

2024年度秋学期

統計学

第7回

データの関係を知る(2)

— 回帰と決定係数



関西大学総合情報学部
浅野 晃

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○℃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°C 下がる

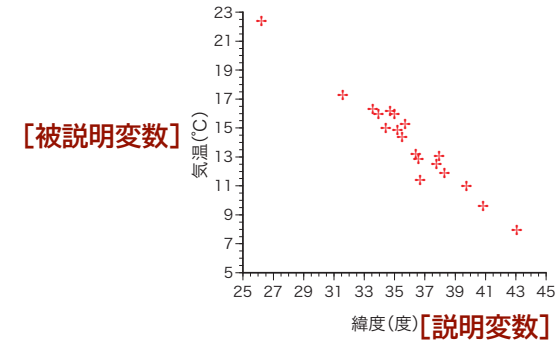
各都市の気温の違いは、緯度によって決まっているという[モデル]を考える

- ※「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
- ※前回の「(学年を無視すれば)成績が体格によって決まっている」というのも、モデルとしてはあり。(学年を無視することが妥当かどうかは別)

統計学では、気温の分散は、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くかを考える

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)

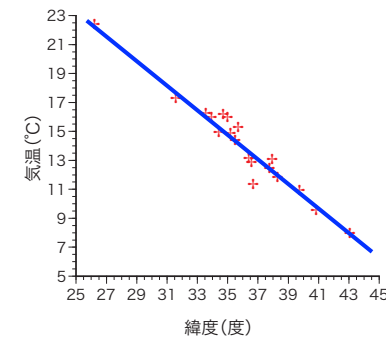


線形単回帰 🤔

線形単回帰

気温の分散は緯度によって説明される

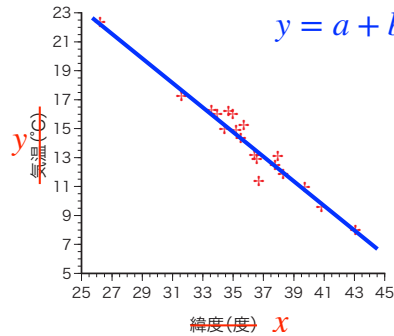
どう説明される? どういうモデルか?



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



$y = a + bx$ という式で表される関係

【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

$y = ax + b$ 降幕(こうべき)順

$y = a + bx$ 昇幕(しょうべき)順

$y = a + bx$?

降幕(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

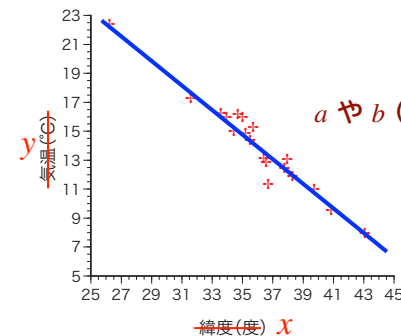
昇幕(しょうべき)順は
説明変数を付け加えて $y = ax + b_2x_2 + b_3x_3 + \dots$
いくことができる 気温 緯度 標高 海からの距離 ...

説明変数が2つ以上ある場合を[重回帰]という

統計学では、昇幕順を使うことが多い

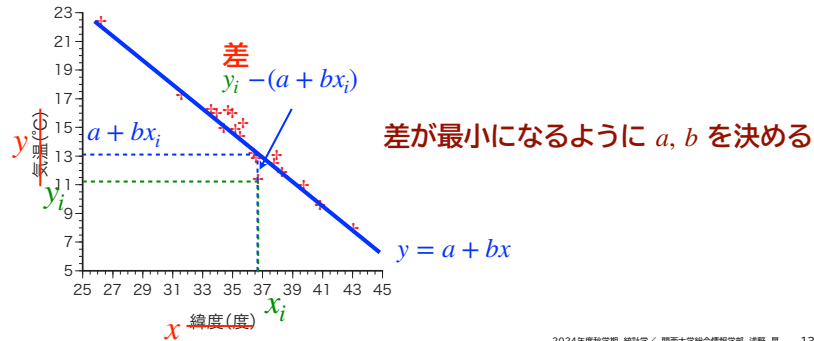
線形単回帰

$y = \textcircled{a} + \textcircled{b}x$ という式で表される関係



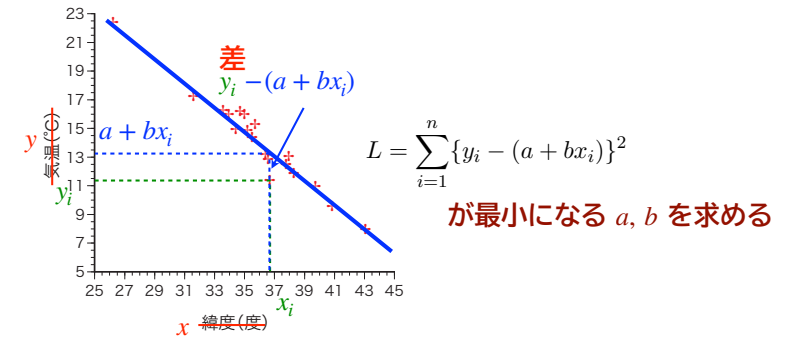
パラメータの決定

$x = x_i$ 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
の2乗



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。

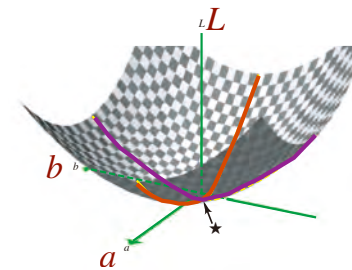
今から、「偏微分による方法」の考え方
(数式そのものではなくて考え方)を説明します。

「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

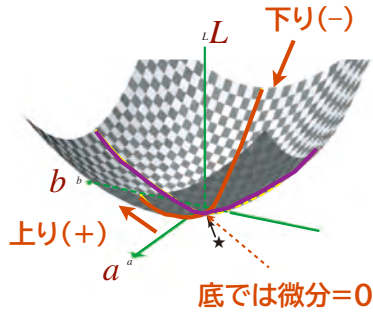


a だけの関数と考えて微分

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

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)

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

σ_{xy} ← x, y の共分散
 σ_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]

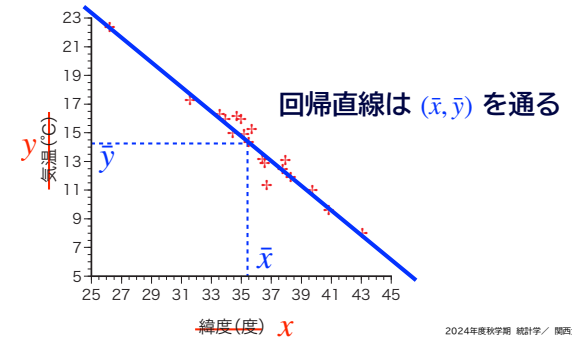
[回帰方程式]あるいは[回帰直線]

ところで

$$y = a + bx$$

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

から $y - \bar{y} = b(x - \bar{x})$



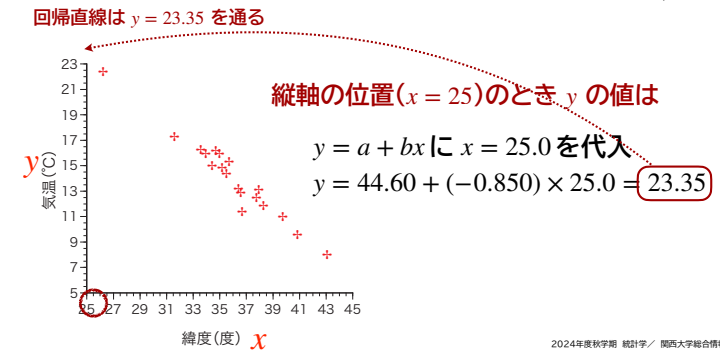
線形単回帰の結果を使う💡

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

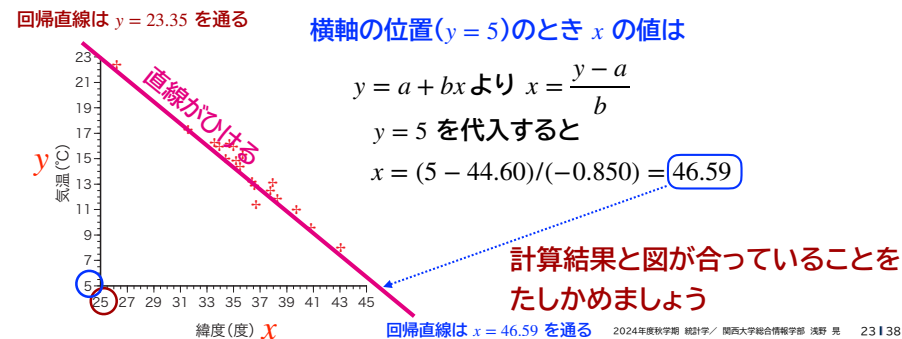
$$\rightarrow b = -0.850, a = 44.60$$



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

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

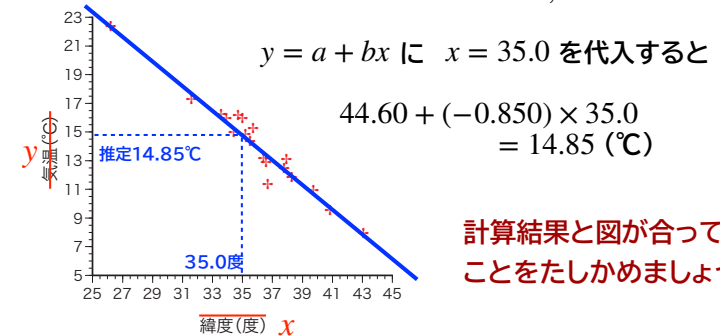


求めた回帰直線を使って

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

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



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

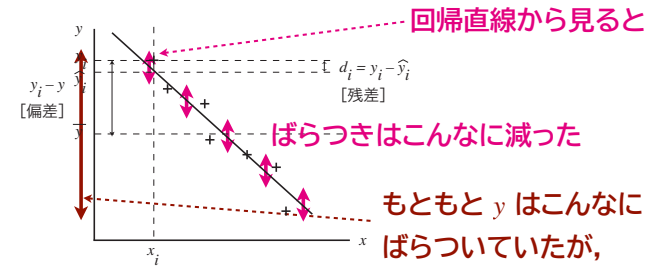
y の偏差の2乗の平均 = y の分散

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」

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

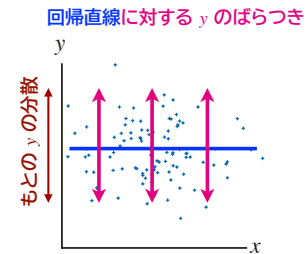
回帰直線からのばらつき
決定係数

y のもともとのばらつき

決定係数 = 回帰直線によるばらつきの縮小の割合
 = 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」

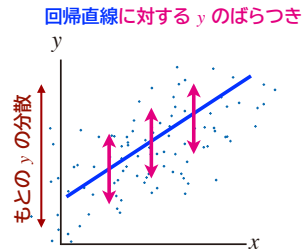
相関係数 = 0, すなわち 決定係数 = 0 のとき



回帰直線に対する y のばらつきは
 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」

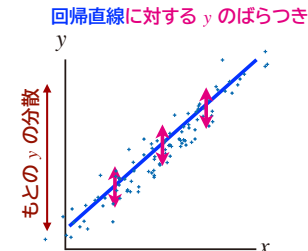
相関係数 = 0.7 すなわち 決定係数 \doteq 0.5 のとき



回帰直線に対する y のばらつきは
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」

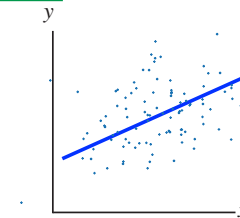
相関係数 = 0.9 すなわち 決定係数 \doteq 0.8 のとき



回帰直線に対する y のばらつきは
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

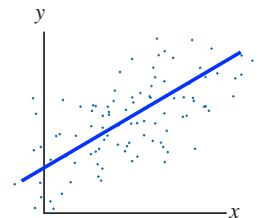
ところで、前回の講義で
言いかけていたことですが💬💧

「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

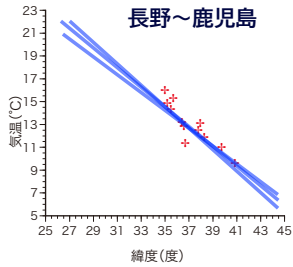


相関係数0.7
決定係数0.49

回帰直線でもとの y の分散の
50%を説明している

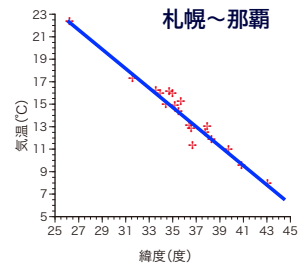
こちらのほうが、中くらいの相関関係
(分散の説明という意味では)

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

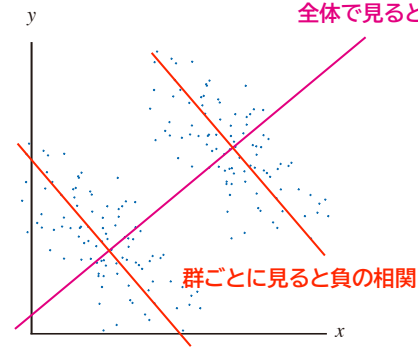


決定係数0.949

平均から離れた個体がある
と安定する

注意すべき例

こういう分布だと



相関係数や回帰直線は
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。