2016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

A.Asano, Kansai Univ.

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
緯度があがると，気温が下がる
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度あがると，気温が◯℃下がる

A.Asano, Kansai Univ.

回帰分析とは

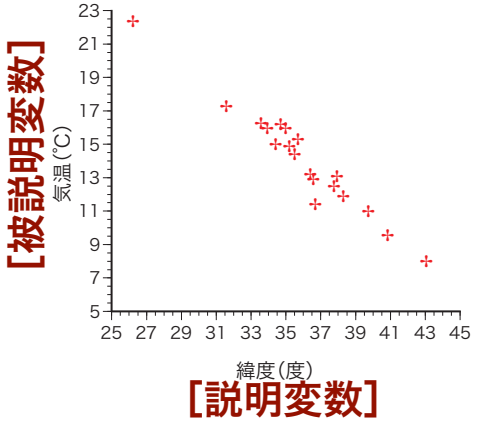
緯度が上がるから気温が下がると考える
緯度が１度あがると，気温が◯℃下がる

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）

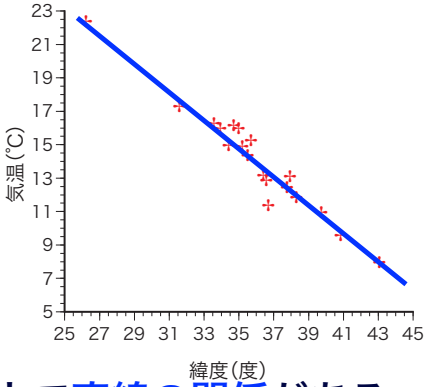


［説明変数］

線形単回帰

線形単回帰

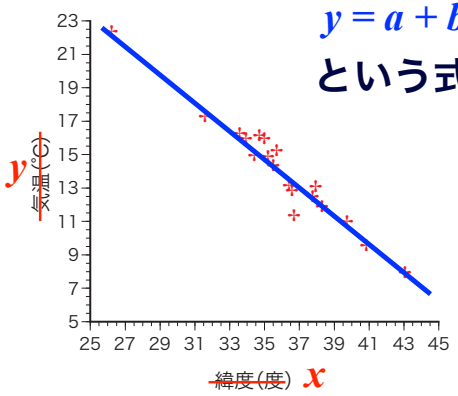
気温は緯度によって説明される
どう説明される？



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

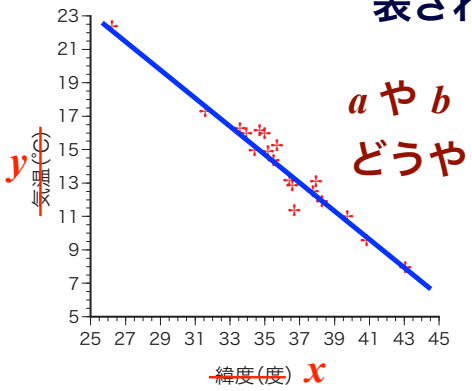
［線形単回帰］

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

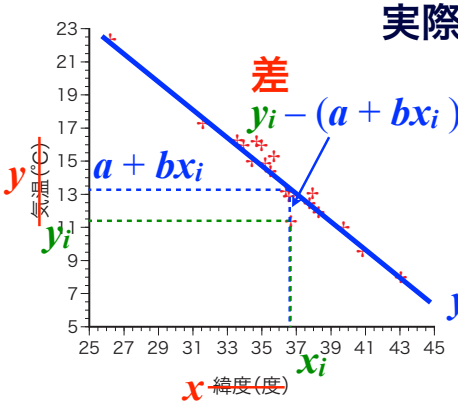
どうやって求める？

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



$y = a + bx$

差が最小に

なるように

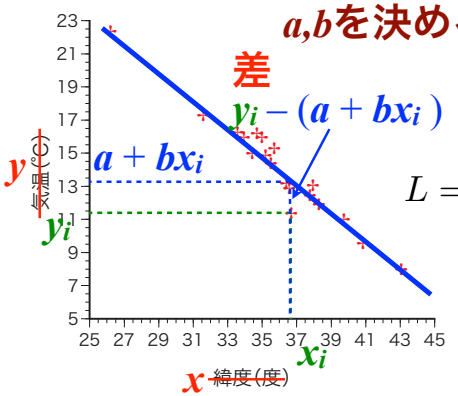
a,bを決める

パラメータの決定

すべてのx_iについて，

差^{2}の合計が最小になるように

a,bを決める



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

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

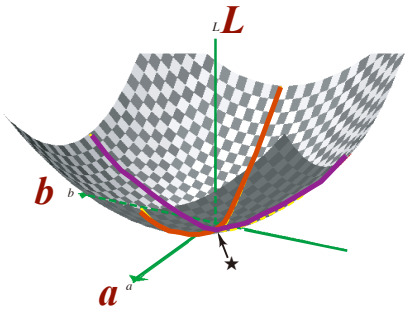
- 偏微分による方法（付録１）
- 「２次関数の最大・最小」に
  よる方法（付録２）

「偏微分」による方法

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

a,bの２次関数

が最小になる
a,bを求める

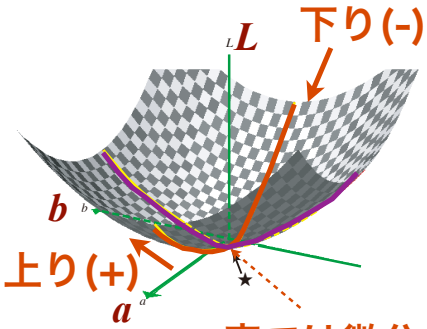


aだけの関数
と考えて微分

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

底では微分=0

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

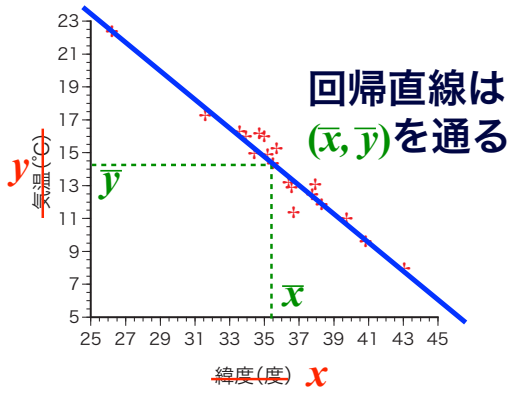
［回帰直線］

［回帰係数］

ところで

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

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



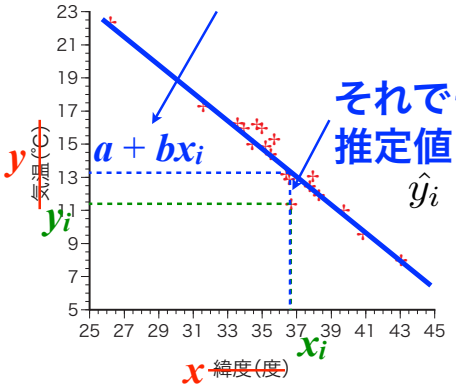
決定係数

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

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



［残差］という

d_i

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

決定係数が１に近づくほど
残差の２乗和が0に近づく

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

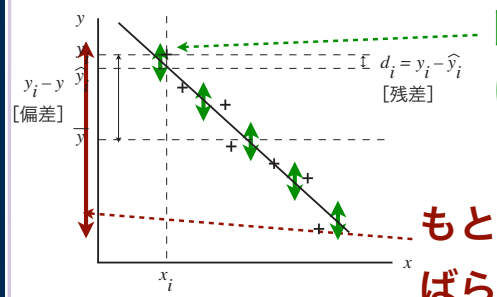
yの偏差の２乗の平均
＝yの分散

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）



回帰直線からの
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

決定係数の意味

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

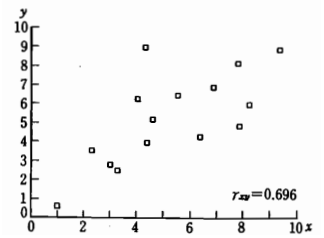
決定係数

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

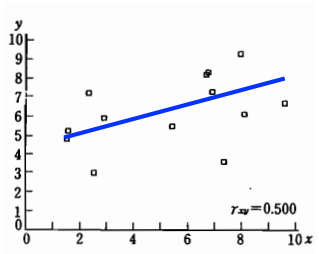
「中くらいの相関」とは

「統計学入門」(東京大学出版会)より



相関係数0.7
決定係数0.49

こちらが中くらいの相関関係

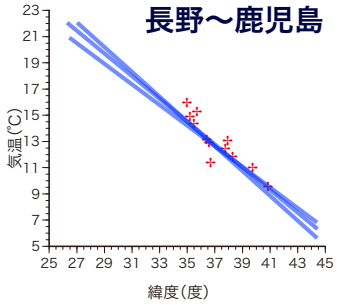


相関係数0.5
決定係数0.25

回帰直線ではばらつきの25%しか説明できない

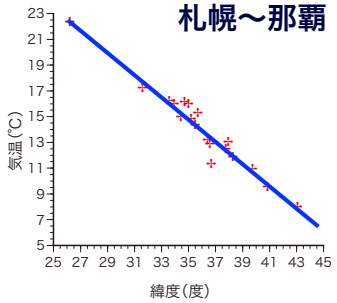
A.Asano, Kansai Univ.

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

平均から離れたデータがあると安定する

A.Asano, Kansai Univ.