

2020年度秋学期 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🤔

## 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? 🤔

## 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  
傾向があることを見いだす

緯度と気温の、どちらがどちらに影響しているかは考えない

### 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 $\circ\text{C}$ 下がる

## 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 $0^{\circ}\text{C}$ 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

※「決まっている」というのは、緯度によって気温が決まるメカニズムがあるという意味ではなく、緯度の違いによって気温の違いが推測できる、という意味

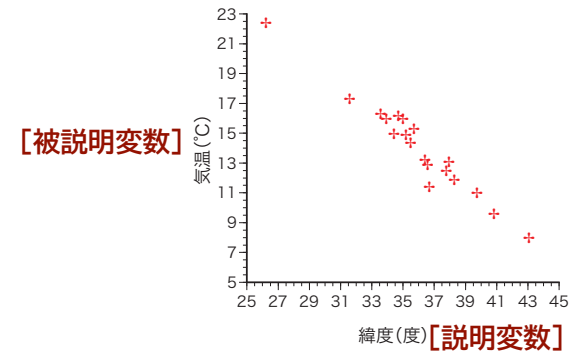
統計学では、

気温がばらついていることは、緯度によって[説明]されるという

そして、そのモデルでどの程度説明がつかかを考える

## 説明変数・被説明変数

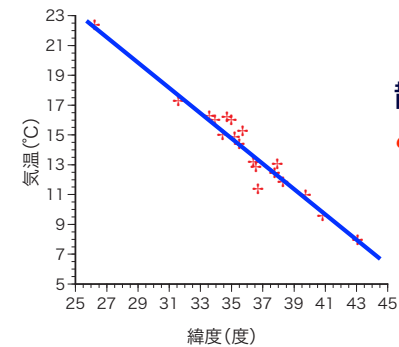
気温は緯度によって説明される(というモデル)



## 線形単回帰 🤔

## 線形単回帰

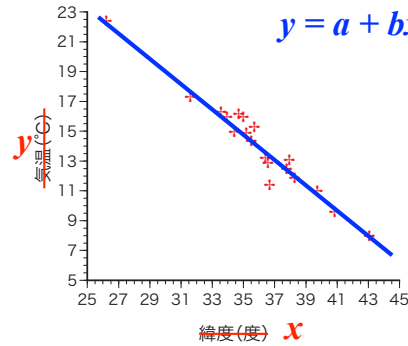
気温は緯度によって説明される どう説明される? どういうモデルか?



散布図上で直線の関係がある、  
というモデルを考える

## 線形単回帰

散布図上で直線の関係がある



【線形単回帰】という

$$y = a + bx ?$$

直線の式は  $y = ax + b$  と習ったような🤔

どちらも正解です

$$y = ax + b \quad \text{降冪(こうべき)順}$$

$$y = a + bx \quad \text{昇冪(しょうべき)順}$$

$$y = a + bx ?$$

降冪(こうべき)順は  $y = ax + b$  ただちに1次関数とわかる

何次関数かすぐわかる  $y = ax^2 + bx + c$  これは2次関数

昇冪(しょうべき)順は  
説明変数を付け加えて  
いくことができる  $y = a + bx + b_2x_2 + b_3x_3 + \dots$

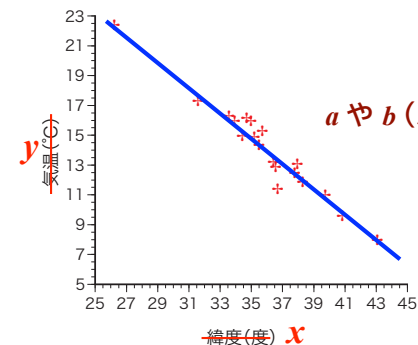
気温 緯度 標高 海からの距離

説明変数が2つ以上ある場合を重回帰という

統計では、昇冪順を使うことが多い

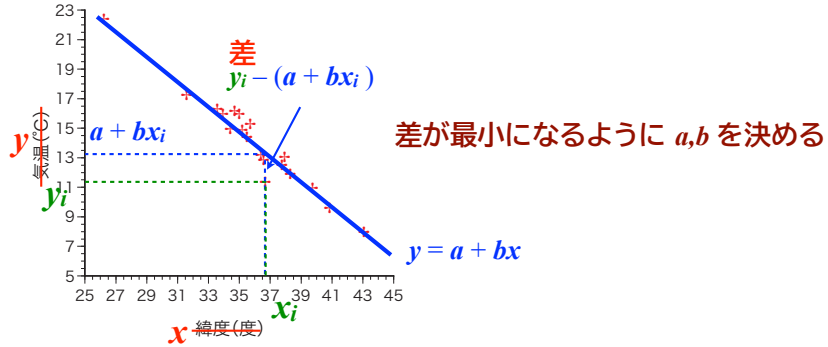
## 線形単回帰

$y = (a) + (b)x$  という式で表される関係



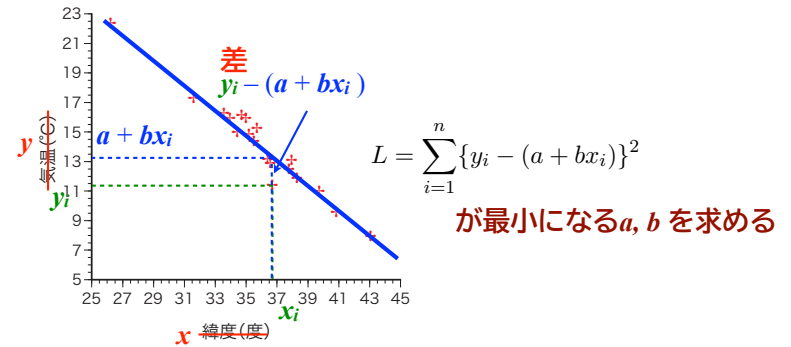
## パラメータの決定

$x = x_i$  のとき モデルによれば  $y = a + bx_i$  実際は  $y_i$



## 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



## 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)

付録に収録してある数式の展開は、試験の範囲には含みません。

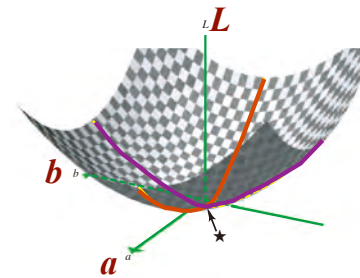
今から、「偏微分による方法」の考え方を説明します。

## 「偏微分」による方法

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

$a, b$  の2次関数

が最小になる  $a, b$  を求める

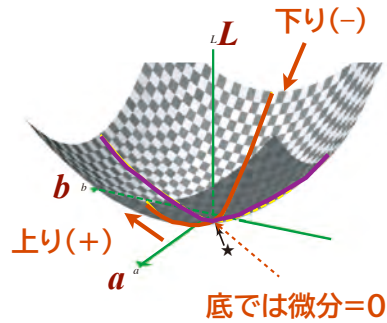


$a$  だけの関数と考えて微分

$b$  だけの関数と考えて微分

微分? 😓

## 微分?



$a$  だけの関数と考える微分

微分は、傾きを求める計算

底では微分=0

$b$  についても同じ、底では微分=0

底で  $L$  が最小だから、  
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)

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

$\sigma_{xy}$  ←  $x, y$  の共分散  
 $\sigma_x^2$  ←  $x$  の分散

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

$\bar{y}$  ←  $y$  の平均  
 $\bar{x}$  ←  $x$  の平均

## 最小二乗法

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

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

$$y = a + bx$$

[回帰係数]

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

を最小にしたので[最小二乗法]

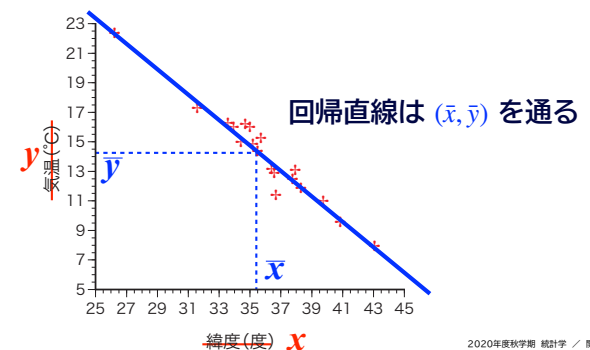
[回帰方程式]あるいは[回帰直線]

## ところで

$$y = a + bx$$

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

から  $y - \bar{y} = b(x - \bar{x})$



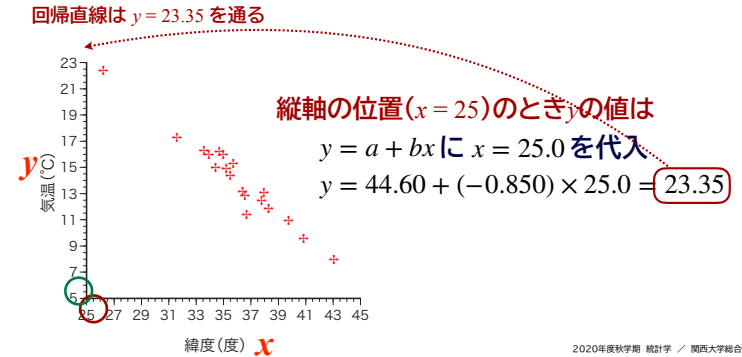
## 線形単回帰の結果を使う💡

## 緯度と気温(前回の講義)の例で

散布図上に回帰直線をひく

緯度を $x$ , 気温を $y$ として回帰直線  $y = a + bx$  を求めると

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



## 散布図上に回帰直線をひく

緯度を $x$ , 気温を $y$ として回帰直線  $y = a + bx$  を求めると

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

回帰直線は  $y = 23.35$  を通る

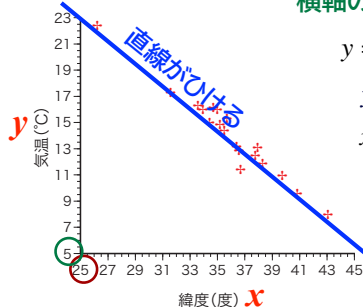
横軸の位置( $y = 5$ )のとき  $x$  の値は

$$y = a + bx \text{ より } x = \frac{y - a}{b}$$

$y = 5$  を代入すると

$$x = (5 - 44.60) / (-0.850) = 46.59$$

計算結果と図が合っていることを  
たしかめましょう



## 求めた回帰直線を使って

緯度35.0度の都市の気温は何°Cかを推定する

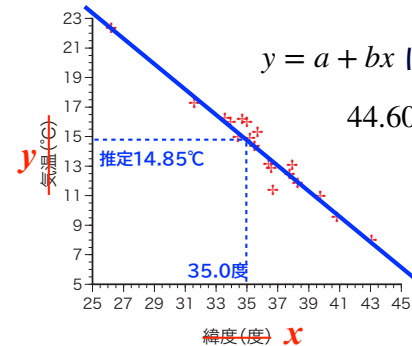
緯度を $x$ , 気温を $y$ として回帰直線  $y = a + bx$  を求めると

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

$y = a + bx$  に  $x = 35.0$  を代入すると

$$44.60 + (-0.850) \times 35.0 = 14.85 \text{ (°C)}$$

計算結果と図が合っている  
ことをたしかめましょう

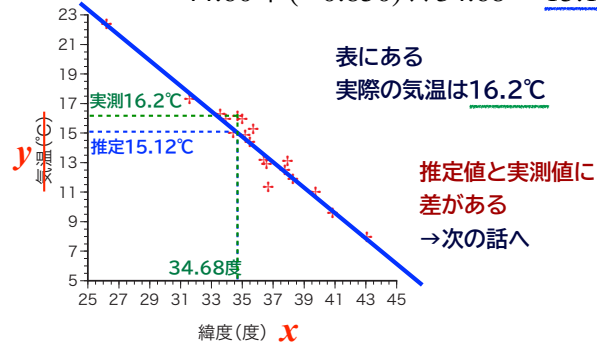


## 求めた回帰直線を使って

表の中にある大阪市(緯度34.68度)の気温を推定

$$y = a + bx \text{ に } x = 34.68 \text{ を代入}$$

$$44.60 + (-0.850) \times 34.68 = \underline{15.12} (^{\circ}\text{C})$$

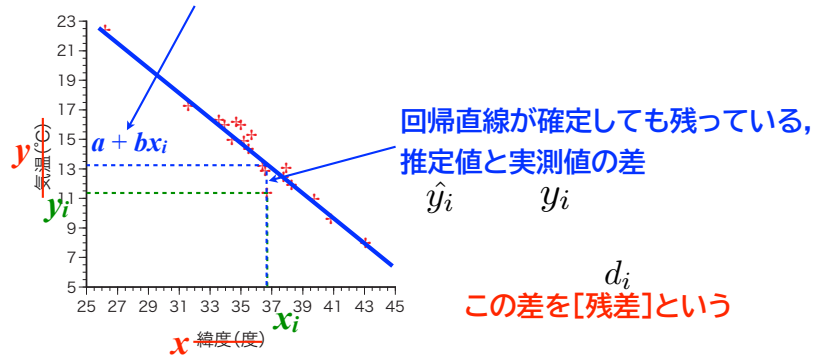


## 決定係数と「説明」🤔

## 残差

$a, b$  が求められて、回帰直線が確定したとき

$$x_i \text{ に対する, 回帰直線による } y \text{ の推定値 } \hat{y}_i = a + bx_i$$



## 残差と決定係数

残差は、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、次の関係がなりたつ(付録3)

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

残差      相関      相関係数の2乗  
係数      係数      [決定係数]



## 決定係数の意味

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

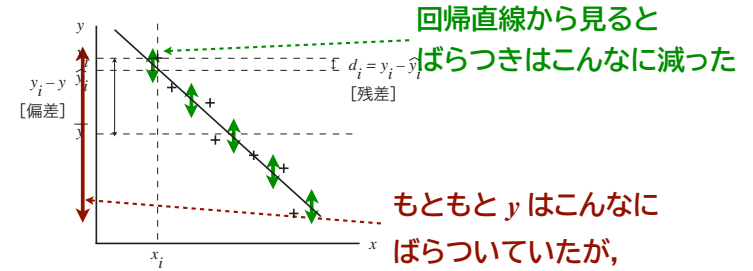
$$1 - r_{xy}^2 = \frac{\sum d_i^2 / n}{\sum (y_i - \bar{y})^2 / n}$$

残差の2乗の平均  
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

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

残差の2乗の平均  
決定係数 y の偏差の2乗の平均 (y の分散)



## 決定係数の意味と「説明」

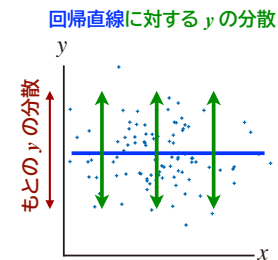
$$1 - r_{xy}^2 = \frac{\sum d_i^2 / n}{\sum (y_i - \bar{y})^2 / n}$$

回帰直線からのばらつき  
決定係数 y のもとのばらつき

決定係数 = 回帰直線によるばらつきの減少の割合  
 = 回帰直線によって、ばらつきの何%が「説明」できたか

## 決定係数の意味と「説明」

相関係数 = 0, すなわち 決定係数 = 0 のとき

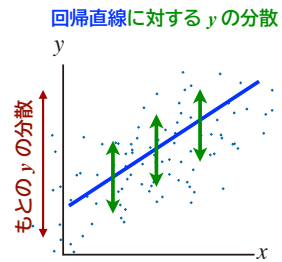


回帰直線に対する y の分散は  
 もとの y の分散とまったく変わらない  
 「回帰直線のまわりに散らばっている」と説明したところで、  
 全く説明になっていない



## 決定係数の意味と「説明」

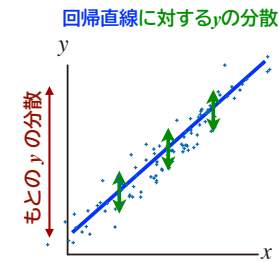
相関係数 = 0.7 すなわち 決定係数  $\doteq$  0.5 のとき



回帰直線に対する  $y$  の分散は  
もとの  $y$  の分散 に比べて半分になっている  
「回帰直線のまわりに散らばっている」と  
説明したことで、  
もとの  $y$  の分散の半分を説明した

## 決定係数の意味と「説明」

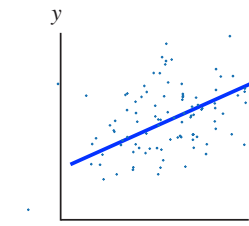
相関係数 = 0.9 すなわち 決定係数  $\doteq$  0.8 のとき



回帰直線に対する  $y$  の分散は  
もとの  $y$  の分散 に比べて20%に減っている  
「回帰直線のまわりに散らばっている」と  
説明したことで、  
もとの  $y$  の分散の80%を説明した

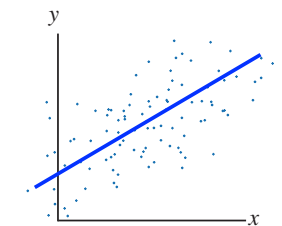
ところで、前回の講義で  
言いかけていたことですが💬💧

## 「中くらいの相関」とは



相関係数0.5  
決定係数0.25

回帰直線ではもとの  $y$  の分散の  
25%しか説明できていない

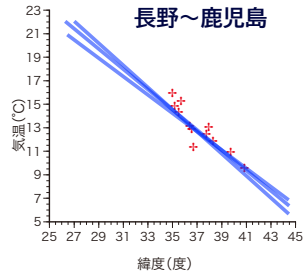


相関係数0.7  
決定係数0.49

回帰直線でもとの  $y$  の分散の  
50%を説明している

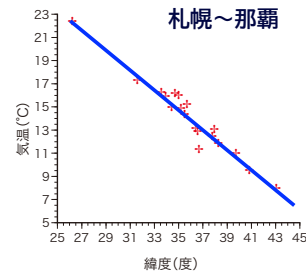
こちらのほうが、中くらいの相関関係

## 緯度と気温の例で



決定係数0.712

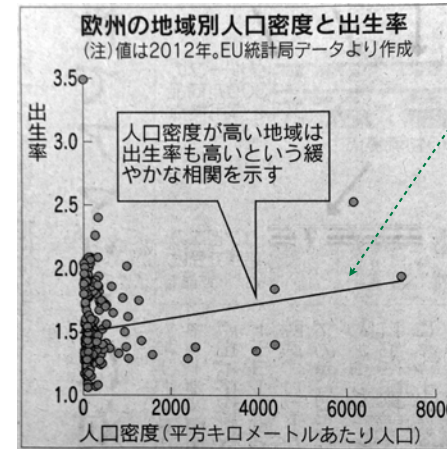
平均付近に密集して  
いると不安定



決定係数0.949

平均から離れた個体がある  
と安定する

## 注意すべき例



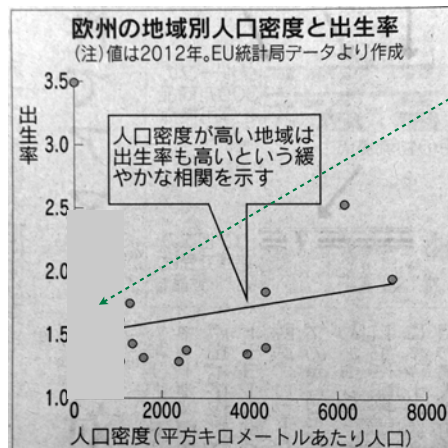
回帰直線をひいているけれど…

回帰係数は、今回説明した式で計算  
すればそれなりに求まるけれど、

そもそも、こんなの  
「人口密度と出生率に相関がある」  
といえるの？

(2014年10月29日 日経新聞)

## 注意すべき例



ここを隠すと相関があるように見える  
([選択バイアス]という)

回帰直線は  
どんなデータであっても計算  
「できてしまう」ことに注意

得られた回帰直線は、  
それが意味のあるものかどうか、  
よく考えましょう。

(2014年10月29日 日経新聞)