シミュレーション論 Ⅱ

第12回

データの分類と数値処理（1）

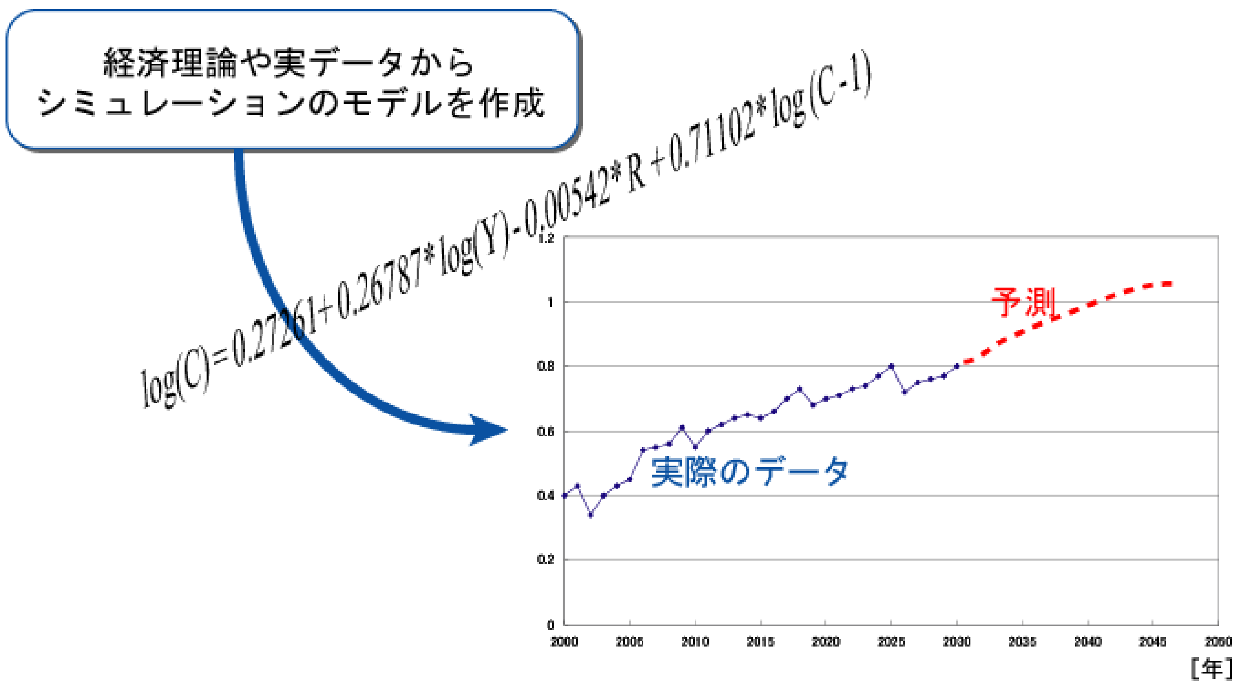
今回の内容

- 現実のデータや実験結果から数理モデルを作成するにはどのような手法があるのか？
- シミュレーション結果からデータの傾向、規則性を読み取るにはどうすればよいか？

⇒データ間の相互関係を分析する手法「回帰分析」

データの相互関係

- 実験やシミュレーション結果、実際の観察データ、種々の統計データなどからデータ間の関係を推定・整理する
- 原因-結果の相互関係が明確になり、モデルの作成や将来の予測が可能となる



回帰分析

- **回帰分析**(regression analysis)とは、**従属変数**(**目的変数**)と連続尺度の**独立変数**(**説明変数**)の間に式を当てはめ、従属変数が説明変数によってどれくらい説明できるのかを定量的に分析することである

  (Wikipediaより引用)

- 複数の変数間の関係をなんとか式であらわしたい！
  - **価格と需要の関係**
  - **所得と消費の関係**
  - **気温とおでんの売上の関係**

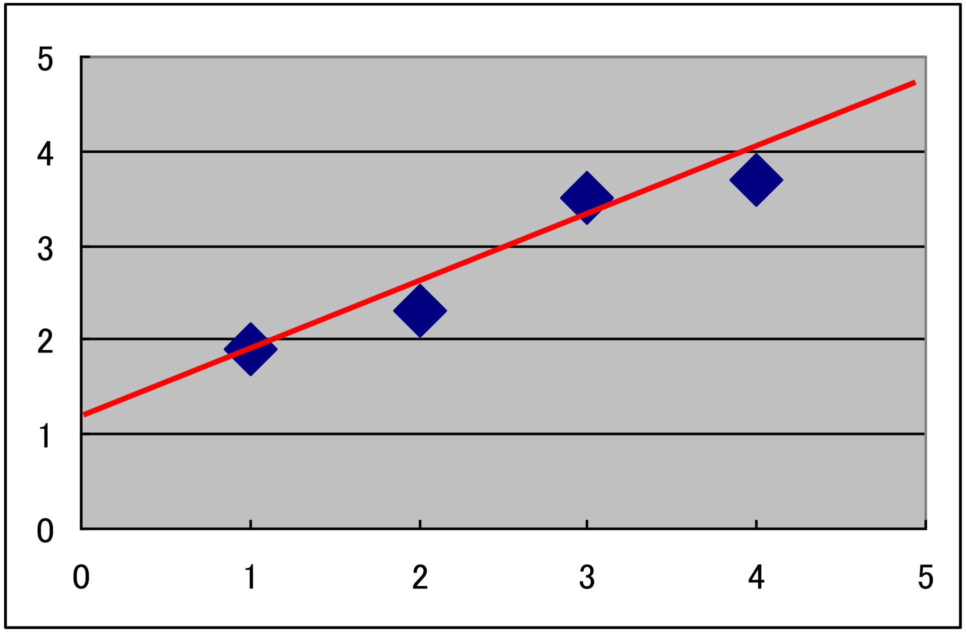
  など

直線回帰(線形回帰)

- 2つの変数間の関係を一次方程式で表すような回帰のことを直線回帰(線形回帰)という
- 2つの変数 X, Y の関係を一次方程式　$Y = aX + b$　で表すこと
- 複数のデータの組から回帰式を作成する場合には、もっとも「それらしい」式を作成してやる必要がある

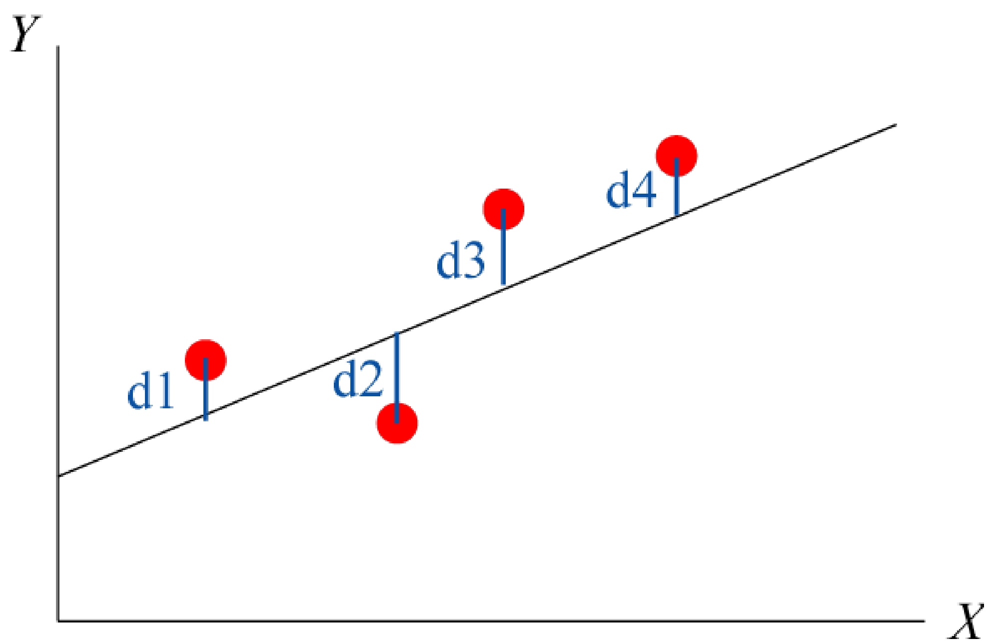
データと回帰直線

- 2つの変数間に線形の関係があれば、データは直線上に並ぶはず
- 実際は測定誤差などがあるため、ばらつきを含む
- 直線回帰では、「もっともらしい」直線（線形の回帰式）を作成する必要がある



最小二乗法

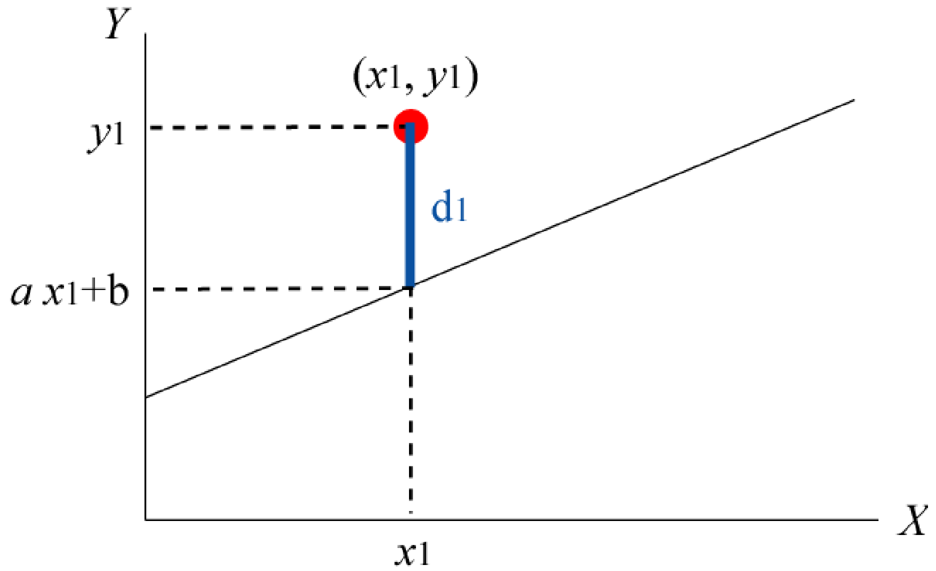
- 回帰直線を求めるもっとも一般的な方法
- 直線とデータとの誤差の二乗和を最小にすることで、回帰直線を求める



直線とデータとの誤差(d1,d2,d3,d4)の二乗和 $d1^2 + d2^2 + d3^2 + d4^2$ を最小にする

具体的な計算方法

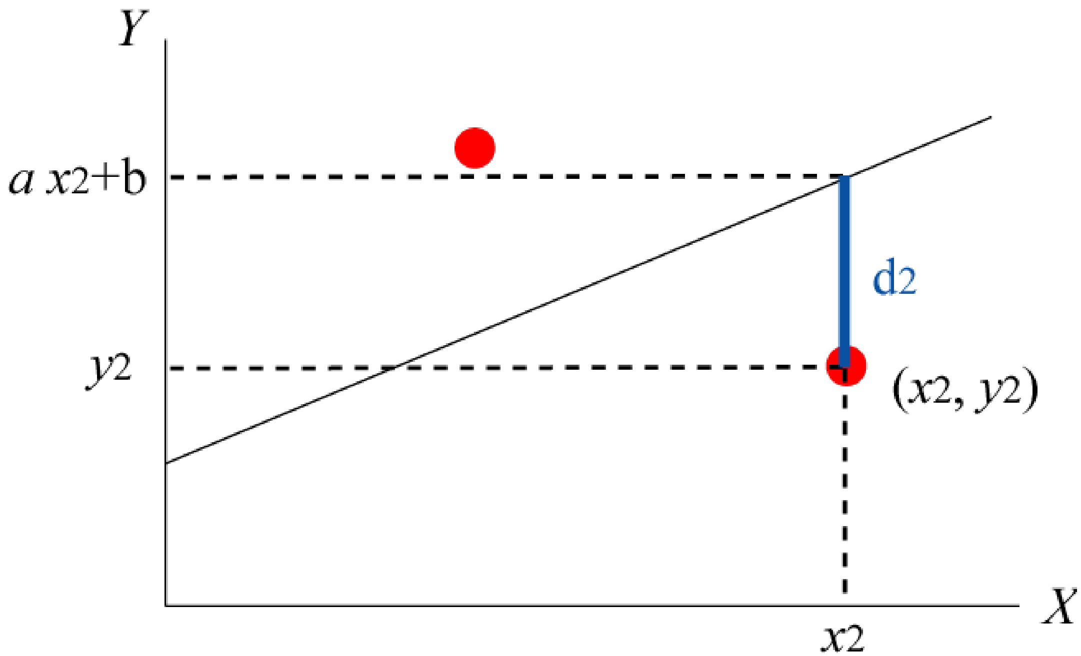
- 2変数X, Yの関係を　Y = aX + b （直線）と仮定する
- データと直線との誤差を求め、二乗する



$$d_1^2 = (y_1 - ax_1 - b)^2$$

具体的な計算方法(2)

- d2についても計算してみよう



$$d_1^2 = (ax_2 + b - y_2)^2 = \{-(y_2 - ax_2 - b)\}^2$$
$$= (y_2 - ax_2 - b)^2$$

具体的な計算方法(3)

- 他の誤差についても同様に計算できる

$$d_1^2 = (y_1 - ax_1 - b)^2$$
$$d_2^2 = (y_2 - ax_2 - b)^2$$
$$d_3^2 = (y_3 - ax_3 - b)^2$$
$$d_4^2 = (y_4 - ax_4 - b)^2$$

- よって誤差の二乗和をSとおくと

$$S = d_1^2 + d_2^2 + d_3^2 + d_4^2$$

となり、この S を最小にするような a, b を求めればよい

具体的な計算方法(4)

- Sを展開して整理すると

$$S=(y_1^2+y_2^2+y_3^2+y_4^2)+a^2(x_1^2+x_2^2+x_3^2+x_4^2)+4b^2$$
$$-2b(y_1+y_2+y_3+y_4)-2a(x_1y_1+x_2y_2+x_3y_3+x_4y_4)$$
$$+2ab(x_1+x_2+x_3+x_4)$$

- ここで、式を簡単にするために以下のようにおく

$$A=(y_1^2+y_2^2+y_3^2+y_4^2)$$
$$B=(x_1^2+x_2^2+x_3^2+x_4^2)$$
$$C=(y_1+y_2+y_3+y_4)$$
$$D=(x_1y_1+x_2y_2+x_3y_3+x_4y_4)$$
$$E=(x_1+x_2+x_3+x_4)$$

具体的な計算方法(5)

- 先ほどの置き換えで

$$S = A + a^2 B + 4b^2 - 2bC - 2aD + 2abE$$

と書ける。

- A, B, C, D, E は全てデータから計算できる定数なので、変数は a, b の2つとなる
- b を定数と仮定すれば、Sはaの二次方程式とみなせる
- a を定数と仮定すれば、Sはbの二次方程式とみなせる
- S は二乗の和だから常に0以上の値
- S を a, b で偏微分して0とおいた連立方程式を作り、最小値を求める

具体的な計算方法(6)

- S を a, b で偏微分して0とおくと

$$\frac{\partial S}{\partial a} = 2aB - 2D + 2bE = 0$$

$$\frac{\partial S}{\partial b} = 8b - 2C + 2aE = 0$$

が得られる。両辺を2で割って、結局

$$\begin{cases} aB - D + bE = 0 \\ aE - C + 4b = 0 \end{cases}$$

という連立方程式を解けば a, b が求められる

具体的な計算方法(7)

- 先ほどの連立方程式を解いて

$$\boxed{\begin{cases} a = \dfrac{4D - CE}{4B - E^2} \\ b = \dfrac{BC - DE}{4B - E^2} \end{cases}}$$

が得られる。このようにして、回帰直線 Y = aX + b の係数 a, b が決定される。ただしB, C, D, Eは以下のとおり。

$$B = (x_1^2 + x_2^2 + x_3^2 + x_4^2)$$

$$C = (y_1 + y_2 + y_3 + y_4)$$

$$D = (x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4)$$

$$E = (x_1 + x_2 + x_3 + x_4)$$

一般的な場合

- これまでの計算はデータが4組の場合だったが、一般に n 組のデータ $(x_1, y_1),\ (x_2, y_2),\ (x_3, y_3),\ \Lambda\ ,\ (x_n, y_n)$ がある場合にも同様の計算で a, b が求められる

$$\boxed{\begin{cases} a = \dfrac{nD - CE}{nB - E^2} \\ b = \dfrac{BC - DE}{nB - E^2} \end{cases}}$$

- ただし

$$B = \sum_{i=1}^{n} x_i^2 ,\qquad C = \sum_{i=1}^{n} y_i$$

$$D = \sum_{i=1}^{n} x_i y_i ,\qquad E = \sum_{i=1}^{n} x_i$$

一般的な場合(2)

- B, C, D, Eを元の式に戻して、一般的に最小二乗法によって n 組のデータを Y = aX + b の形に線形回帰すると a, b は

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

参考

- 線形回帰の傾き a および切片 b は以下のようにに計算することもできる

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

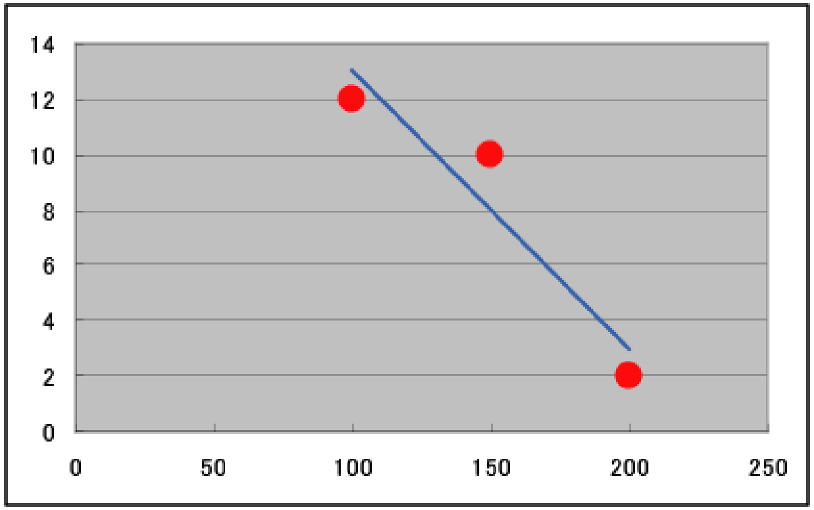
$$b = \bar{y} - a\bar{x}$$

ただし　$\bar{x}, \bar{y}$　は　$x,\ y$　の平均値

例：需要関数の推定

- あるパン屋で、同じパンを100円、150円、200円で販売したときの1日の売上はそれぞれ12個、10個、2個だった。
- 需要関数（価格と売上の関係）が一次方程式 Y = aX + b で表せると仮定して線形回帰をおこなってみよう。

価格(円)	売上(個)
100	12
150	10
200	2



例：需要関数の推定(2)

- この例題ではデータ数 n = 3 で、データの組はそれぞれ
  $(x_1, y_1) = (100,12), \quad (x_2, y_2) = (150,10), \quad (x_3, y_3) = (200,2)$
  a, b は以下の式で求められる

$$\boxed{\begin{cases} a = \dfrac{nD - CE}{nB - E^2} \\ b = \dfrac{BC - DE}{nB - E^2} \end{cases}}$$

- ただし

$$B = \sum_{i=1}^{n} x_i^2, \quad C = \sum_{i=1}^{n} y_i$$

$$D = \sum_{i=1}^{n} x_i y_i, \quad E = \sum_{i=1}^{n} x_i$$

例：需要関数の推定(3)

- まずB～Eを求めると

$$B=\sum_{i=1}^{n}x_i^2=100^2+150^2+200^2=72500$$

$$C=\sum_{i=1}^{n}y_i=12+10+2=24$$

$$D=\sum_{i=1}^{n}x_iy_i=100\times12+150\times10+200\times2=3100$$

$$E=\sum_{i=1}^{n}x_i=100+150+200=450$$

例:需要関数の推定(4)

- これを a, b の式に代入して

$$a = \frac{nD - CE}{nB - E^2} = \frac{3 \times 3100 - 24 \times 450}{3 \times 72500 - 450^2} = -0.1$$

$$b = \frac{BC - DE}{nB - E^2} = \frac{72500 \times 24 - 3100 \times 450}{3 \times 72500 - 450^2} = 23$$

- よって需要関数の推定値(回帰直線)は Y = −0.1X + 23
- たとえば50円で売ったとすると Y = −0.1 × 50 + 23 = 18で18個売れると予想される