統計学 01

早稲田大学政治経済学部

第18回

西郷 浩

本日の目標

- 回帰分析とは
  - 二次元データ
  - データ発生の仕組みとしての回帰モデル
- 回帰モデルの推定
  - 回帰係数の推定方法（最小二乗法）
  - 残差

二次元のデータ(1)

- 教科書 表13.1（p. 258）
  - (X_i, Y_i)
    = (第 i -1日の福岡の気圧, 第 i 日の東京の気圧）
  - 散布図を見ると右上がり。
    - 前日の福岡の気圧が高い（低い）と東京の気圧も高い（低い）。
    - ただし、その関係は関数（X が決まればY が一意的に決まる）といえるほど強くはない。
      - X が同じような値であっても Y の値にはバラツキがある。
    - 次の2つ現象を同時に扱えるような仕組み（モデル）は？
      - Xの値が増えると、平均的にYの値が増える。
      - Xの値が同程度でも、Yの値にはバラツキがある。

二次元のデータ(2)

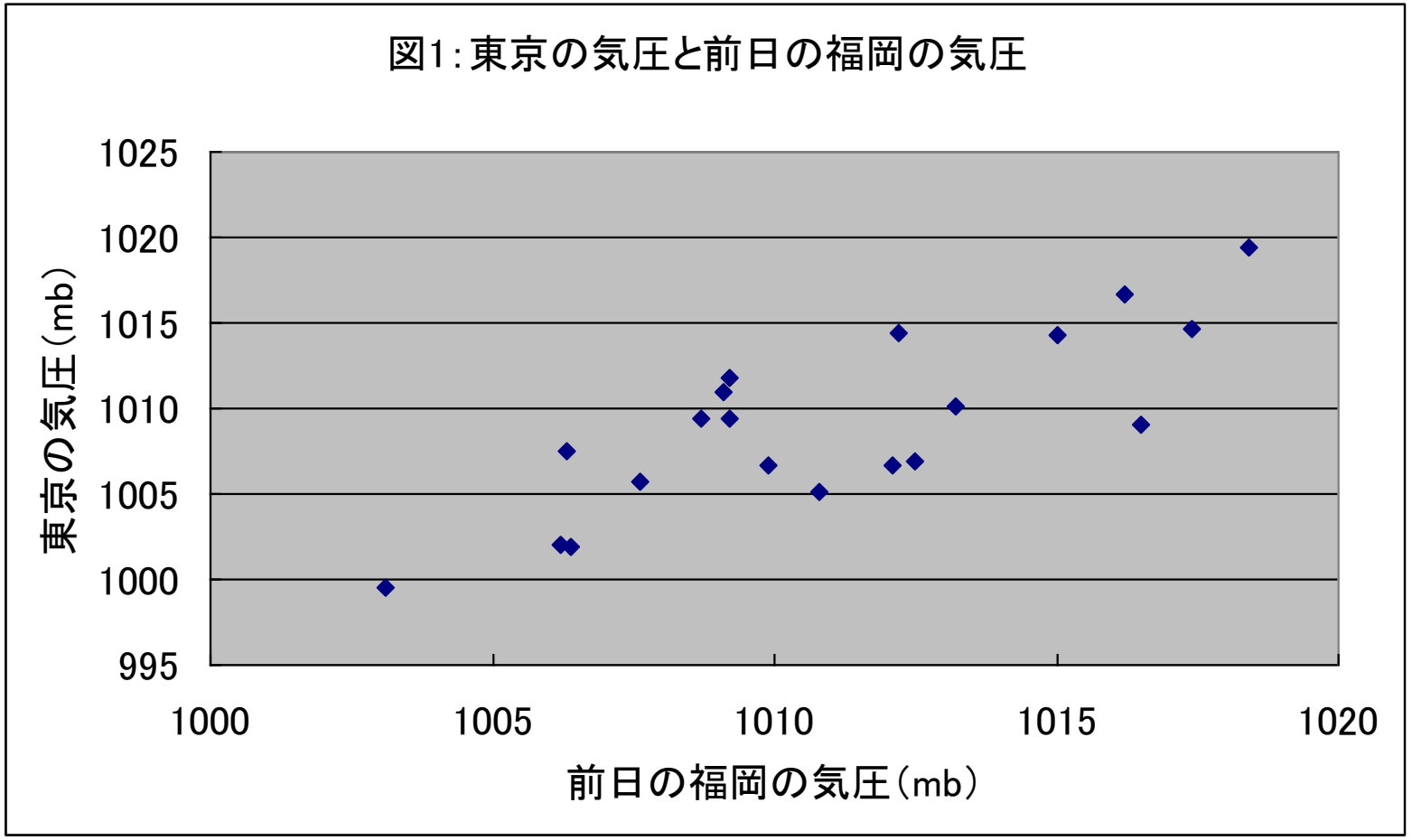
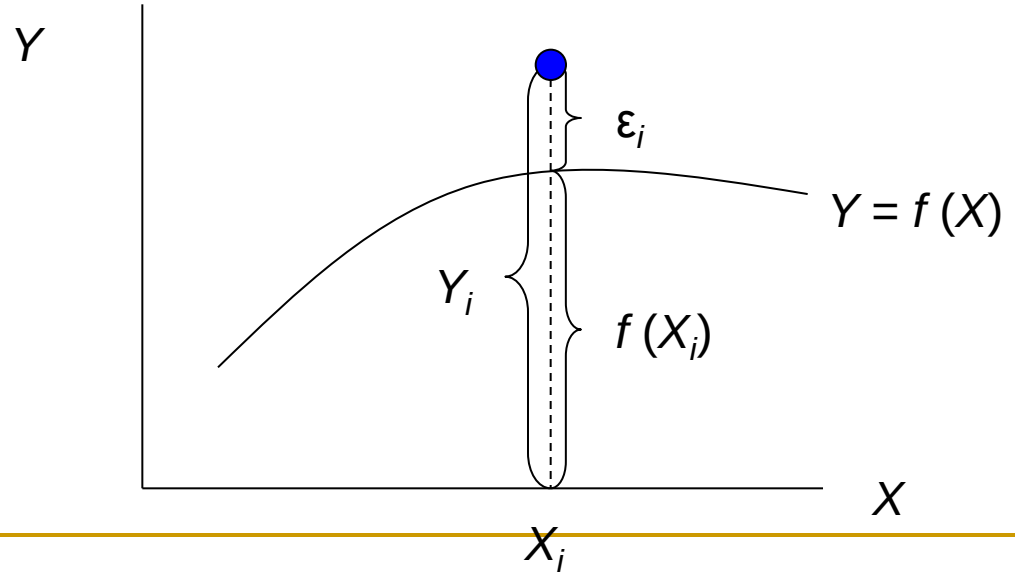


図1：東京の気圧と前日の福岡の気圧
東京の気圧（mb）
1025
1020
1015
1010
1005
1000
995
1000
1005
1010
1015
1020
前日の福岡の気圧（mb）

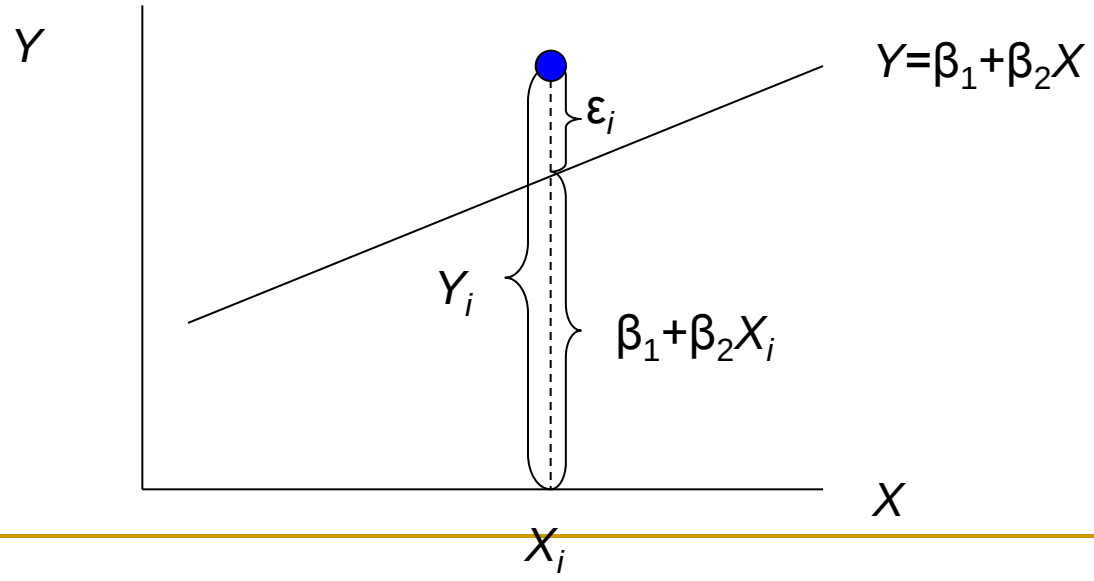
データ発生の仕組みとしての回帰モデル(1)

- 2つの現象を以下のように捉えることにする。
  - 回帰モデル：$Y_i = f(X_i) + \varepsilon_i$
    - $f(X_i)$：Y と X との平均的な関係をあらわす部分（回帰関数）
    - ε_i ：平均的関係からの乖離を表す偶然的変動（誤差項）



データ発生の仕組みとしての回帰モデル(2)

- とくに $f(X) = \beta_1+\beta_2 X$ とした場合
  - 線形回帰モデル：$Y_i = \beta_1+\beta_2 X_i + \varepsilon_i$
    - $Y_i = \beta_1+\beta_2 X_i$ ：Y と X との平均的な関係（回帰直線）
    - ε_i ：平均的関係からの乖離を表す偶然的変動（誤差項）



データ発生の仕組みとしての回帰モデル(3)

- $f(X)$ として直線が多用される理由
  - 数学的な扱いが容易である。
  - たとえ曲線的な関係であっても、変数変換を援用して直線的な関係に置き換えられる場合が多い。
    - $Y'_i = g(Y_i)$, $X'_i = f(X_i)$ として、$Y'_i = \beta_1 + \beta_2 X'_i + \varepsilon_i$
- 用語の整理
  - $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$

被説明変数　母回帰係数　説明変数　誤差項

データ発生の仕組みとしての回帰モデル(4)

- モデルの偶然的(確率的)変動 ← 誤差項
  - 今の事例においては、福岡の気圧は前日に決まっており、当日の東京の気圧を考えるときには所与(確定値)と考えることができる。
  - データにふくまれる偶然的な変動がすべて誤差項からもたらされているのだから、誤差項(確率変数)の性質がわからなければ、確率モデルとしての回帰モデルの性質を明らかにできない。
  - 誤差項の備えるべき性質は？

データ発生の仕組みとしての回帰モデル(5)

- 誤差項 ε_i の備えるべき性質
  - 期待値 0：$E(\varepsilon_i)=0$
    - Y_i は $f(X_i)$ よりも大きかったり小さかったりするけれども、Y_i の平均的な値は $f(X_i)$ に等しくなる。
      - 「$Y=f(X)$ が Y_i と X_i との平均的な関係を表す」のであれば、誤差項の平均的な値が0 となるように $f(X_i)$ を選ぶべきだ。
  - 分散一定：$V(\varepsilon_i)=\sigma^2$
    - 回帰関数 $Y=f(X)$ のまわりの（縦軸方向で測った）バラツキが X の水準によらず一定である。
  - 誤差同士の無相関性：$Cov(\varepsilon_i, \varepsilon_j)=0$
    - 特定の誤差同士が関係をもたない。

回帰係数の推定(1)

- 問題の所在
  - 設定：回帰モデルにしたがってデータは発生している。
    - $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$
  - しかし、われわれに観察できるのは発生したデータ（散布図に描き出された点）だけである。
    - 回帰係数 β_1, β_2 や誤差項のバラツキ σ^2 は未知である。
      - 「観察点の背後に直線 $Y = \beta_1 + \beta_2 X$ があり、偶然的な変動のためにその平均的な関係から若干ずれてデータが発生している」ということは知っているけれども、直線がどこにあるか（回帰係数 β_1, β_2 の値）や偶然変動の程度（誤差項の分散σ^2 ）は未知である。
    - シミュレーション

回帰係数の推定(2)

- 推定の方針
  - データ (X_i, Y_i) に、データ発生の仕組み（回帰モデル）が反映されている。
  - 誤差項の平均的な値（$E(\varepsilon_i)$）が0であり、誤差項の分散が有限（大きな誤差は出にくい）のであるから、回帰直線 $Y = \beta_1 + \beta_2 X$ はデータの中心部を通っていると期待できる。
  - したがって、データ全体と直線とのズレが最小になるように直線の位置（回帰係数β_1, β_2）を推定するのが賢明である。

回帰係数の推定(3)

最小二乗推定量 $\hat{\beta}_1, \hat{\beta}_2$: 以下の最小化問題の解

$$\min_{\beta_1,\beta_2} \sum_{i=1}^{n} \left(Y_i - \beta_1 - \beta_2 X_i\right)^2$$

その解は下の連立方程式（正規方程式）の解と同等。

$$\begin{cases} n\hat{\beta}_1 + \left(\sum_i X_i\right)\hat{\beta}_2 = \left(\sum_i Y_i\right) \\ \left(\sum_i X_i\right)\hat{\beta}_1 + \left(\sum_i X_i^2\right)\hat{\beta}_2 = \left(\sum_i X_i Y_i\right) \end{cases}$$

さらに計算を進めると

$$\hat{\beta}_2 = \sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \Big/ \sum_i \left(X_i - \overline{X}\right)^2$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$$

回帰係数の推定(4)

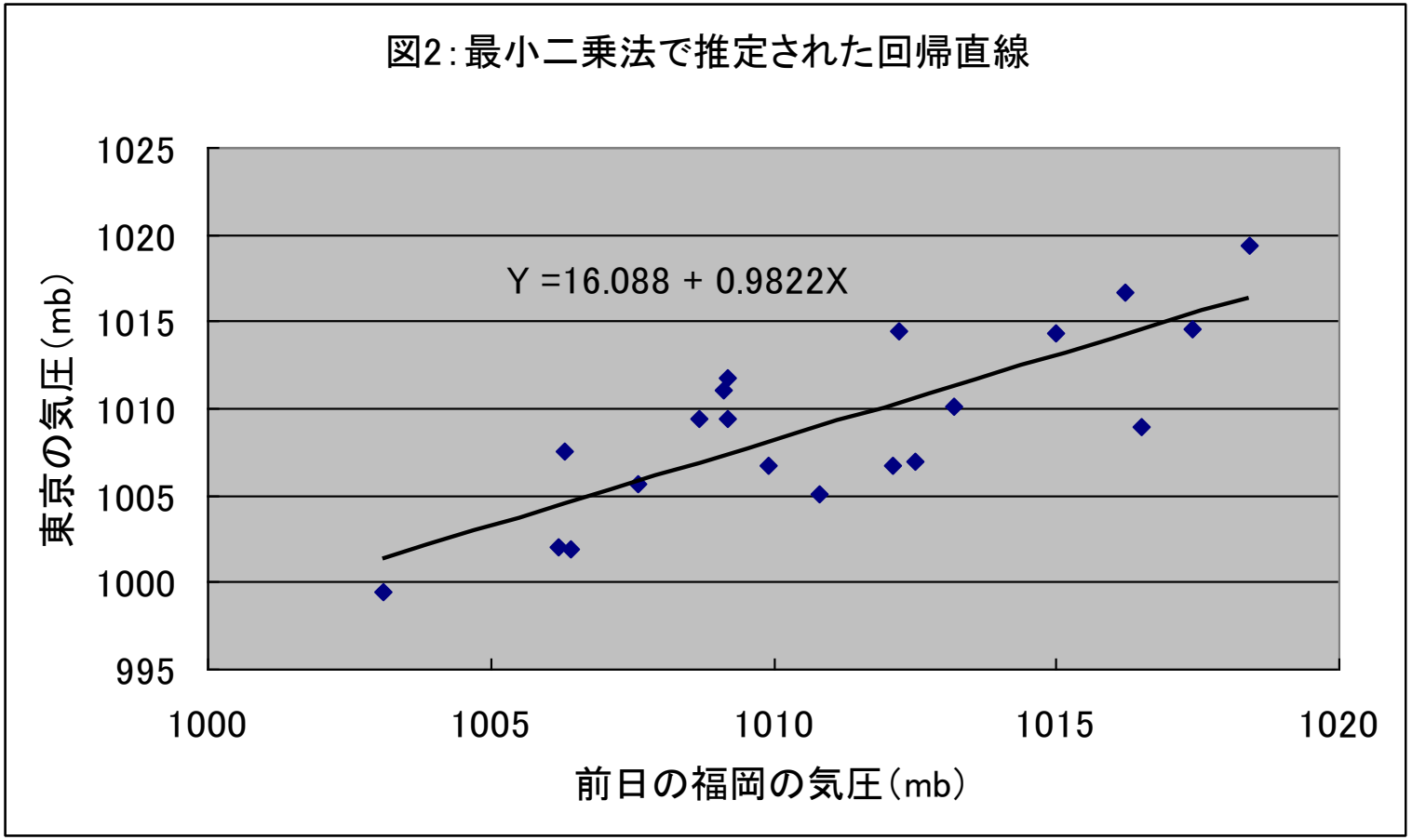


図2：最小二乗法で推定された回帰直線
Y =16.088 + 0.9822X
東京の気圧（mb）
前日の福岡の気圧（mb）
1025
1020
1015
1010
1005
1000
995
1000
1005
1010
1015
1020

回帰係数の推定(5)

$$\begin{cases} \text{母回帰係数：} \beta_1, \beta_2 \\ \text{最小二乗推定量：} \hat{\beta}_1, \hat{\beta}_2 \end{cases}$$

両者の関係は？

- シミュレーション
  - 推定された回帰直線（標本回帰直線）は、母回帰直線の近辺に出現するように見える。
  - その出方は誤差項によってもたらされる偶然変動に影響を受ける。

誤差項に関する推定(1)

$$\begin{cases} \text{誤差項：} \varepsilon_i = Y_i - \beta_1 - \beta_2 X_i \\ \text{推定された誤差項（残差）：} \hat{\varepsilon}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \end{cases}$$

残差の性質：

$$\begin{cases} \sum_i \hat{\varepsilon}_i = 0 \quad \left(\Leftrightarrow \bar{\hat{\varepsilon}} = \frac{1}{n} \sum_i \hat{\varepsilon}_i = 0 \right) \\ \sum_i \hat{\varepsilon}_i X_i = 0 \end{cases}$$

誤差項の分散σ^2の(不偏)推定量：

$$s^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2} \quad \left(E\left(s^2\right) = \sigma^2 \text{という性質がある。} \right)$$

誤差項に関する推定(2)

- 気圧データに関して計算すれば、
  - $s^2 = 9.911$
- 推定された回帰直線と母回帰直線とのズレ
  ← 誤差項によってもたらされる。
  - そのズレは、回帰係数の最小二乗推定量と(真の)回帰係数とのズレに対応している。
    - つまり、最小二乗推定量が確率変数である。
  - したがって、回帰係数の最小二乗推定量の分散(バラツキ)は、誤差項の分散σ^2と関係があるはずである。

推定された回帰係数の分散

傾きの最小二乗推定量の分散

$$V\left(\hat{\beta}_2\right)=\frac{\sigma^2}{\sum_i\left(X_i-\overline{X}\right)^2}$$

この式からわかること：

1. 誤差項の分散σ^2が小さいほど、
2. $\sum_i\left(X_i-\overline{X}\right)^2$つまりXのバラツキが大きいほど、

傾きが精確に推定できる。