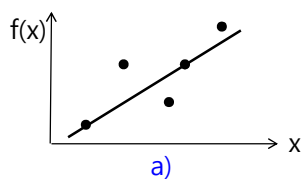


14. Linear Regression

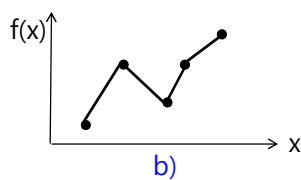
HoHee Kim

Curve Fitting : Discrete data 들로부터 값을 추정해야 할 때, 측정된 data 들을 직선 또는 곡선으로 연결하는 것



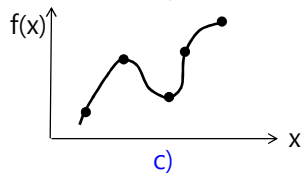
a) 오차가 큰 data 의 일반적인 경향을 직선 또는 곡선으로 나타내는 것 → regression

14장, 15장



b), c) 정확한 data 들을 모두 지나는 곡선으로 나타내는 것 → interpolation

17장, 18장



수치해석-14장

경북대 전자공학부 김호희

2

Statistics Review

어떤 실험에서 여러 번 측정하여 7개의 측정값을 얻었다 가정

| | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|
| 6.395 | 6.555 | 6.555 | 6.625 | 6.625 | 6.655 | 6.775 |
|-------|-------|-------|-------|-------|-------|-------|

- Arithmetic mean(산술평균) : n 개의 data들의 평균 $\bar{y} = \frac{\sum y_i}{n}$
- Median : data 들을 오름차순으로 나열했을 때 중앙에 있는 값
 - data 의 수가 홀수일 때 중앙에 있는 값
 - data 의 수가 짝수일 때 중앙에 있는 2개의 data 의 산술평균값
- Mode : data들 중에서 가장 빈번하게 발생하는 data 중 가장 낮은 값
- Range : 가장 큰 값과 가장 작은 값의 차이
- data 와 평균의 차이 (Residual) 의 제곱의 합 : $S_t = \sum (y_i - \bar{y})^2$

수치해석-14장

경북대 전자공학부 김호희

3

- 한 Sample 에 대한 standard deviation (표준편차) :

$$s_y = \sqrt{\frac{S_t}{n-1}}$$

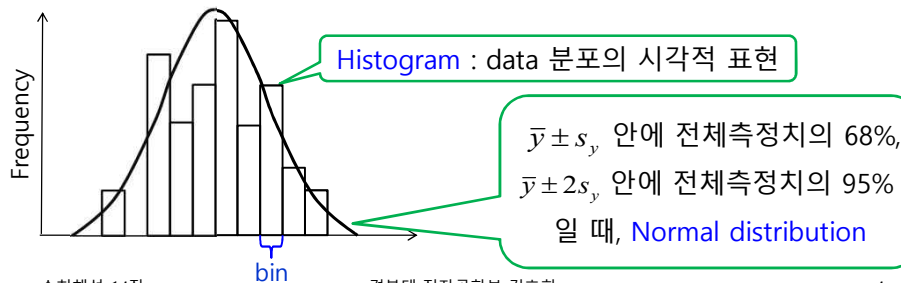
Degrees of freedom(자유도)이 n-1 이므로

- Variance (분산) :

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{\sum y_i^2 + n \left(\frac{\sum y_i}{n} \right)^2 - 2 \left(\frac{\sum y_i}{n} \right) \sum y_i}{n-1} = \frac{\sum y_i^2 - (\sum y_i)^2 / n}{n-1}$$

- Coefficient of variation (분산계수) : data 의 분포를 수치화한 통계값

$$\text{c.v.} = \frac{s_y}{\bar{y}} \times 100\%$$



수치해석-14장

경북대 전자공학부 김호희

4

```
>> s = [6.395 6.555 6.555 6.625 6.625 6.655 6.775];
```

```
>> mean(s), median(s), mode(s)
```

```
ans = 6.5979
```

```
ans = 6.6250
```

```
ans = 6.5550
```

가장 빈번한 것 중 가장 낮은 값

```
>> range = max(s) - min(s)
```

```
range =
```

```
0.38
```

variance & standard deviation

```
>> var(s), std(s)
```

```
ans = 0.0135
```

```
ans = 0.1161
```

n 은 각 bin 사이에 있는 데이터 수

x 는 각 bin 의 중간값을 의미

hist(s) 만 치면 그래프만 나옴

```
>> [n,x] = hist(s)
```

```
n = 1
```

```
0
```

```
0
```

```
0
```

```
2
```

```
0
```

```
3
```

```
0
```

```
0
```

```
1
```

```
x = 6.4140
```

```
6.4520
```

```
6.4900
```

```
6.5280
```

```
6.5660
```

```
6.6040
```

```
6.6420
```

```
6.6800
```

```
6.7180
```

```
6.7560
```

수치해석-14장

경북대 전자공학부 김호희

5

Linear Least-Squares Regression

data 점들을 시각적으로 조사한 후 최적의 선을 결정하는데,

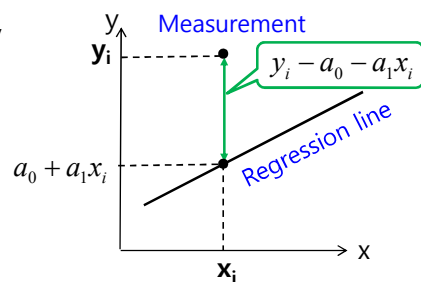
→ data 점들과 직선 사이의 차이를 최소화시키는 선으로 정함

- n 개의 점 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$,

$$y = a_0 + a_1x + e$$

관측치와 모델값 사이의 오차,
Residual

$$e = y - a_0 - a_1x$$



$$S_r = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1x_i)^2$$

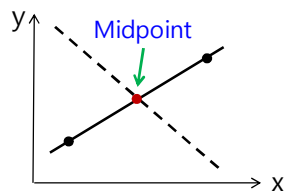
Residual 의 제곱의 합이 최소화 되는 조건을 least squares 라 함

☞ S_r 이 최소화 되도록 a_0, a_1 를 결정하여 유일한 직선을 유도

수치해석-14장

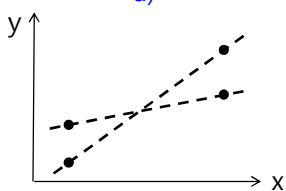
경북대 전자공학부 김호희

6



a) $\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - a_0 - a_1 x_i) : \text{residual 의 합}$

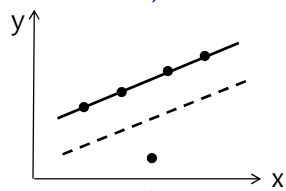
→ 중간 점을 지나는 모든 직선이 residual 합이 0



b) $\sum_{i=1}^n |e_i| = \sum_{i=1}^n |y_i - a_0 - a_1 x_i|$

: residual 의 절대치의 합

→ 점선 안의 모든 직선이 residual 절대치의 합이 최소화



c) 각 점의 최대오차를 최소화

→ 오차가 너무 큰 점 발생

☞ a), b), c) 모두 부적합

수치해석-14장
경북대 전자공학부 김호희
7

$$\left[\begin{aligned} \frac{\partial S_r}{\partial a_0} &= -2 \sum (y_i - a_0 - a_1 x_i) = 0 \\ \frac{\partial S_r}{\partial a_1} &= -2 \sum [(y_i - a_0 - a_1 x_i) x_i] = 0 \end{aligned} \right.$$

← 최소가 되려면

← 미분 값 = 0

$$\left[\begin{aligned} 0 &= \sum y_i - \sum a_0 - \sum a_1 x_i \\ 0 &= \sum x_i y_i - \sum a_0 x_i - \sum a_1 x_i^2 \end{aligned} \right.$$

a_0, a_1 의 선형 방정식으로

$$\left[\begin{aligned} na_0 + (\sum x_i) a_1 &= \sum y_i \quad \heartsuit \\ (\sum x_i) a_0 + (\sum x_i^2) a_1 &= \sum x_i y_i \end{aligned} \right.]$$

→

slope

$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

intercept

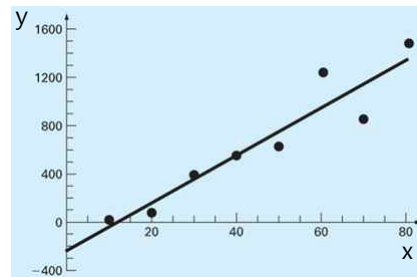
$$a_0 = \bar{y} - a_1 \bar{x} \quad \heartsuit$$

Normal equation

수치해석-14장
경북대 전자공학부 김호희
8

Ex) Fit a straight line to the values in the following table.

| i | x_i | y_i | x_i^2 | $x_i y_i$ |
|----------|-------|-------|---------|-----------|
| 1 | 10 | 25 | 100 | 250 |
| 2 | 20 | 70 | 400 | 1400 |
| 3 | 30 | 380 | 900 | 11400 |
| 4 | 40 | 550 | 1600 | 22000 |
| 5 | 50 | 610 | 2500 | 30500 |
| 6 | 60 | 1220 | 3600 | 73200 |
| 7 | 70 | 830 | 4900 | 58100 |
| 8 | 80 | 1450 | 6400 | 116000 |
| Σ | 360 | 5135 | 20400 | 312850 |



linear regression

$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{8(312850) - 360(5135)}{8(20400) - (360)^2} = 19.47024$$

$$a_0 = \bar{y} - a_1 \bar{x} = \left(\frac{5135}{8}\right) - 19.47024 \left(\frac{360}{8}\right) = -234.2857$$

$$\Rightarrow y = -234.2857 + 19.47024x$$

수치해석-14장

경북대 전자공학부 김호희

9

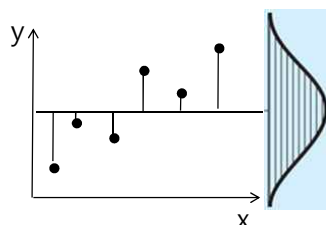
Quantification of Error of Linear Regression :

Linear regression 의 오차를 수치로 나타내는 것

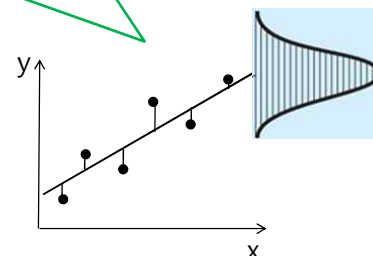
- data와 직선과의 차이는 data 범위에 걸쳐 비슷한 크기 갖고, 직선을 중심으로 한 data들의 분포가 정규분포를 가진다면 Linear Regression 은 최적의 직선 나타냄 → maximum likelihood principle

$$S_t = \sum (y_i - \bar{y})^2$$

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



regression 전



regression 후

수치해석-14장

경북대 전자공학부 김호희

10

- Standard error of the estimate:

$$s_{y/x} = \sqrt{\frac{S_r}{n-2}}$$

a_0, a_1 을 유도하려면 최소한
2개의 data는 있어야 하므로
자유도가 $n-2$

(y/x 는 특정 x 에 대한 y 의 예상 값에 대한 오차를 의미)

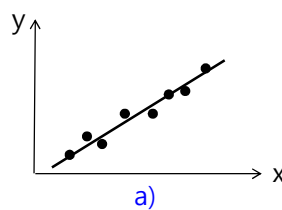
- Coefficient of determination(결정 계수), r^2 :

$$r^2 = \frac{S_t - S_r}{S_t}$$

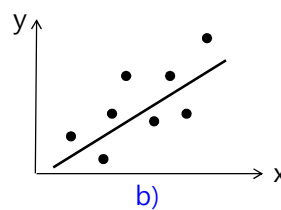
$S_r=0, r^2=1$: 완벽한 fitting

$S_t=S_r, r^2=0$: regression 해도 개선되지 않음

(r : correlation coefficient (상관계수))



수치해석-14장



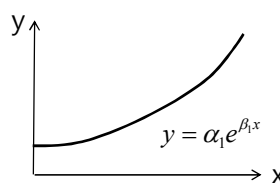
경북대 전자공학부 김호희

Residual error 가
a) small
b) large
경우

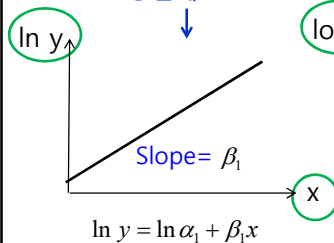
11

Linearization of Nonlinear Relationships

linear regression 은 독립변수(x)와 종속변수(y)가 선형관계일 때 적합,
비선형관계일 때도 선형으로 변환하여 a_0, a_1 를 구한 뒤 α, β 를 구함

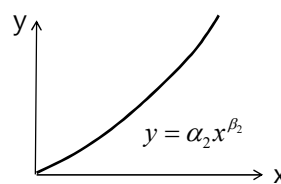


양변에 \ln

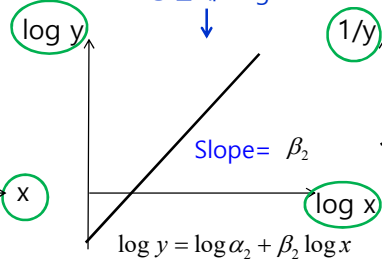


Exponential

수치해석-14장

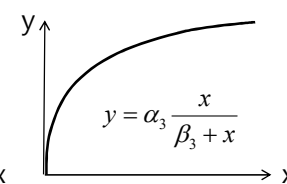


양변에 \log

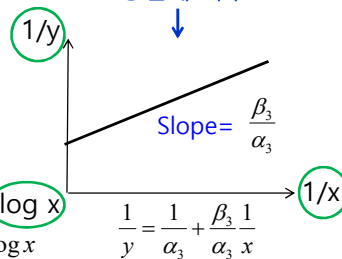


Power

경북대 전자공학부 김호희



양변에 역수



Saturation-growth-rate

12

Ex) Fit $y = \alpha x^\beta$ to the data in the table using a logarithmic transformation.

| i | x_i | y_i | $\text{Log } x_i$ | $\text{Log } y_i$ | $(\text{Log } x_i)^2$ | $\text{Log } x_i \text{ Log } y_i$ |
|----------|-------|-------|-------------------|-------------------|-----------------------|------------------------------------|
| 1 | 10 | 25 | 1.000 | 1.396 | 1.000 | 1.398 |
| 2 | 20 | 70 | 1.301 | 1.845 | 1.693 | 2.401 |
| 3 | 30 | 380 | 1.477 | 2.580 | 2.182 | 3.811 |
| 4 | 40 | 550 | 1.602 | 2.740 | 2.567 | 4.390 |
| 5 | 50 | 610 | 1.699 | 2.785 | 2.886 | 4.732 |
| 6 | 60 | 1220 | 1.778 | 3.086 | 3.162 | 5.488 |
| 7 | 70 | 830 | 1.845 | 2.919 | 3.404 | 5.386 |
| 8 | 80 | 1450 | 1.903 | 3.161 | 3.622 | 6.016 |
| Σ | | | 12.606 | 20.515 | 20.516 | 33.622 |

$$a_1 = \frac{n \sum \log x_i \log y_i - \sum \log x_i \sum \log y_i}{n \sum (\log x_i)^2 - (\sum \log x_i)^2} = \frac{8(33.622) - 12.606(20.515)}{8(20.516) - (12.606)^2} = 1.9842$$

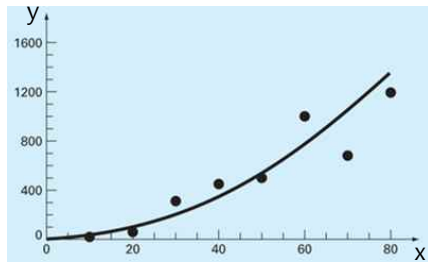
$$a_0 = \overline{\log y_i} - a_1 \overline{\log x_i} = -0.5620 \quad \Rightarrow \log y = -0.5620 + 1.9842 \log x \quad (\text{linear})$$

$$\Rightarrow y = \alpha x^{1.9842} \quad (-0.5620 = \log \alpha) \Rightarrow y = 0.2741x^{1.9842} \quad (\text{nonlinear})$$

수치해석-14장

경북대 전자공학부 김호희

13



nonlinear regression 하기 위해
각 점들을 log 형태 점으로 변환하여
linear regression 의 직선을 구한 뒤
nonlinear regression 의 곡선으로 변경

```
>> x = [ 10 20 30 40 50 60 70 80 ];
>> y = [25 70 380 550 610 1220 830 1450];
>> a = polyfit(x,y,1)
a = 19.4702 -234.2857
>> z = polyval(a,45)
z = 641.8750
>> b = polyfit(log10(x),log10(y),1)
b = 1.9842 -0.5620
>> rand(1,3)
ans = 0.2785 0.5469 0.9575
```

1차 다항식으로 fitting

a 가 계수인 다항식에 45를 대입

균일 분포 0~1 사이 값으로 이루어진
1X3 행렬 생성하는 built-in 함수

수치해석-14장

경북대 전자공학부 김호희

14