DF.mean()
```

MedInc	3.870671
HouseAge	28.639486
AveRooms	5.429000
AveBedrms	1.096675
Population	1425.476744
AveOccup	3.070655
Latitude	35.631861
Longitude	-119.569704

Predict([DF.mean()]) =

각 변수의 평균 값을 토대로 예측했을 때의 결과 값

```
Model.predict([DF.mean()])  
array([2.06855817])
```