2019年度春学期　統計学　　第７回

データの関係を知る(2)—回帰分析

浅野　晃
関西大学総合情報学部

回帰分析とは🤔

A. Asano, Kansai Univ.

回帰分析とは

多変量データがあるとき
ある変量の変化を他の変量の変化で
［説明］する方法

説明？🤔

A. Asano, Kansai Univ.

回帰分析とは

緯度と気温のデータを例にとると

相関分析
「緯度が上がると，気温が下がる」という
傾向がはっきりしている

回帰分析
「緯度が上がるから気温が下がる」と考える
緯度が１度上がると，気温が◯℃下がる

A. Asano, Kansai Univ.

回帰分析とは

緯度が上がるから気温が下がると考える
緯度が１度上がると，気温が〇℃下がる

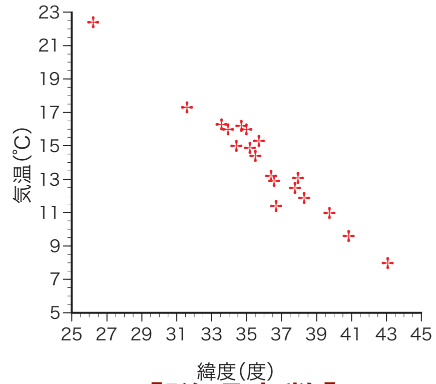
各都市の気温の違いは，緯度によって決まっているという［モデル］を考える

統計学では，
気温（のばらつき）は，緯度によって［説明］されるという

説明変数・被説明変数

気温は緯度によって説明される
（というモデル）

［被説明変数］

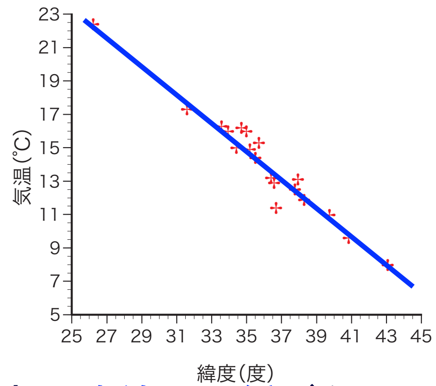


［説明変数］

線形単回帰🤔

線形単回帰

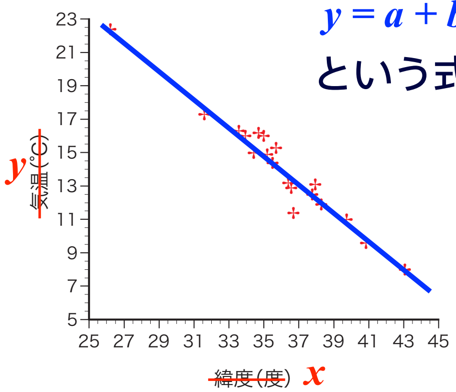
気温は緯度によって説明される
どう説明される？



散布図上で直線の関係がある，と考える

線形単回帰

散布図上で**直線の関係**がある



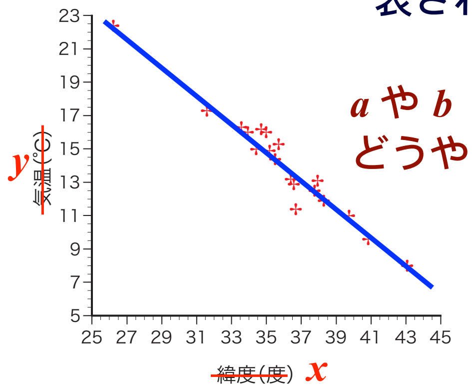
$y = a + bx$

という式で表される関係

[線形単回帰]
という

線形単回帰

$y = a + bx$ という式で
表される関係

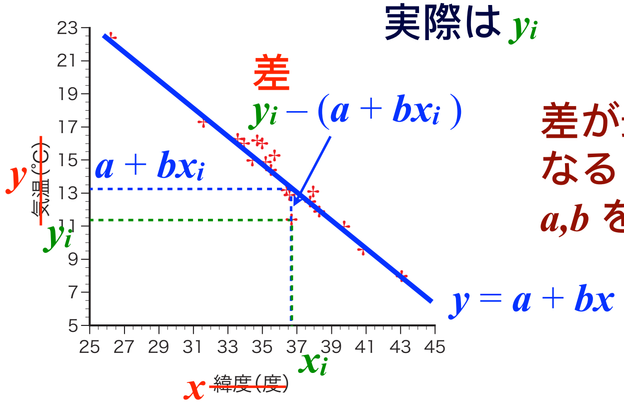


a や b （パラメータ）は
どうやって求める？

パラメータの決定

$x = x_i$ のとき

モデルによれば $a + bx_i$
実際は y_i

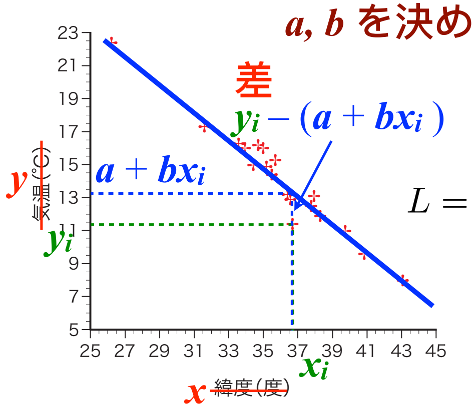


差が最小に
なるように
a,b を決める

パラメータの決定

すべての x_i について，
差の合計が最小になるように
a, b を決める

2



$$L = \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2$$

が最小になる
a, b を求める

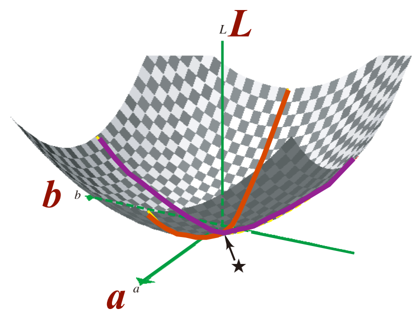
Lが最小になるa,bを求める

- 偏微分による方法（付録１）
- 「２次関数の最大・最小」による方法（付録２）

「偏微分」による方法

$$L = \underbrace{\sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2}_{a,\, b\ \text{の２次関数}}$$

が最小になる
a, *b* を求める

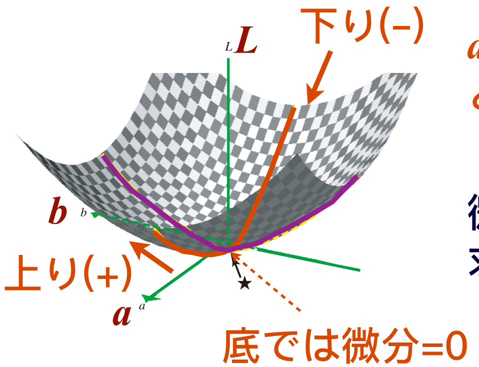


a だけの関数
と考えて微分

b だけの関数
と考えて微分

微分？

微分?



a だけの関数
と考えて微分

微分は，傾きを
求める計算

b についても同じ，
底では微分=0

これらから
a, *b* を求める

計算はともかく結論は

- 偏微分による方法（付録１）
- 「２次関数の最大・最小」による方法（付録２）

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

x, *y* の共分散
x の分散

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

y の平均
x の平均

最小二乗法

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

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

$$L = \sum_{i=1}^{n}\{y_i - (a + bx_i)\}^2$$

を最小にしたので
[最小二乗法]

$$y = a + bx$$

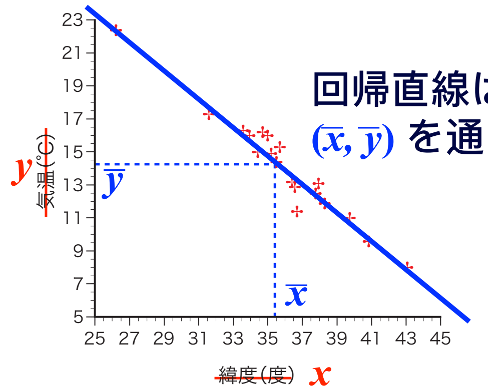
[回帰方程式] あるいは
[回帰直線]

[回帰係数]

ところで

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

から $y - \bar{y} = b(x - \bar{x})$

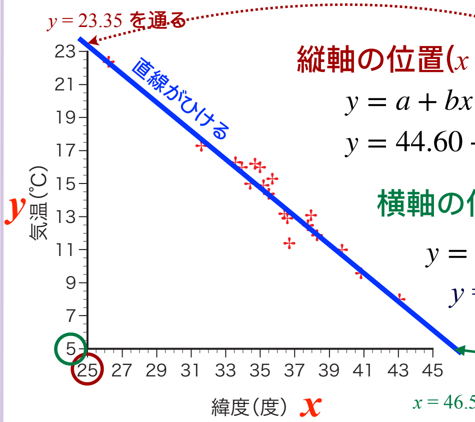


回帰直線は
$(\bar{x}, \bar{y})$ を通る

緯度と気温の例では

緯度をx
気温をy として回帰直線 $y = a + bx$ を求める

$\rightarrow b = -0.850, \quad a = 44.60$



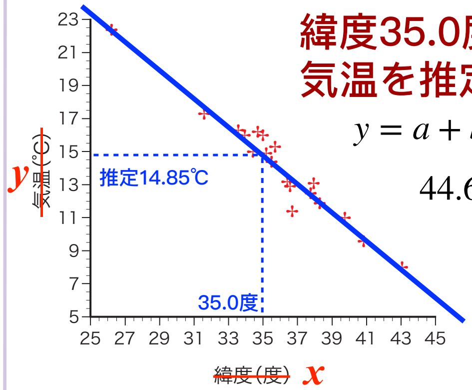
縦軸の位置($x = 25$)のときyの値は
$y = a + bx$ に $x = 25.0$ を代入
$y = 44.60 + (-0.850) \times 25.0 = \boxed{23.35}$

横軸の位置($y = 5$)のときxの値は
$y = a + bx$ より $x = \dfrac{y - a}{b}$
$y = 5$ を代入すると
$x = (5 - 44.60)/(-0.850)$
$= \boxed{46.59}$

緯度と気温の例では

緯度をx
気温をy として回帰直線 $y = a + bx$ を求める

$\rightarrow b = -0.850, \quad a = 44.60$



緯度35.0度の都市の
気温を推定

$y = a + bx$ に $x = 35.0$ を代入

$44.60 + (-0.850) \times 35.0$
$= 14.85$ (℃)

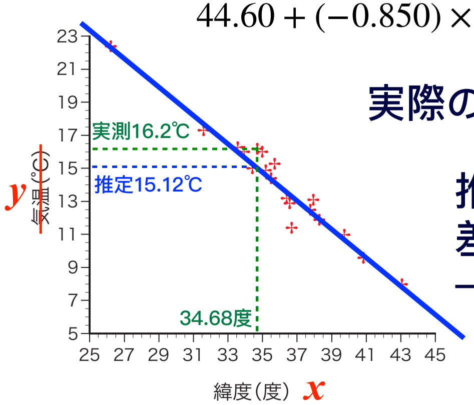
A. Asano, Kansai Univ.

緯度と気温の例では

表にある大阪市(緯度34.68度)の気温を推定

$y = a + bx$ に $x = 34.68$ を代入

$44.60 + (-0.850) \times 34.68 = 15.12$ (℃)



実際の気温は16.2℃

推定値と実測値に差がある
→次の話へ

A. Asano, Kansai Univ.

決定係数🤔

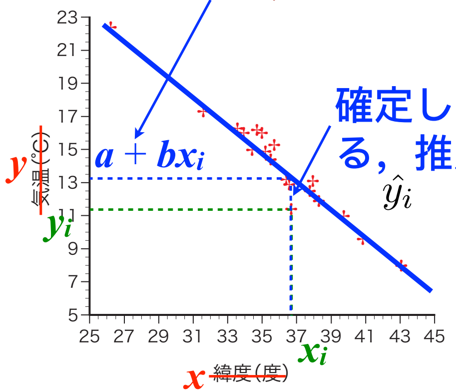
A. Asano, Kansai Univ.

残差

a, b が求められて，回帰直線が確定

x_i に対する，回帰直線による y の推定値

$\hat{y}_i = a + bx_i$



確定しても残っている，推定値と実際の差

$\hat{y}_i$　　y_i

［残差］という

d_i

A. Asano, Kansai Univ.

残差と決定係数

回帰方程式を使って y_i を予測したときの，予測によって表現できなかった部分

残差について（付録３）

$$\sum d_i^2 = (1 - r_{xy}^2) \sum (y_i - \bar{y})^2$$

残差　　相関係数　　決定係数

決定係数の意味

$\sum d_i^2 = (1 - r_{xy}^2) \sum (y_i - \bar{y})^2$ より

残差の２乗の平均

$$1 - r_{xy}^2 = \frac{\sum d_i^2 / n}{\sum (y_i - \bar{y})^2 / n}$$

決定係数

y の偏差の２乗の平均

＝ y の分散

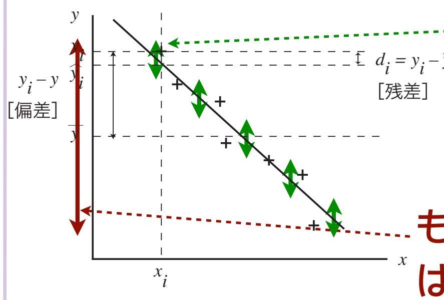
決定係数の意味

$$1 - r_{xy}^2 = \frac{\sum d_i^2 / n}{\sum (y_i - \bar{y})^2 / n}$$

決定係数

残差の２乗の平均

y の偏差の２乗の平均（y の分散）



回帰直線から見ると
ばらつきは
こんなに減った

もともと y はこんなに
ばらついていたが，

決定係数の意味

$$1 - r_{xy}^2 = \frac{\sum d_i^2 / n}{\sum (y_i - \bar{y})^2 / n}$$

決定係数

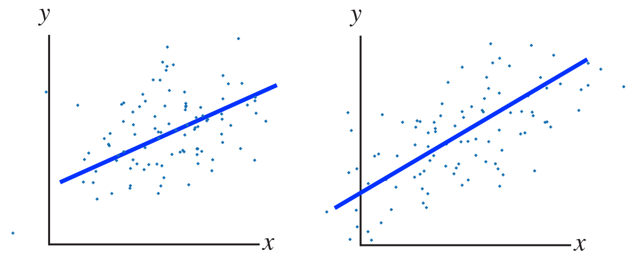
回帰直線からの
ばらつき

y のもともとの
ばらつき

決定係数 ＝回帰直線によるばらつきの
減少の度合い

＝回帰直線によって，
ばらつきの何%が説明できたか

「中くらいの相関」とは



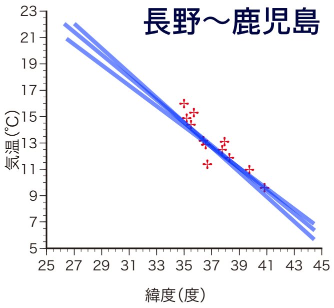
相関係数0.5
決定係数0.25

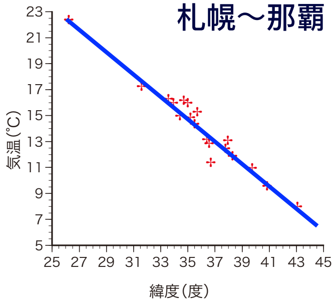
相関係数0.7
決定係数0.49

回帰直線では
ばらつきの25%
しか説明できない

こちらが，
中くらいの相関関係

前回の演習問題の例





決定係数0.712

平均付近に密集していると不安定

決定係数0.949

平均から離れた個体があると安定する

A. Asano, Kansai Univ.