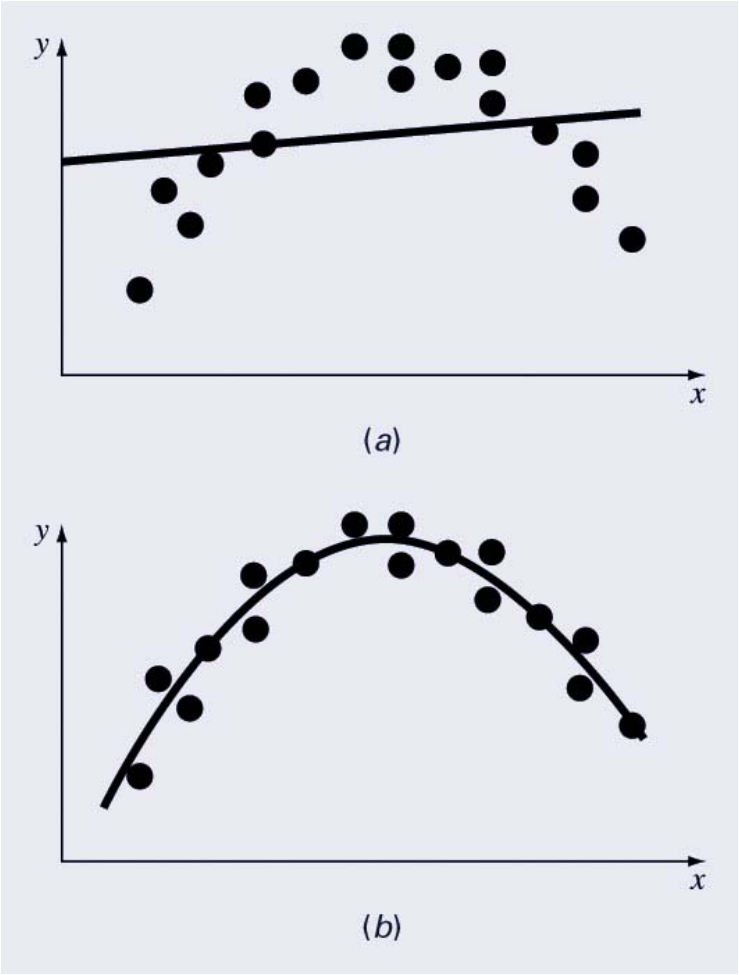
15장 일반적인 선형최소제곱과 비선형회귀분석

15.1 다항식 회귀분석
15.2 다중 선형회귀분석
15.3 일반적인 선형최소제곱
15.4 QR 분해법과 역슬래시 연산자
15.5 비선형회귀분석

15.1 다항식 회귀분석 (1/4)

앞 장에서 최소제곱 기준을 사용하여 직선 식을 유도하였으나, 일부 데이터들은 직선으로 표현하기에는 불충분하며, 이러한 경우 곡선을 이용하여 데이터를 표현하는 것이 낫다.

다항식 회귀분석 : 최소제곱 방법을 확장하여 고차다항식을 데이터에 접합



15.1 다항식 회귀분석 (3/4)

위 식들을 0으로 놓고 다시 정리하면, 다음과 같은 정규 방정식을 구할 수 있다.

$$
\begin{aligned}
(n)a_0 + \left(\sum x_i\right)a_1 + \left(\sum x_i^2\right)a_2 &= \sum y_i \\
\left(\sum x_i\right)a_0 + \left(\sum x_i^2\right)a_1 + \left(\sum x_i^3\right)a_2 &= \sum x_i y_i \\
\left(\sum x_i^2\right)a_0 + \left(\sum x_i^3\right)a_1 + \left(\sum x_i^4\right)a_2 &= \sum x_i^2 y_i
\end{aligned}
$$

미지수 3
방정식 3
의 선형시스템

계수값들은 측정값들로부터 바로 계산할 수 있다.

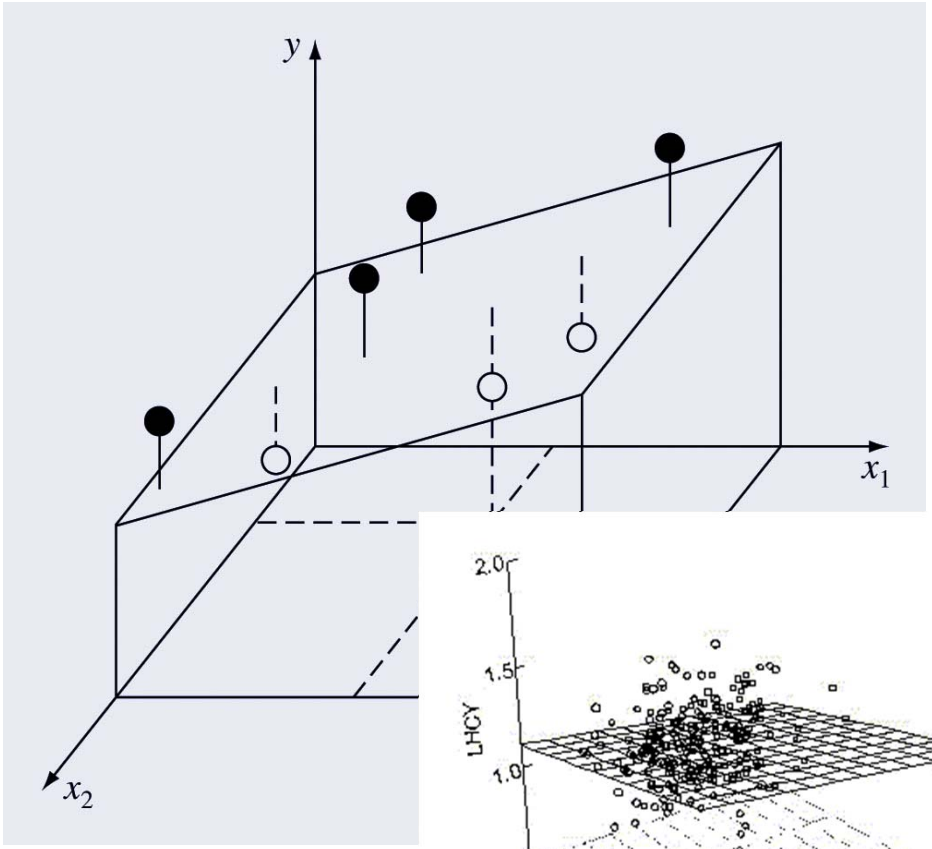
따라서 최소제곱 2차 다항식을 결정하는 문제는 세 개의 선형 연립방정식을 푸는 문제와 같다.

15.2 다중 선형회귀분석 (1/3)

y가 두 개 이상의 독립변수(x_1, x_2, …)에 대해 선형함수인 경우를 보자.

– *회귀분석 평면*

$$y = a_0 + a_1 x_1 + a_2 x_2 + e$$



앞의 경우: 차원은 1이고 독립변수의 order가 높아짐
여기서는: 차원 자체가 증가. order는 1.

15.2 다중 선형회귀분석 (2/3)

계수에 대해 미분하면,

$$\frac{\partial S_r}{\partial a_0} = -2\sum (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i})$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum x_{1,i} (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i})$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum x_{2,i} (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i})$$

잔차의 제곱합을 최소로 하는 계수는 각 미분값을 0으로 놓을 때 얻어지며, 이를 행렬식으로 표현하면

$$\begin{pmatrix} n & \sum x_{1,i} & \sum x_{2,i} \\ \sum x_{1,i} & \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{2,i} & \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{pmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_{1,i} y_i \\ \sum x_{2,i} y_i \end{Bmatrix}$$

15.2 다중 선형회귀분석 (3/3)

m차원 문제로의 확장은 다음 식을 사용한다.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m + e$$

미지수가 m+1개이니

이 경우 표준오차와 결정계수는 각각 다음과 같다.

$$s_{y/x} = \sqrt{\frac{S_r}{n-(m+1)}} \qquad r^2 = \frac{S_t - S_r}{S_t}$$

자유도는 m+1을 뺌

15.3 일반적인 선형최소제곱 (1/3)

$a_0 + a_1 x + a_2 x^2 + \ldots$ $a_0 + a_1 x_1 + a_2 x_2 + \ldots$

단순선형회귀분석, 다항식 회귀분석, 그리고 다중선형 회귀분석은 모두 *일반적인 선형최소제곱 모델*에 속한다.

15.3 일반적인 선형최소제곱 (1/3)

앞의 수식을 행렬식으로 표현해 보자:

(Ex.) 다항식 회귀분석: 측정 점 4개, 2차 다항식

$$y = a_0 z0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

$$y = a_0 + a_1 x + a_2 x^2 \rightarrow z0 = 1, z_1 = x, z_2 = x^2$$

$$\left.\begin{aligned}
(x_1, y_1) &\rightarrow y_1 = a_0 + a_1 x_1 + a_2 x_1^2 \\
(x_2, y_2) &\rightarrow y_2 = a_0 + a_1 x_2 + a_2 x_2^2 \\
(x_3, y_3) &\rightarrow y_3 = a_0 + a_1 x_3 + a_2 x_3^2 \\
(x_4, y_4) &\rightarrow y_4 = a_0 + a_1 x_4 + a_2 x_4^2
\end{aligned}\right\} \rightarrow \mathbf{y} = \mathbf{Za}$$

일반적으로
방정식의 수 > 미지수의 수
4 3
(과결정 시스템)
over-constrained

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} z_{01} & z_{11} & z_{21} \\ z_{02} & z_{12} & z_{22} \\ z_{03} & z_{13} & z_{23} \\ z_{04} & z_{14} & z_{24} \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$$

15.3 일반적인 선형최소제곱 (2/3)

일반적인 선형최소제곱 방정식은 다음과 같은 행렬식으로 쓸 수 있다.

$$\{y\} = [Z]\{a\} + \{e\}$$

여기서 y = 종속변수의 측정값, a = 미지 계수, e = 잔차, 그리고 z는,

$$[Z] = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{m1} \\ z_{02} & z_{12} & \cdots & z_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{0n} & z_{1n} & \cdots & z_{mn} \end{bmatrix}$$

행의 수: 측정 점의 수
열의 수: 파라메터의 수

점의 수 >= 파라메터 수

z_{ji} 는 i 점에서 계산된 j번째 기저 함수이다.

잔차의 제곱합 :

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} a_j z_{ji} \right)^2$$

15.3 일반적인 선형최소제곱 (3/3)

각 계수에 대하여 미분을 취하고, 그 결과를 0으로 놓으면, 다음과 같은 정규방정식(normal eq.)을 얻을 수 있다.

$$\left[[Z]^T[Z]\right]\{a\} = \left\{[Z]^T\{y\}\right\}$$

결정계수 :

$$r^2 = \frac{S_t - S_r}{S_t} = 1 - \frac{S_r}{S_t} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

$\hat{y}$ = 최소제곱접합의 예측값

최적접합 곡선과 데이터 사이의 잔차 :

$$\{y\} - [Z]\{a\}$$

Solve Ex) 15.3

15.4 QR 분해법과 역슬래시 연산자

네 점을 지나는 포물선: $f(x) = a_1 x^2 + a_2 x + a_3$

(Ex.) (2,1.12), (3,0.616), (4, 0.525), (5, 0.457)을 지나는 포물선을 구하시오.

$$\left.\begin{aligned} 1.120 &= a_1(2)^2 + a_2(2) + a_3 \\ 0.616 &= a_1(3)^2 + a_2(3) + a_3 \\ 0.525 &= a_1(4)^2 + a_2(4) + a_3 \\ 0.457 &= a_1(5)^2 + a_2(5) + a_3 \end{aligned}\right\} \rightarrow \begin{bmatrix} 4 & 2 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \\ 25 & 5 & 1 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} 1.120 \\ 0.616 \\ 0.525 \\ 0.457 \end{Bmatrix}$$

Cond(Z) = 216.4

Z행렬은 수치적으로 불안정 (큰 조건수)
그런데, Z^TZ행렬은 이 불안정성을 더 증가시킴

➔ 부록 A의 QR 분해를 이용하면 Z^TZ 의 연산이 불필요하다.

15.4 QR 분해법과 역슬래시 연산자

정규방정식은 불량조건을 가질 수 있으므로 반올림오차[1] 에 민감할 수 있다. 이런 면에서 두 가지 고급해석 기법인 QR 분해법[2] 과 특이값 분리는 더욱 강건한 방법들이다.

다음의 경우 MATLAB에서 QR 분해법이 자동적으로 실행된다.

(1) 다항식을 접합시키기 위해 내장함수 polyfit을 실행할 때
(2) 과결정 시스템을 왼쪽 나눗셈으로 풀 때

1) http://imre.polik.net/wp-content/uploads/polik_ces703_4.pdf
http://www.cs.uleth.ca/~holzmann/notes/illconditioned.pdf

15.5 비선형 회귀분석 (1/2)

일반적인 공학과 과학 분야의 문제에서 데이터를 접합할 때, *일반적인 선형최소제곱 모델*의 형태에 맞게 조작할 수 없는 경우가 많다.

$$y = a_0(1 - e^{-a_1 x}) + e$$

: 선형화 불가능

비선형모델은 매개변수가 비선형적으로 종속된다.

비선형회귀분석도 잔차의 제곱합을 최소화시키는 매개변수를 구하는 데 근거하지만, 비선형 경우의 해는 반복적인 방법에 의해서만 구할 수 있다.

15.5 비선형 회귀분석 (2/2)

비선형회귀분석을 위한 방법 :

(a) Gauss-Newton법[1] (Chapra and Canale, 2002)
(b) 최소제곱접합을 직접 구하기 위하여 최적화기법을 사용함 :
- 잔차의 제곱합을 계산하기 위하여 위 식을 다음과 같은 함수로 표현한다.

$$f(a_0, a_1) = \sum_{i=1}^{n} \left[y_i - a_0 (1 - e^{-a_1 x_i}) \right]^2$$

- 함수를 최소화시키는 a0와 a1을 결정하기 위하여 최적화 기법을 사용한다.

```
[x, fval] =fminsearch(fun, x0, options, p1, p2, ···)
```

(Nelder-Mead Simplex method)[2]

1) http://blog.naver.com/PostView.nhn?blogId=sdland85&logNo=90108105130
2) https://en.wikipedia.org/wiki/Nelder%E2%80%93Mead_method

예제 15.5 (MATLAB을 이용한 비선형회귀분석) (1/3)

Q. 예제 14.4에서 로그를 이용한 선형화를 통하여 표 14.1의 데이터에 멱모델을 접합시켰음을 기억하라. 이 멱모델은 다음과 같다.

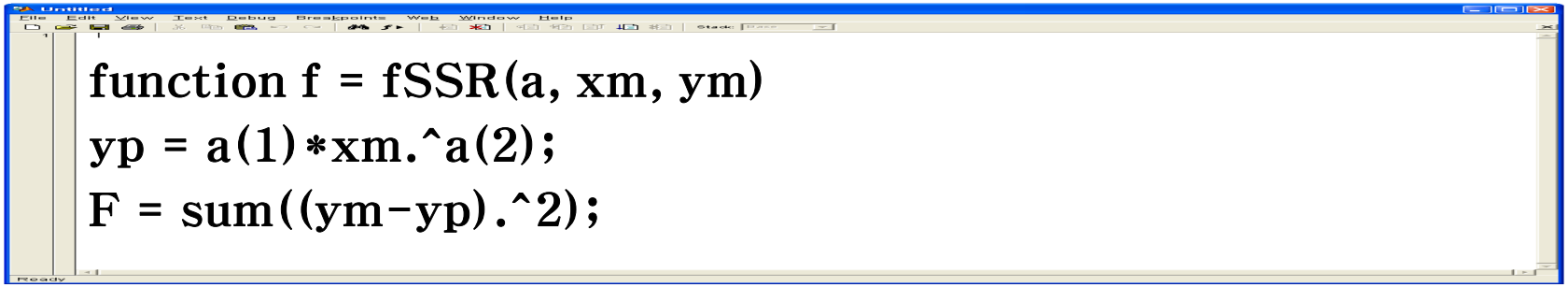
$$F = 0.2741\, v^{1.9842}$$

이 예제를 반복하되 비선형회귀분석을 사용하라. 계수에 대한 초기조건은 1을 사용한다.

예제 15.5 (MATLAB을 이용한 비선형회귀분석) (2/3)

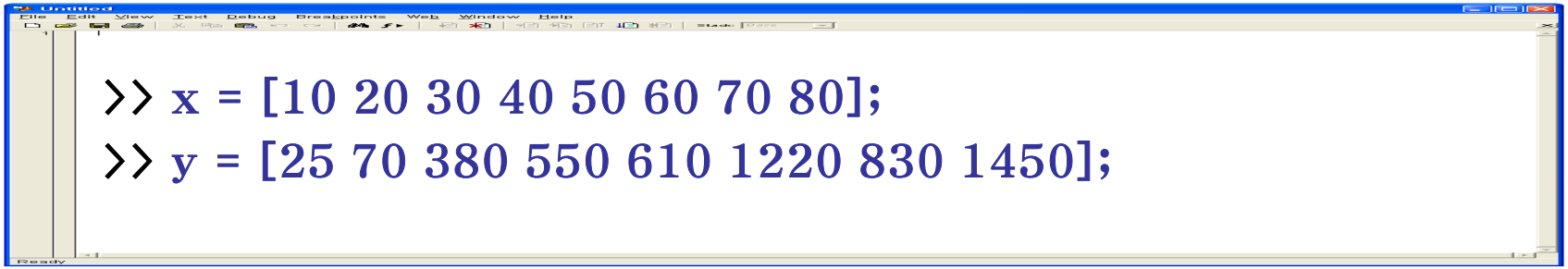
풀이)

제곱합을 계산하기 위한 M-파일 함수



```
function f = fSSR(a, xm, ym)
yp = a(1)*xm.^a(2);
F = sum((ym-yp).^2);
```

데이터 입력



```
>> x = [10 20 30 40 50 60 70 80];
>> y = [25 70 380 550 610 1220 830 1450];
```

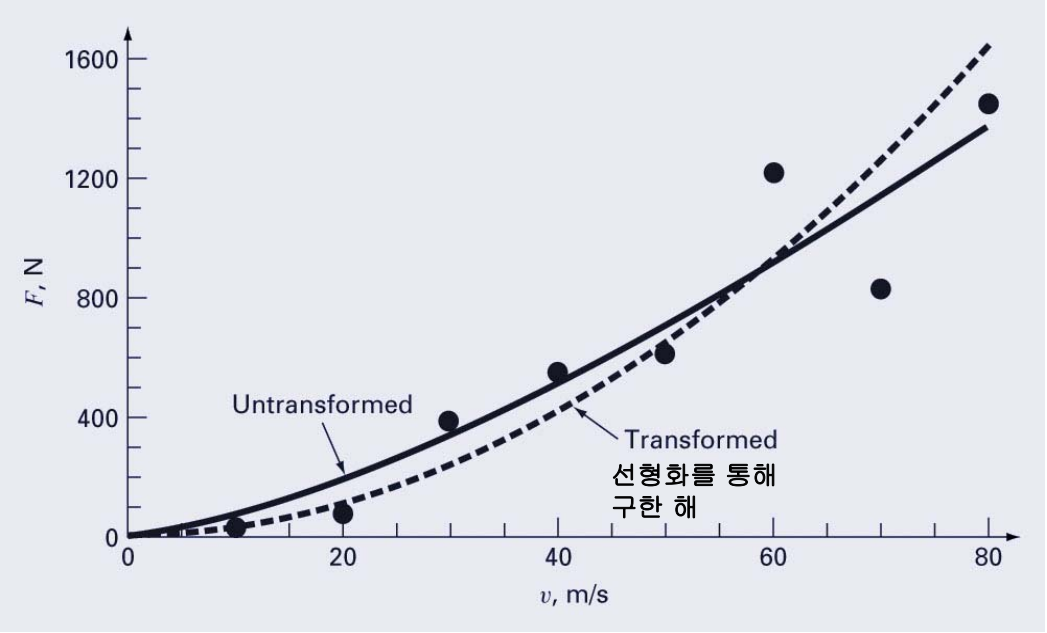
예제 15.5 (MATLAB을 이용한 비선형회귀분석) (3/3)

함수의 최소화

```
>> fminsearch(@fSSR, [1, 1], [ ], x, y)
ans =
    2.5384    1.4359
```

최적접합 모델:

$$F = 2.5384\, v^{1.4359}$$



부록 A. QR분해

QR분해

$$\mathbf{Z} = \mathbf{q} \cdot \mathbf{r} = \begin{bmatrix} \text{orthonormal} \\ \text{matrix} \end{bmatrix} \cdot \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}$$

q r

정규직교행렬 상삼각행렬

For non-symmetric matrix $(m \neq n, m > n)$

$$\mathbf{Z} = \mathbf{qr} = \mathbf{q}\begin{bmatrix} \mathbf{r}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{bmatrix}\begin{bmatrix} \mathbf{r}_1 \\ 0 \end{bmatrix} = \mathbf{q}_1\mathbf{r}_1$$

m-n개의 열

m-n개의 행

$$\mathbf{Z} \cdot \mathbf{a} = \mathbf{y} \rightarrow (\mathbf{q}_1\mathbf{r}_1) \cdot \mathbf{a} = \mathbf{y}$$

$$\mathbf{q}_1{}^T\mathbf{q}_1\mathbf{r}_1 \cdot \mathbf{a} = \mathbf{I} \cdot \mathbf{r}_1 \cdot \mathbf{a} = \mathbf{q}_1{}^T\mathbf{y}$$

$$\mathbf{r}_1 \cdot \mathbf{a} = \mathbf{q}_1{}^T\mathbf{y} \quad \text{(후진대입)}$$

(Ex.) 앞 예제의 Matlab 풀이

```
>> Z =
    4    2    1
    9    3    1
   16    4    1
   25    5    1
>> [q,r]=qr(Z)  % QR분해
q =
    0.1279    0.6602    0.7055   -0.2236
    0.2878    0.5717   -0.3746    0.6708
    0.5116    0.2043   -0.4965   -0.6708
    0.7994   -0.4422    0.3397    0.2236
r =
   31.2730    7.1627    1.7267
         0    1.6417    0.9940
         0         0    0.1742
         0         0         0
>> q1=q(:, 1:3); % q1만 잘라냄(slicing)
>> r1=r(1:3,:); % r1만 잘라냄(slicing)
>>q1*r1 % 곱해보면 Z가 나옴
 ans =
    4.0000    2.0000    1.0000
    9.0000    3.0000    1.0000
   16.0000    4.0000    1.0000
   25.0000    5.0000    1.0000
>>a =r1\(q1'*y)  % r1*a=q1'*y, 후진 대입
a =  (0.1090  -0.9710  2.6055)'  % solution
>> Z*a  % (1.12 0.616 0.525 0.457), 검산
ans = (1.100  0.6745 0.4665 0.4765)' % y와 비슷
```

m=4,
n=3의 경우

m-n개의 열

m-n개의 행