多変量解析

第6回　単回帰分析

萩原•篠田
情報理工学部

回帰分析

マンション価格は広さと築年数から予測可能か？

量的変数　　量的変数

回帰式：$\hat{y} = 1.02 + 0.067x_1 - 0.081x_2$

サンプル No.	広さ（m^2） x_1	築年数（年） x_2	価格（千万円） y
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

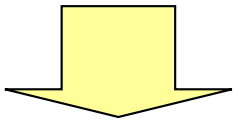
keywords

目的変数、説明変数、
線形回帰、残差、
最小二乗法、
決定係数（寄与率）、
分散共分散行列

回帰分析

回帰分析とは、目的変数（従属変数）と連続尺度の説明変数（独立変数）の間に式を当てはめ、目的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。

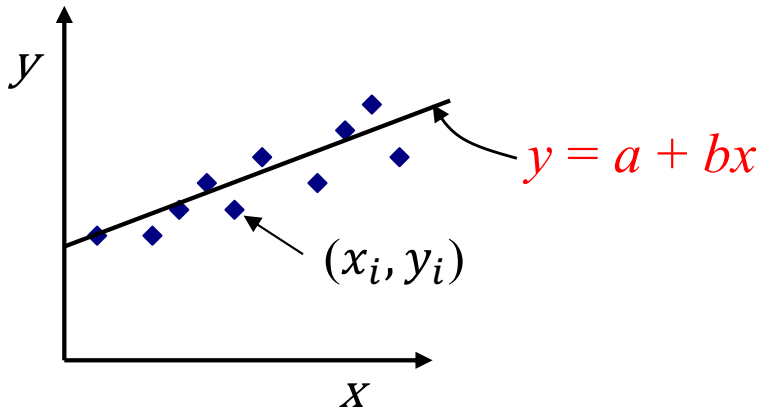
2つの変量(x , y)の関係について、x は指定できる変数（独立変数）であり、指定された x に対して y があるばらつきをもって決まる場合、x と y の関係を単回帰という。



両変数間の数的関係を回帰直線で表し、ある x が指定されたときに y がいくつになるかを求めることを主な目的とする。

回帰直線

平面上の n 個の点 (x_i, y_i) $(i = 1,2,\ldots..,n)$にあてはめた直線を回帰直線と呼ぶ。
回帰直線は目的変数yを1つの説明変数xから推定しようとする単回帰分析



その式を $y = a + bx$ とするとその係数 a,b は次式で与えられる。

傾き $$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\sum x_i y_i-(\sum x_i \sum y_i)/n}{\sum x_i^2-(\sum x_i)^2/n} \quad \cdots\cdots ①$$

切片 $$a = \bar{y} - b\bar{x} = \frac{1}{n}(\sum y_i - b\sum x_i) \quad \cdots\cdots ②$$

S_{xx} ：x の偏差平方和

S_{xy} ：xとy の偏差積和

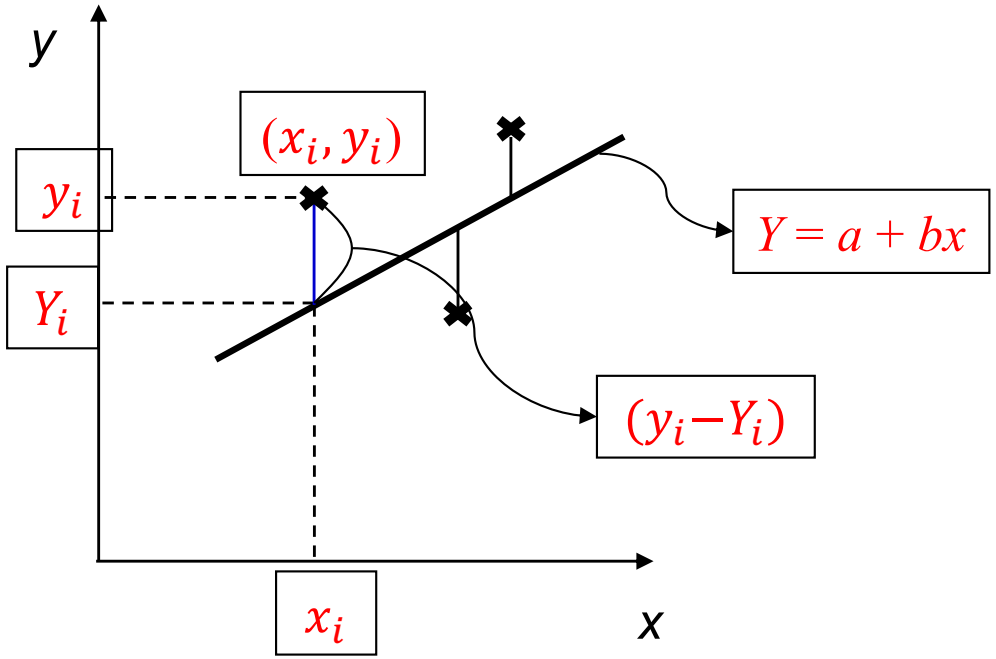
最小二乗法の原理と回帰直線(1)

n 個の点に対する回帰直線はたくさん引ける。その中でデータに最もフィットしたものとして各点 (x_i, y_i) から回帰直線までの垂線距離の2乗和S（回帰からの偏差平方和）が最小となる場合の直線の式を求める

求める回帰式を $y = a + bx$ としてn個の点総てについて実測値 y_i と回帰直線上の推定値 Y_i との差（回帰残差）$y_i - Y_i$の平方和を求める

$$S = \sum(y_i - Y_i)^2 = \sum(y_i - a - bx_i)^2$$

$$Y_i = a + bx_i$$



最小二乗法の原理と回帰直線(2)

S を最小にする a , b を求めるには、S を a , b の関数と見て a , b で偏微分し
その関数$\partial S/\partial a$, $\partial S/\partial b$ がそれぞれ0になる場合を計算すればよい

$T = y_i - a - bx_i$ とおいてaについて偏微分

$$\frac{\partial S}{\partial a} = \frac{\partial \sum (y_i - a - bx_i)^2}{\partial a}$$
$$= \frac{\partial \sum T^2}{\partial T} \times \frac{\partial T}{\partial a}$$
$$= 2\sum (y_i - a - bx_i) \cdot (-1)$$
$$= -2\sum (y_i - a - bx_i) = 0$$

bについて偏微分

$$\frac{\partial S}{\partial b} = \frac{\partial \sum (y_i - a - bx_i)^2}{\partial b}$$
$$= \frac{\partial \sum T^2}{\partial T} \times \frac{\partial T}{\partial b}$$
$$= 2\sum (y_i - a - bx_i) \cdot (-x_i)$$
$$= -2\sum (x_i y_i - ax_i - bx_i^2) = 0$$

整理すると

$$n \cdot a + (\sum x_i) b = \sum y_i$$
$$(\sum x_i)\, a + (\sum x_i^2)\, b = \sum x_i y_i$$

この連立方程式よりa, bをとくと①、②が求められる。

$$b = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad \cdots\cdots ①$$

$$a = \frac{1}{n}(\sum y_i - b \sum x_i) \quad \cdots\cdots ②$$

$\sum x_i^2 - \frac{(\sum x_i)^2}{n} = \sum (x_i - \bar{x})^2, \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = \sum (x_i - \bar{x})(y_i - \bar{y})$ を証明してみよう！

次の10点を通る回帰直線 $y = a + bx$ を最小二乗法で求めよ

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

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

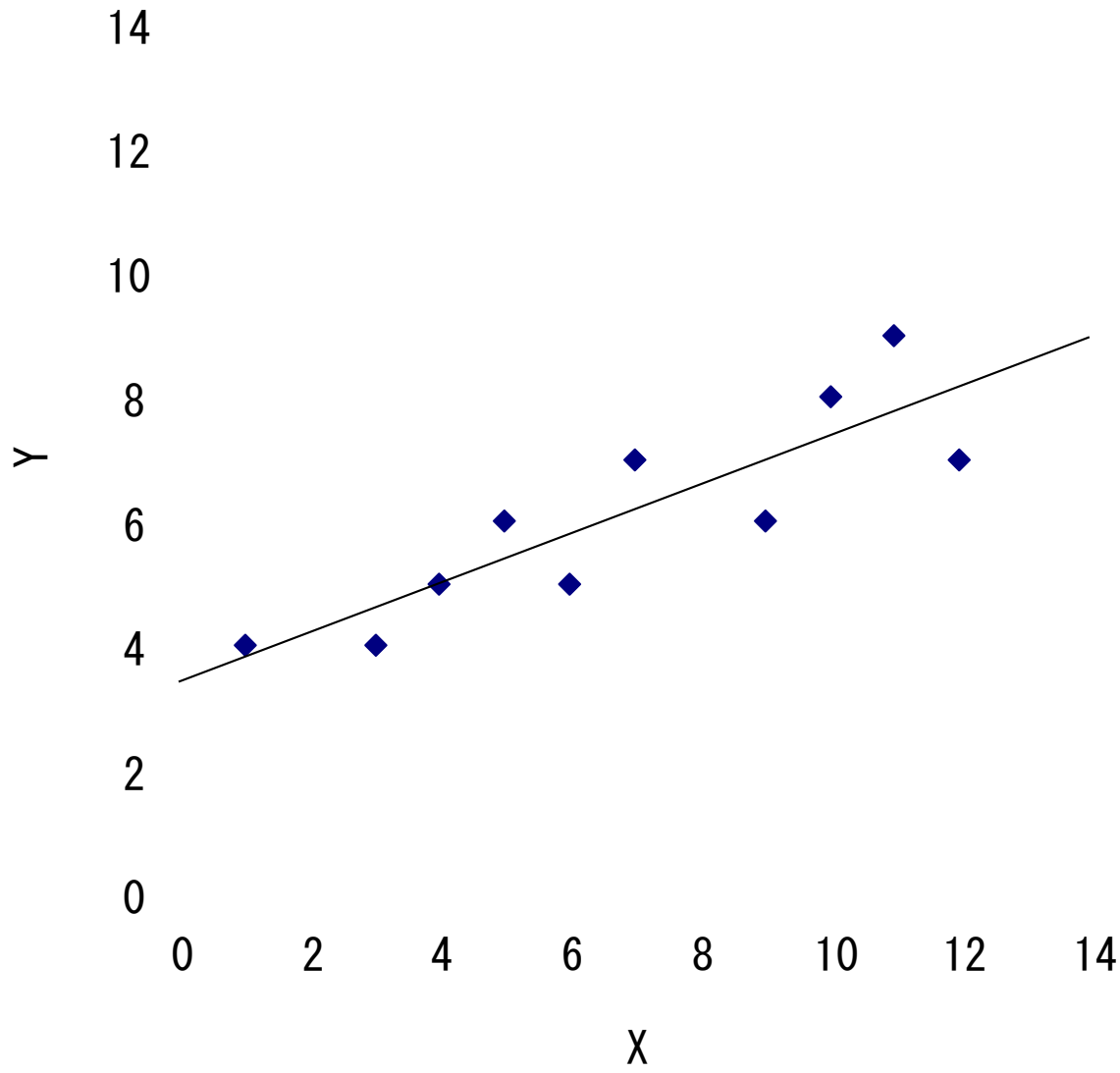
$$= \frac{462 - 68 \times 61/10}{582 - 68^2/10} = \frac{47.2}{119.6} = 0.39$$

$$a = \bar{y} - b\bar{x} = \frac{1}{n}(\sum y_i - b \sum x_i)$$

$$= (61 - 0.39 \times 68)/10 = 3.4$$

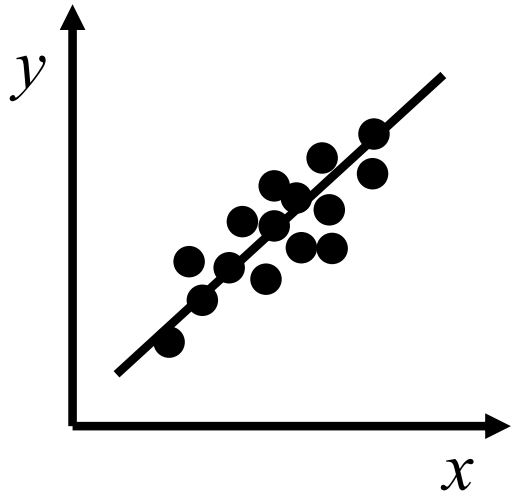
x	y	x^2	xy	y^2
1	4	1	4	16
3	4	9	12	16
4	5	16	20	25
5	6	25	30	36
6	5	36	30	25
7	7	49	49	49
9	6	81	54	36
10	8	100	80	64
11	9	121	99	81
12	7	144	84	49
Σx_i	Σy_i	Σx_i^2	$\Sigma x_i y_i$	Σy_i^2
68	61	582	462	397

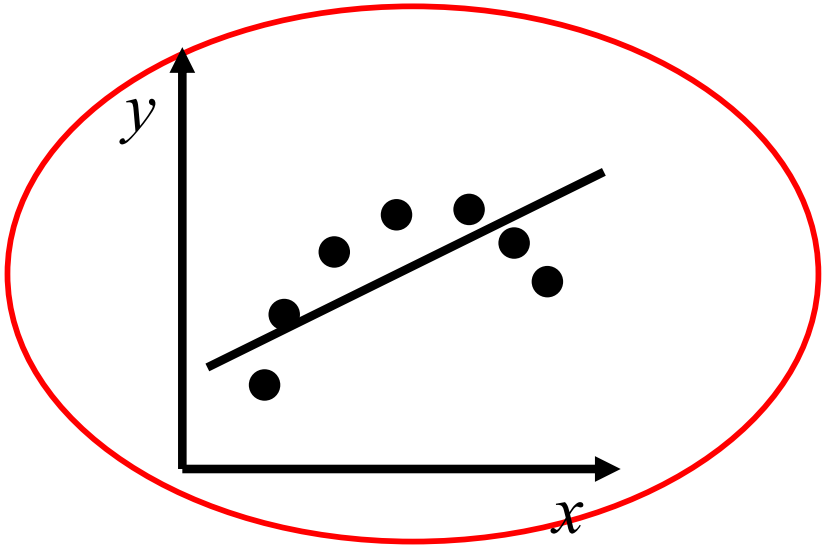
従って求める回帰式は　$y = 3.4 + 0.39x$


14
12
10
8
6
4
2
0
Y
0
2
4
6
8
10
12
14
X

回帰直線の計算（チェックポイント）

（1） 直線関係と考えてよいか？



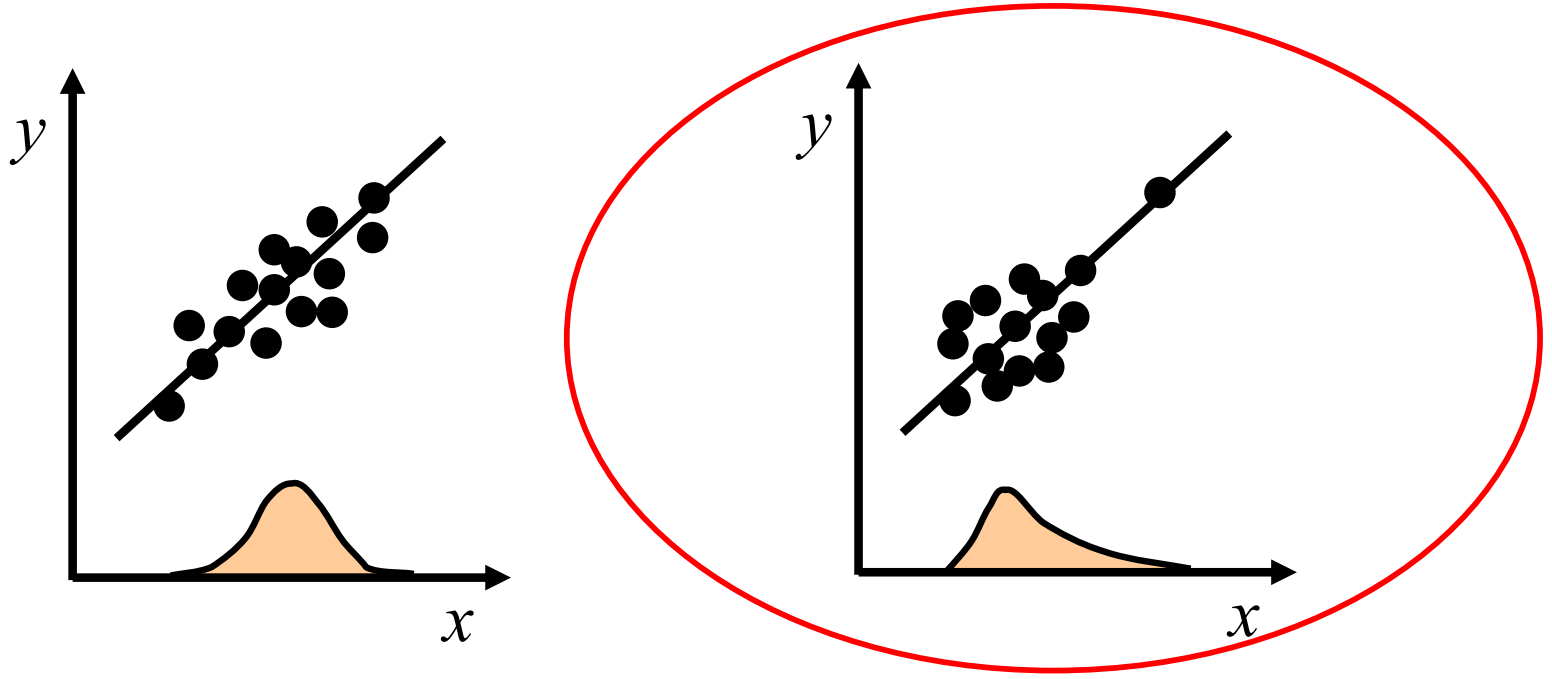


≪対策≫

- y を変数変換して直線化　ex. $\log(y) = a + bx$
- 曲線回帰を試みる　ex. 2次多項式　$y = a + bx + cx^2$
  整次多項式　$y = a + bx + cx^2 + dx^3$

回帰直線の計算（チェックポイント）

（2） x の分布に偏りはないか？

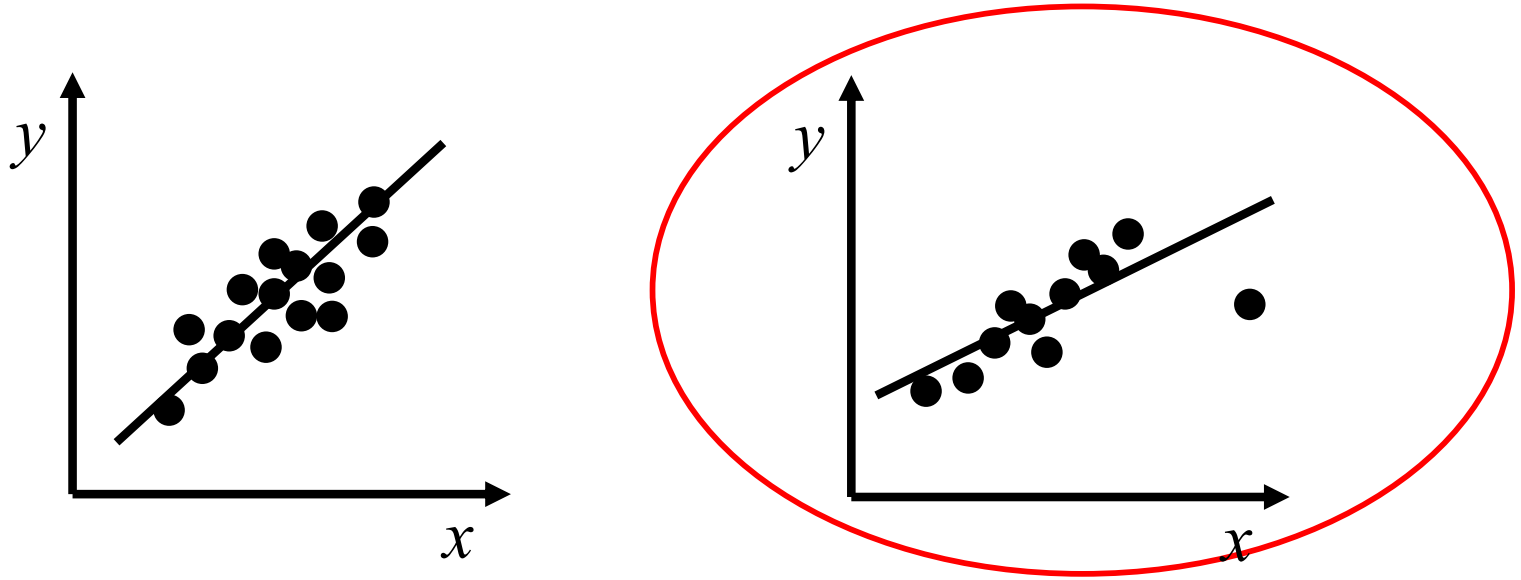


≪対策≫

・できるだけ対称な分布となるよう x を変数変換

回帰直線の計算（チェックポイント）

（3） 飛び離れ点はないか？

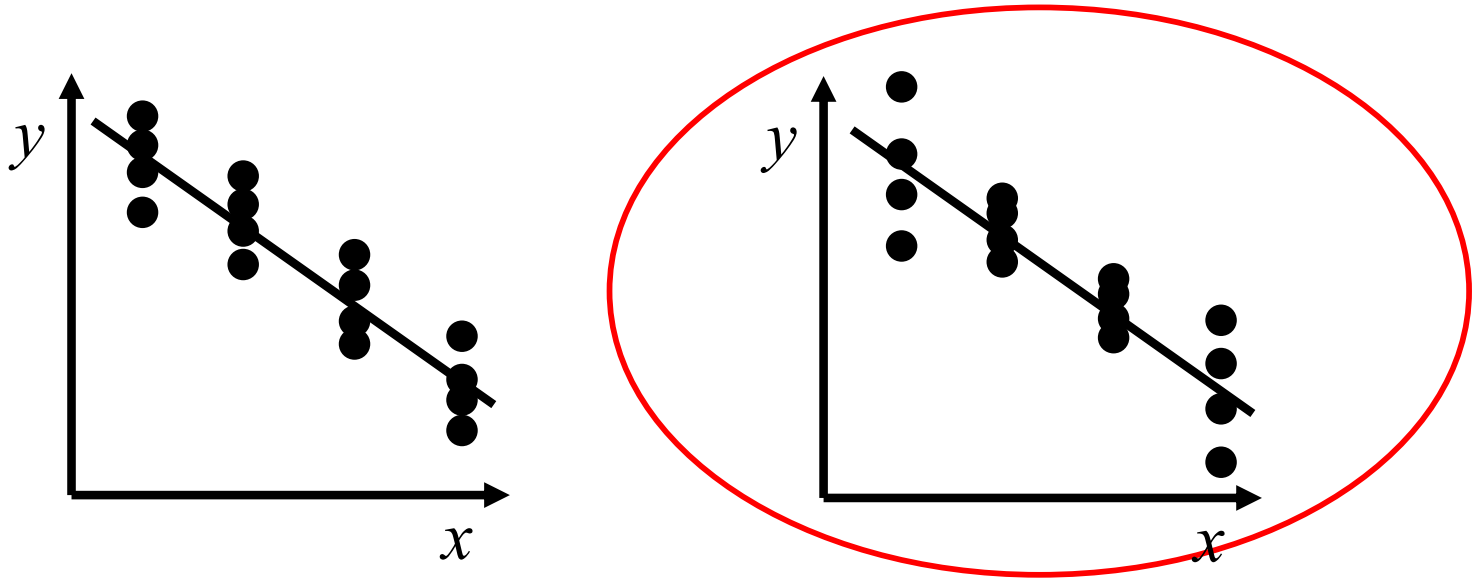


≪対策≫

・データの点検、棄却検定

回帰直線の計算（チェックポイント）

（4） y の分散は、x の値によらず均一とみなせるか？

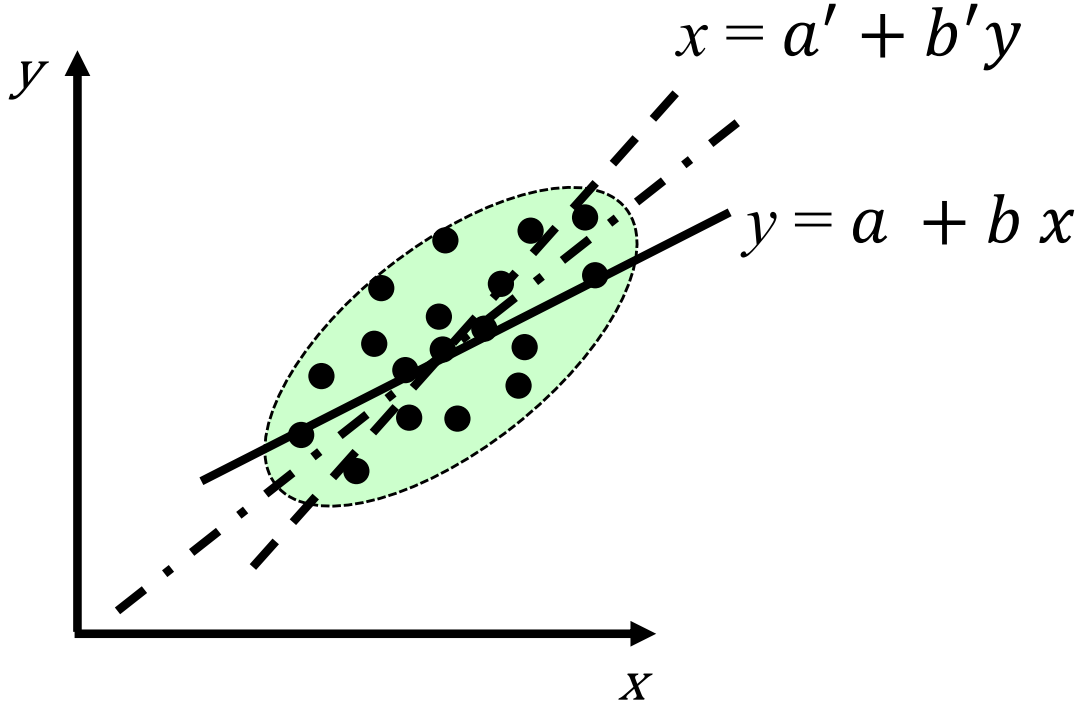


≪対策≫

- 重み付き回帰を試みる
- y の変数変換による場合そのまま曲線回帰

回帰直線の計算（チェックポイント）

（5） 変数 x , y のとり方は妥当か？



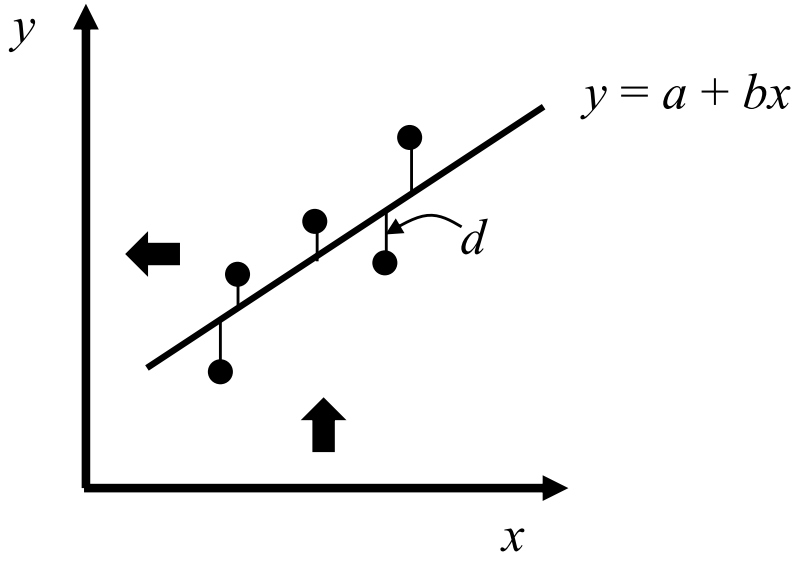
- x から y を予測するのか？
- y から x を予測するのか？
- 楕円の長軸を求めるのか？

回帰の方向性と求心性

・何を説明変数にするかで、回帰式が変わる

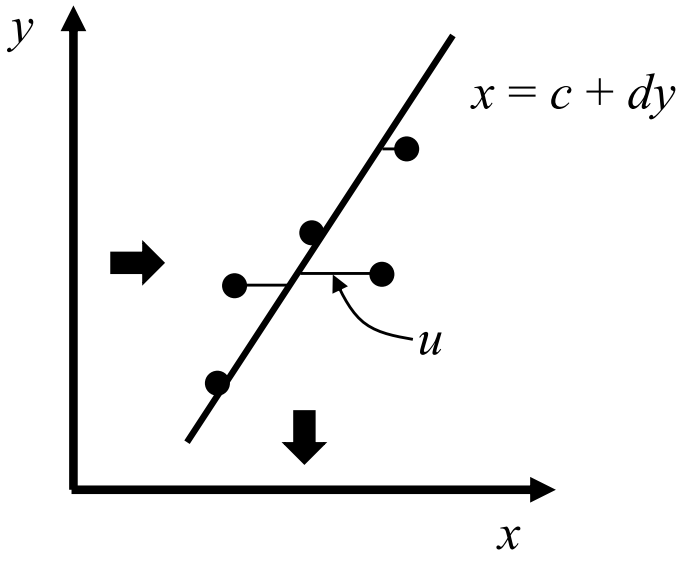
（回帰の方向性）

x から y を回帰



$S = \sum d^2$ を最小にする
a, b を求める

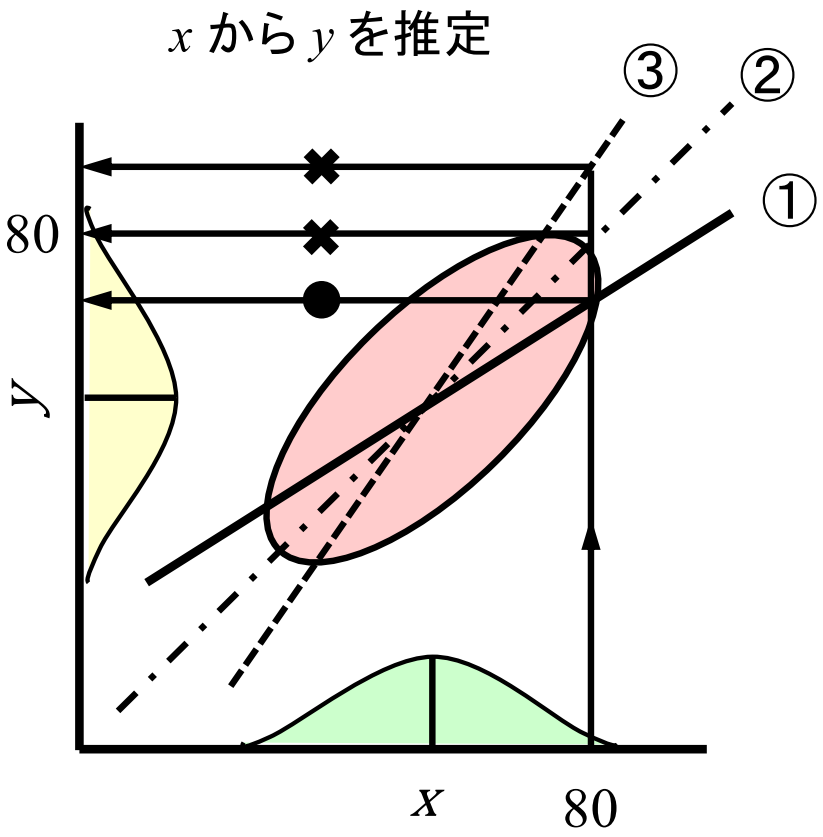
y から x を回帰

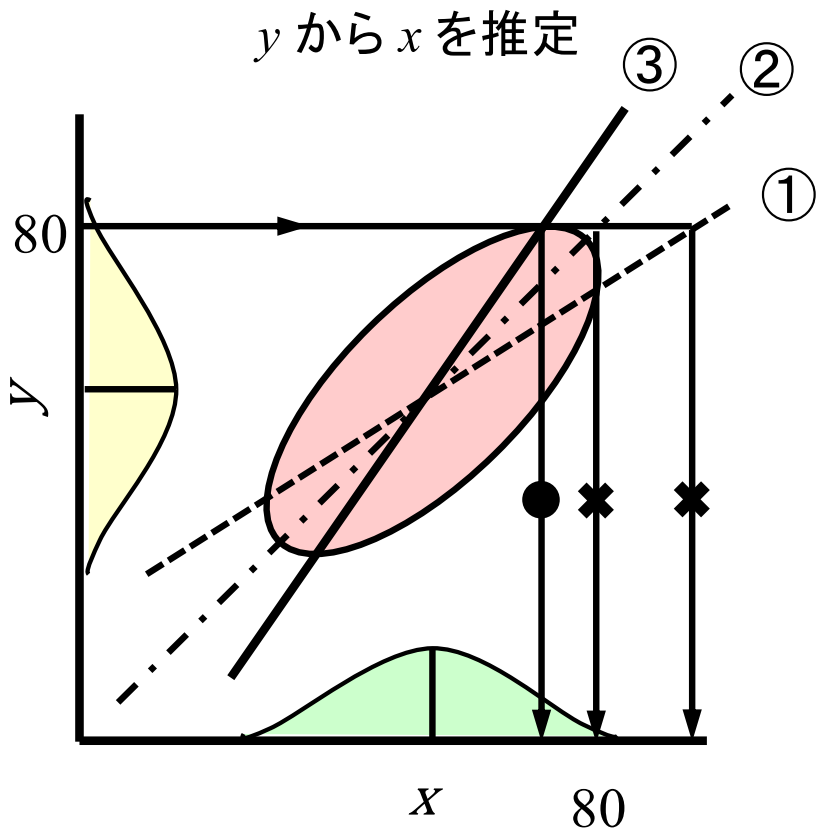


$S = \sum u^2$ を最小にする
c, d を求める

回帰の方向性と求心性

・回帰直線を2本引ける意味（回帰の求心性）



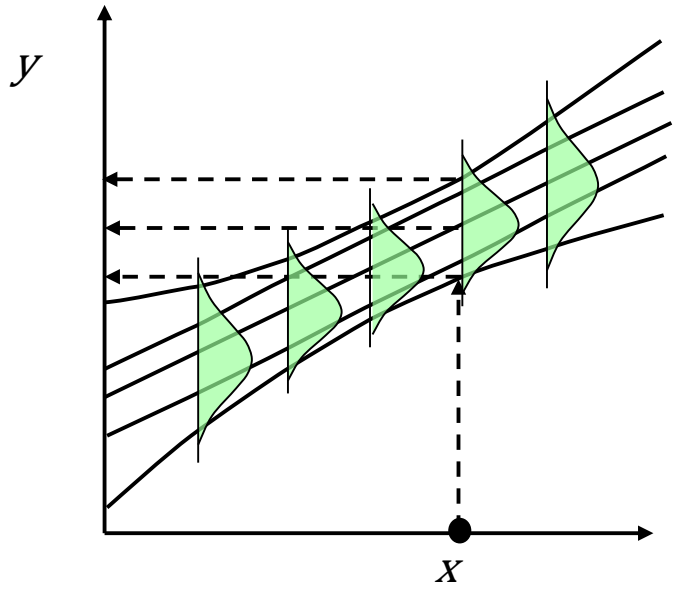


回帰直線の信頼区間

- データをサンプリングするために、(標本)回帰直線も変わる
- しかし、真の(母)回帰直線は、一定の範囲にいるはず

- 回帰直線の信頼域
  - xを基準にしたyの母平均など、母数に対する確率的な領域を指す

- 回帰直線の95%信頼区間(Confidence Interval, CI)
  - 95%の確率でyの母平均などが存在すると推定できる区間を指す
    - データが多く密集しているところが幅が狭まる
    - 求め方は省略



母数：確率分布を特定するための定数.

回帰分析と相関分析

・回帰分析

回帰分析とは、目的変数（従属変数）と連続尺度の説明変数（独立変数）の間に式を当てはめ、目的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。

2つの変量(x,y)の関係について、xは指定できる変数（独立変数）であり、指定されたxに対してyがあるばらつきをもって決まる場合、xとyの関係を単回帰という。

両変数間の数的関係を回帰直線で表し、あるxが指定されたときにyがいくつになるかを求めることを主な目的とする。

・相関分析

量的な（順序尺度を含む）2変数間の関係を分析する際に用いる。

2つの変量(x,y)の関係について、x,yともに正規分布にしたがってばらつく量であるときには両者の関係を相関分析する。

相関分析では両変数間の関連の度合いを相関係数で評価することを主な目的とする。

回帰と相関の違い

１．計算目的の違い

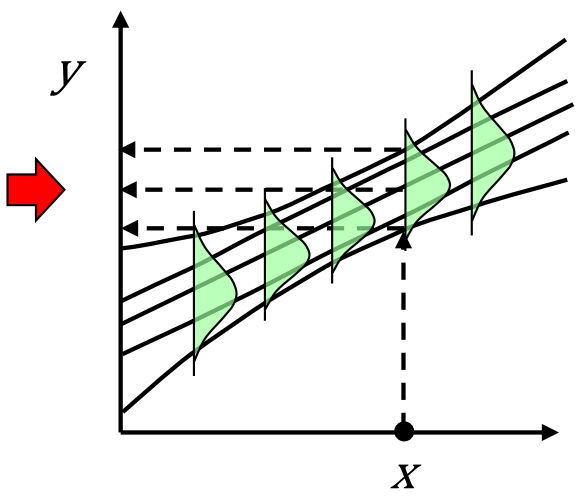
回帰直線：x から y をどのように直線的に関係付けられる（x,yの単位に依存）
相関係数：xとyの相互関係がどの程度直線的か（x,yの単位に依存しない）

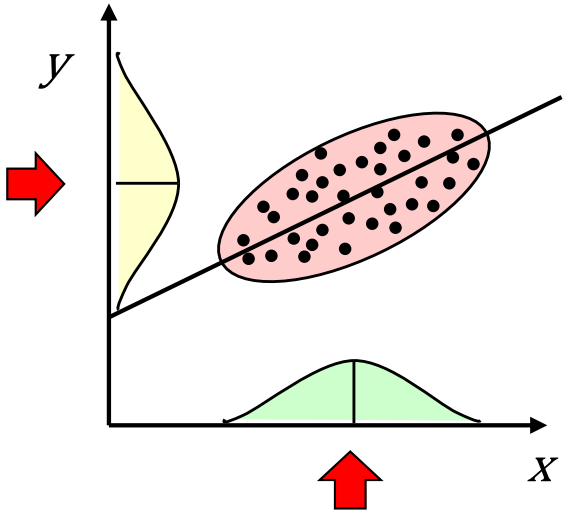
２．誤差の取り扱いの違い：2変量x,yの関係を調べるとき

回帰分析：一方(たとえばx)を基準にして他方(y)を関連付ける
y側だけをばらつきある確率変数として取り扱う
相関分析：x,yともにばらつきのある確率変数とみなし,相互の関係の強さを見る

３．信頼域の違い

回帰分析の信頼域：xを基準にしたyの信頼区間；回帰直線の周りの帯状の領域
相関分析の信頼域：x,y分布の中心からの等確率距離；平面状楕円の領域





単回帰分析

①回帰直線を求めるための最小二乗法の原理を図を用いて説明せよ。

②計測された評価項目間のデータの関連性を求める回帰と相関の違いについて説明せよ。

1. 計算目的の違い
2. 誤差の取り扱いの違い：2変量x,yの関係を調べるとき
3. 信頼域の違い

例題

次の10点に対する回帰直線$y = a + bx$を最小二乗法で求めよ

sample#	x_i	y_i
1	2.2	71
2	4.1	81
3	5.5	86
4	1.9	72
5	3.4	77
6	2.6	73
7	4.2	80
8	3.7	81
9	4.9	85
10	3.2	74
$\sum$	35.7	780