Linear Regression

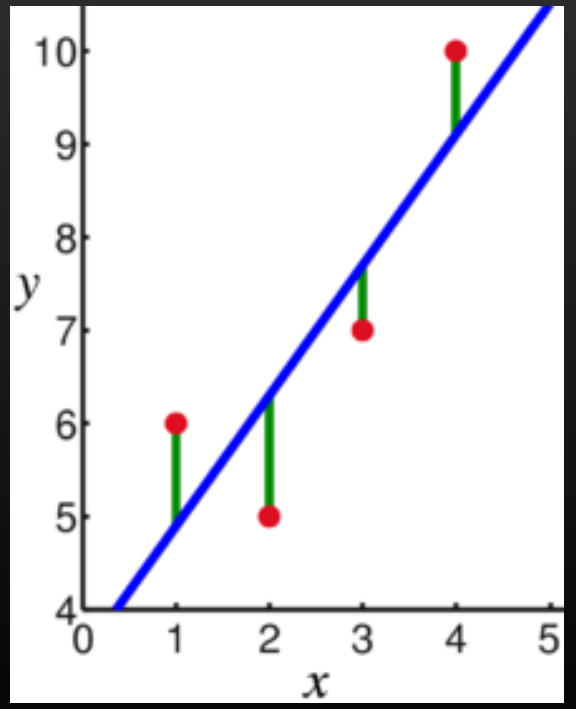
HuStar

2020.07.

이성윤

선형 회귀분석이란?

- 입력과 출력이 선형 관계를 갖고, 그 일반식을 알고 있을 때, 다수의 데이터 벡터를 이용하여 가장 적은 오차를 갖도록 모델링 인자를 추정하는 방법이다.
  - $y = a_1 x + a_0$
  - 측정치 y에 노이즈 또는 오차가 포함되어 있음
  - $y_{meas} = a_1 x_{known} + a_0 + \epsilon$
  - 선형 관계를 나타내는 a_1(기울기)과 a_0(절편)는 미지의 값
  - 다수의 x, y 쌍을 통해 a_1과 a_0를 추정하는 것



선형 회귀분석의 일차함수 예시

$$y_i = a_1 x_i + a_0 + \epsilon_i$$

$$\begin{pmatrix} y_1 = a_1 x_1 + a_0 + \epsilon_1 \\ \vdots \\ y_n = a_1 x_n + a_0 + \epsilon_n \end{pmatrix}$$

$$Y = AX + E$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \qquad A = \begin{bmatrix} a_1 \\ a_0 \end{bmatrix}, \qquad E = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- 연립방정식에서 미지수의 수보다 식의 수가 많으면 해가 없으나, 여기서 A는 정확한 답을 알 수 있는 것이 아니라 데이터 X, Y로부터 계산하는 추정치이다.

선형 회귀분석의 일차함수 예시

$$J = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon_1^2 + \cdots + \epsilon_n^2, \qquad \epsilon_i = y_i - \widehat{a_1} x_i - \widehat{a_0}$$

- J를 최소화하여 오차가 가장 작아지는 a_1과 a_0를 계산해야 하는데, 주어진 식이 ε에 대한 2차식의 형태이므로 편미분하여 0이 되는 해를 구하면 J가 최소인 두 값을 구할 수 있다.

$$\frac{\partial J}{\partial \widehat{a_1}} = \frac{\partial \sum_{i=1}^{n} \epsilon_i^2}{\partial a_1} = \sum_{i=1}^{n} \frac{\partial \epsilon_i^2}{\partial a_1} = \sum_{i=1}^{n} \frac{\partial (y_i - a_1 x_i - a_0)^2}{\partial a_1}$$

$$= \sum_{i=1}^{n} \{-2x_i(y_i - a_1 x_i - a_0)\} = \sum_{i=1}^{n} \{-2x_i y_i + 2x_i^2 a_1 + 2x_i a_0\} = 0$$

$$-2\sum_{i=1}^{n} x_i y_i + 2a_1 \sum_{i=1}^{n} x_i^2 + 2a_0 \sum_{i=1}^{n} x_i = 0$$

$$a_1 \sum_{i=1}^{n} x_i^2 + a_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$\frac{\partial J}{\partial a_0} = \frac{\partial \sum_{i=1}^{n} \epsilon_i^2}{\partial a_0} = \sum_{i=1}^{n} \frac{\partial \epsilon_i^2}{\partial a_0} = \sum_{i=1}^{n} \frac{\partial (y_i - a_1 x_i - a_0)^2}{\partial a_0}$$

$$= \sum_{i=1}^{n} \{-2(y_i - a_1 x_i - a_0)\} = \sum_{i=1}^{n} \{-2y_i + 2a_1 x_i + 2a_0\} = 0$$

$$-2\sum_{i=1}^{n} y_i + 2a_1 \sum_{i=1}^{n} x_i + 2a_0 \sum_{i=1}^{n} 1 = 0$$

$$a_1 \sum_{i=1}^{n} x_i + a_0 \sum_{i=1}^{n} 1 = \sum_{i=1}^{n} y_i$$

선형 회귀분석의 일차함수 예시

$$\begin{pmatrix} a_1\sum_{i=1}^{n}x_i^2 + a_0\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}x_iy_i \\ a_1\sum_{i=1}^{n}x_i + a_0\sum_{i=1}^{n}1 = \sum_{i=1}^{n}y_i \end{pmatrix} \rightarrow \begin{pmatrix} \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}1 \end{pmatrix} \begin{bmatrix} a_1 \\ a_0 \end{bmatrix} = \begin{pmatrix} \sum_{i=1}^{n}x_iy_i \\ \sum_{i=1}^{n}y_i \end{pmatrix}$$

$$\underbrace{\begin{pmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{pmatrix}}_{X^T} \underbrace{\begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}}_{X} = \underbrace{\begin{pmatrix} \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}1 \end{pmatrix}}_{X^TX}, \quad \underbrace{\begin{pmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{pmatrix}}_{X^T} \underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} \sum_{i=1}^{n}x_iy_i \\ \sum_{i=1}^{n}y_i \end{pmatrix}}_{X^TY}$$

$$X^TX\hat{A} = X^TY$$

$$\therefore \hat{A} = \left(X^TX\right)^{-1}X^TY$$

Example 1

- 일차함수 $y = 2x + 1$의 y에 표준편차가 1인 정규분포 난수를 추가하여 데이터를 생성한 후, 회귀분석을 통해 기울기와 절편을 구하고 그래프로 비교해보자.
  - 입력 x의 범위 0~5에서 0.5 간격으로 y를 계산하고 함수 randn을 통해 난수 추가
  - 회귀분석을 통해 기울기(a_1, 기존 값 2)와 절편(a_0, 기존 값 1)의 추정치를 계산
  - 생성한 데이터(점으로 표시), 기존 함수, 추정한 함수 세 가지를 한 그래프에 표시

선형 회귀분석의 일반식 유도

$$y_i = a_j x_{ij} + a_{j-1} x_{i(j-1)} + \cdots + a_1 x_{i1} + \epsilon_i$$

$$\begin{pmatrix} y_1 = a_m x_{1m} + \cdots + a_1 x_{11} + \epsilon_1 \\ \vdots \\ y_n = a_m x_{nm} + \cdots + a_1 x_{n1} + \epsilon_n \end{pmatrix}$$

$$Y = AX + E$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad A = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix}, \quad E = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- 벡터 미분법을 통해 이전에 구했던 추정치 $\hat{A} = \left(X^T X\right)^{-1} X^T Y$를 유도할 수 있다.

선형 회귀분석의 일반식 유도

$$J = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon_1^2 + \cdots + \epsilon_n^2 = E^T E = \left(Y - X\hat{A}\right)^T \left(Y - X\hat{A}\right)$$

$$= Y^T Y - \hat{A}^T X^T Y - Y^T X\hat{A} + \hat{A}^T X^T X\hat{A}$$

$$\frac{\partial J}{\partial \hat{A}} = -Y^T X - Y^T X + \hat{A}^T X^T X + \hat{A}^T X^T X = 0$$

$$2\hat{A}^T X^T X = 2Y^T X$$

$$X^T X\hat{A} = X^T Y$$

$$\therefore \hat{A} = \left(X^T X\right)^{-1} X^T Y$$

- 모든 선형 모델링에서 인자는 위의 식을 통해 데이터로부터 추정할 수 있다.
- 이와 같은 추정 방식을 최소제곱추정(Least Square Estimation)이라고 한다.

삼각함수의 입력에 대한 출력 추정

- LTI 시스템에서 입출력은 주파수 영역에서 모델링할 수 있고, 각 주파수에 대한 응답(크기, 위상)이 시스템을 나타낸다.
- 시스템의 사인파 입력에 대한 출력을 측정하였고, 그 값에 노이즈가 포함되어 있다고 할 때, 선형 회귀분석을 이용하여 크기와 위상을 추정할 수 있다.

$$y_i = a_2 x_{i2} + a_1 x_{i1} + a_0 + \epsilon_i$$

  - 사인파 입력이므로, x_{i2}에 입력 sin(w*t)를 대입한다.
  - 실제 입력은 아니지만, 입력과 동일한 위상을 갖도록 x_{i1}에 cos(w*t)를 대입한다.
  - 출력 yi는 sin(w*t)와 cos(w*t)의 선형 조합으로 나타난다.

$$y_i = a_2 \sin(\omega t_i) + a_1 \cos(\omega t_i) + a_0$$

- 인자 a_2, a_1을 통해 LTI 시스템의 응답 yi의 주파수 응답(크기와 위상)을 계산할 수 있으며, a_0는 오프셋이다.

$$y_i = A \sin(\omega t_i + \phi) + C = a_2 \sin(\omega t_i) + a_1 \cos(\omega t_i) + a_0$$

$$C = a_0, \qquad A = \sqrt{a_2^2 + a_1^2}, \qquad \phi = \tan^{-1} \frac{a_1}{a_2}$$

Example 2

- 전달함수 $\mathrm{H}(s) = \frac{1}{s^2+0.5s+1.2}$에 1rad/sec 사인파 입력에 대한 응답을 표준편차가 0.05인 정규분포 난수를 추가하여 데이터를 생성한 후, 회귀분석을 통해 주파수 응답의 크기와 위상을 구하고 그래프로 비교해보자.
  - 입력 sin(t)로 20초간의 전달함수 응답 y를 계산하고 함수 randn을 통해 난수 추가
  - 회귀분석을 통해 크기와 위상의 추정치를 계산
  - 입력 사인파, 출력 데이터(점으로 표시), 추정한 크기와 위상으로 그린 사인파 세 가지를 한 그래프에 표시

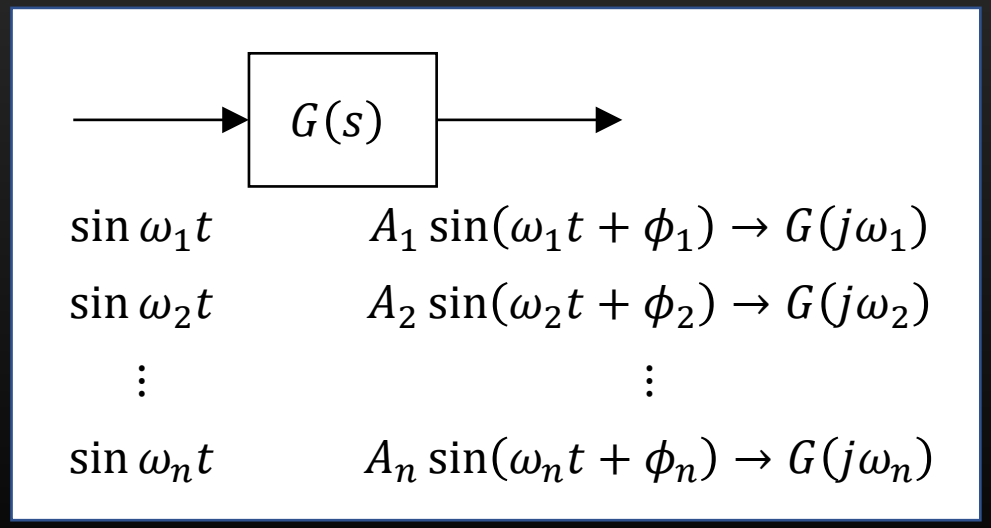
Example 3

- Example 2 전달함수의 impulse response 또는 step response를 통해 steady-state에 가는데 걸리는 시간을 알아본다.
- 그 시간을 바탕으로 필요한 시뮬레이션 시간을 정하고, 다시 회귀분석을 수행해보자.
  - 함수 step 또는 impulse로 steady-state에 도달한 시간 대략적으로 지정
  - 입력 sin(t)의 steady-state에 걸리는 시간+20초간의 전달함수 응답 y를 계산하고 함수 randn을 통해 난수 추가
  - 회귀분석을 통해 크기와 위상의 추정치를 계산
  - 입력 사인파, 출력 데이터(점으로 표시), 추정한 크기와 위상으로 그린 사인파 세 가지를 한 그래프에 표시

시스템 응답으로부터 전달함수 추정

- 전달함수의 형태(분자와 분모 차수)에 대한 정보가 있는 상태에서 여러 주파수 응답 데이터를 얻으면 전달함수의 형태를 추정할 수 있다.

$$G(s) = \frac{b_n s^n + \cdots + b_1 s + b_0}{a_m s^m + \cdots + a_1 s + a_0} \rightarrow numerator\ n-th, denominator\ m-th$$

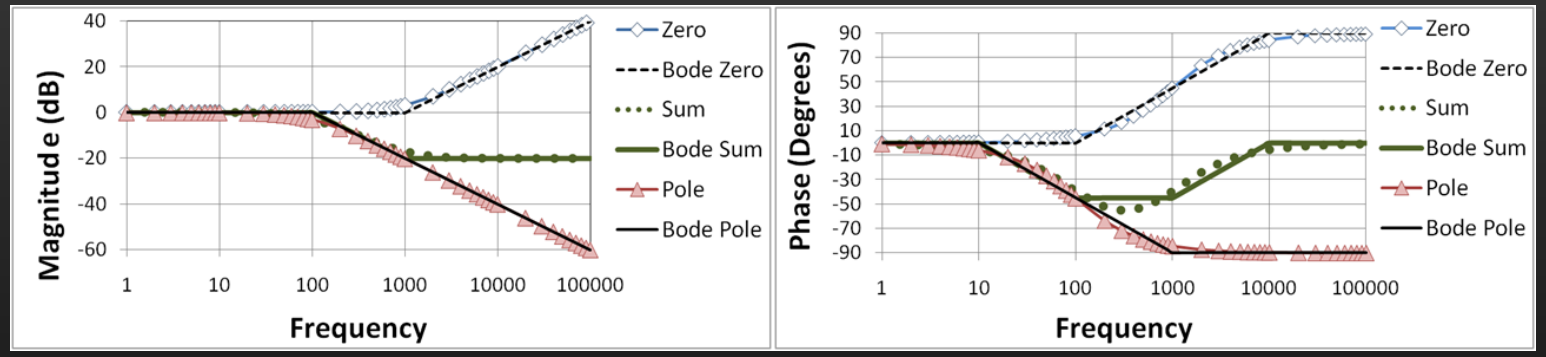


Example 4

- Example 2의 시스템 전달함수 $\mathrm{H}(s) = \frac{1}{s^2+0.5s+1.2}$에 여러 주파수의 사인파 입력에 대한 응답을 구하여 시스템 전달함수를 다시 구해 보자. (표준편차가 0.05인 정규분포 난수 포함)
  - 입력 sin(wt)에서 w를 logscale로 여러 개 선정, 각 주파수에 대한 전달함수의 응답을 구한 뒤 bode plot 형식으로 그래프로 표현
  - 회귀분석을 통해 크기와 위상의 추정치로부터 시스템 계수를 계산(함수 invfreqs)

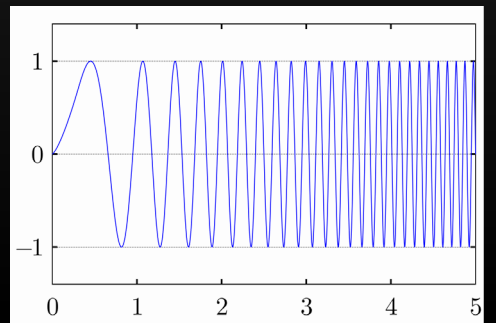
시스템 응답으로부터 전달함수 추정

- 전달함수의 형태(분자와 분모 차수)에 대한 정보와 데이터를 얻을 주파수의 범위는 알고 있는 물리적 지식을 통해 예상하는 방법과, 직접 실험을 통해 얻는 방법이 있다.



- 정확한 형태를 알지 못하더라도 비슷한 형태라면 근사하여 전달함수를 구할 수 있으며, bode plot을 최대한 넓은 입력 주파수 범위에서 그려보고 고주파 영역에서 분모, 분자 차수의 차이와 각 지점의 기울기(차수당 20dB/decade)를 통한 pole/zero의 위치 등을 파악해야 한다.
- 주파수 응답을 얻는 실험의 횟수는 비용에 직결되므로, 주파수 응답을 얻을 주파수를 미리 결정할 때 신중해야 할 수 있다. (실험당 비용이 높다면 – 시간이나 금액)
  - 적절한 속도의 frequency chirp을 입력하는 방법을 사용할 수 있다면 좋다.

$$u_{in}(t) = \sin(\omega(t) \times t)$$



Linear Regression

Example 5

- 미지의 시스템 unknown_sys1의 전달함수 형태를 예상해보자.
  - Frequency chirp을 입력하여 주파수에 대한 크기 응답을 얻고, 주파수 범위를 결정
  - 각 주파수의 입력에 대응하는 주파수 응답의 크기와 위상을 계산
  - 주파수 응답을 bode plot으로 표시
  - Unknown_sys1의 전달함수 형태(차수) 예상

Example 6

- 미지의 시스템 unknown_sys1의 전달함수를 구해보자.
  - 앞서 얻은 frequency response와 예상 차수를 통해 전달함수를 계산
  - 계산한 전달함수의 step response와 unknown_sys1의 step response를 비교