回帰分析

回帰分析

- 回帰分析（regression analysis）
  - 従属変数（目的変数）と独立変数（説明変数）の間に関数式をあてはめ，従属変数と説明変数の関係を定量的に分析すること.
- （単）回帰分析
  - 独立変数（説明変数）が一つだけ.
- 重回帰分析
  - 独立変数（説明変数）が2つ以上.

線形回帰分析

- 従属変数yが独立変数$x_1, x_2, \cdots$の線形関数（一次関数）で与えられる.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots$$

- 多項式近似の場合は

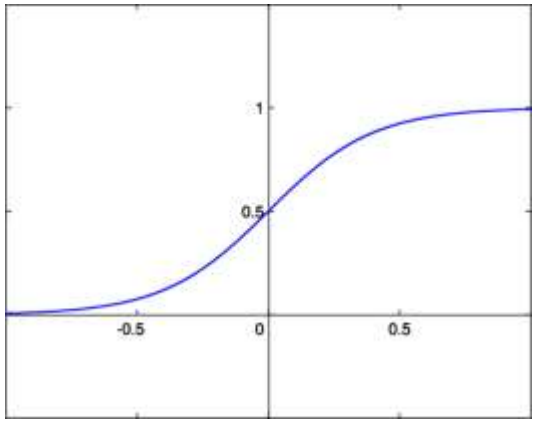
$$y = a_0 + a_1 x + a_2 x^2 + \cdots$$

- ただし，$x_1 = x, x_2 = x^2, \cdots$

非線形回帰分析

- ロジスティック関数（シグモイド関数）

$$\log\left(\frac{y}{1-y}\right) = x \quad \Rightarrow \quad y = \frac{1}{1+e^{-x}}$$

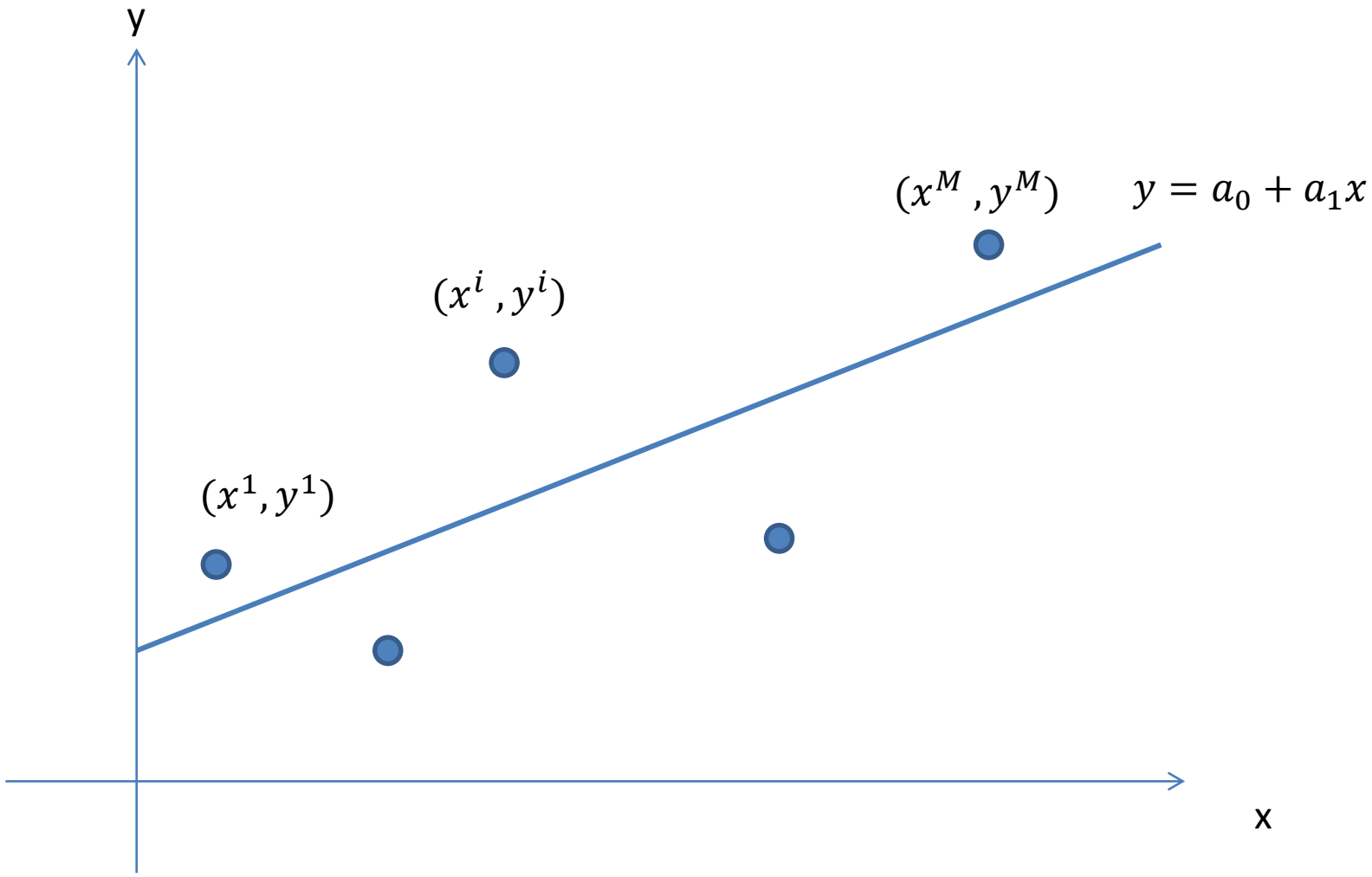


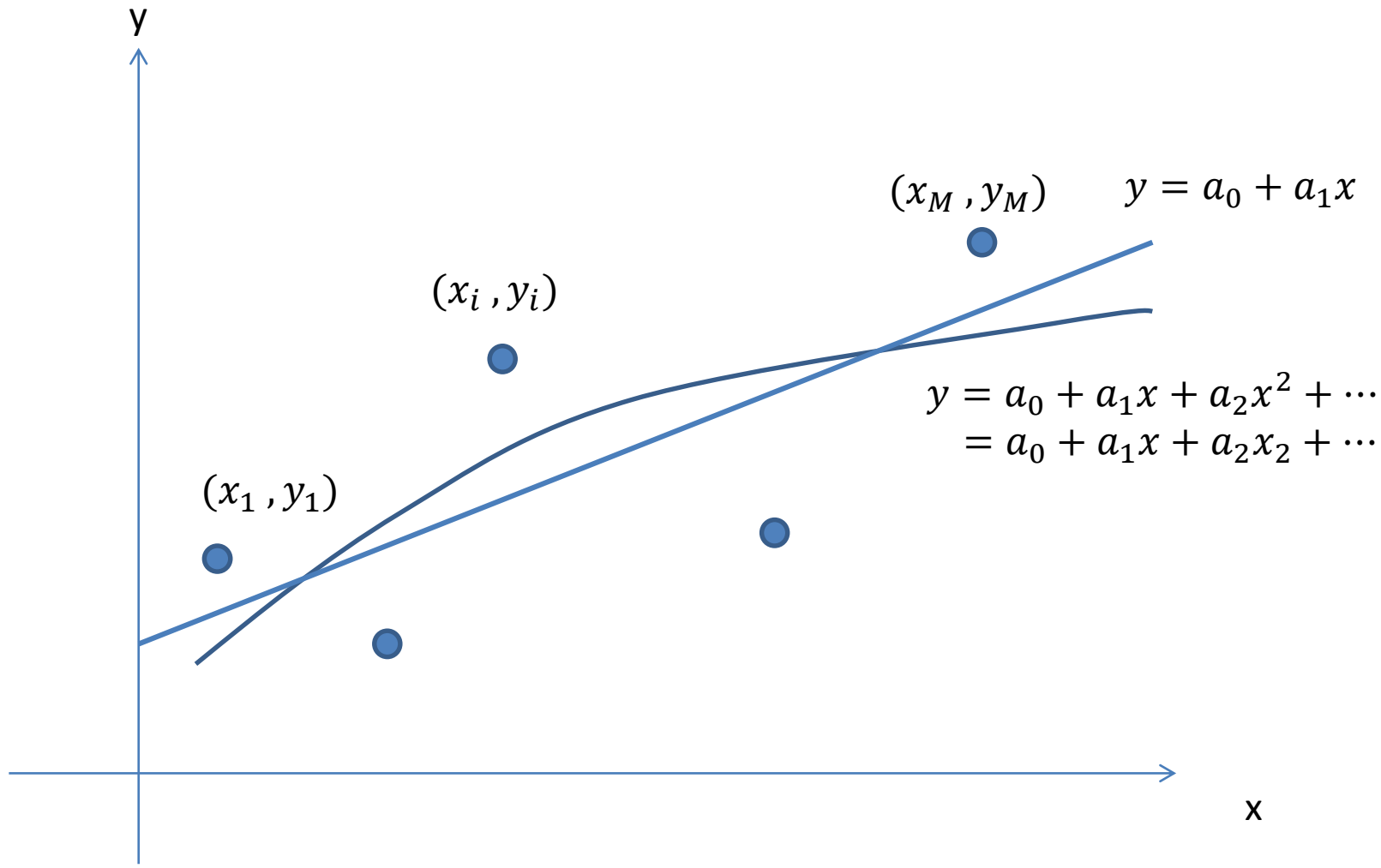
- ロジスティック回帰モデル

$$\log\left(\frac{y}{1-y}\right) = a_0 + a_1x_1 + a_2x_2 + \cdots$$

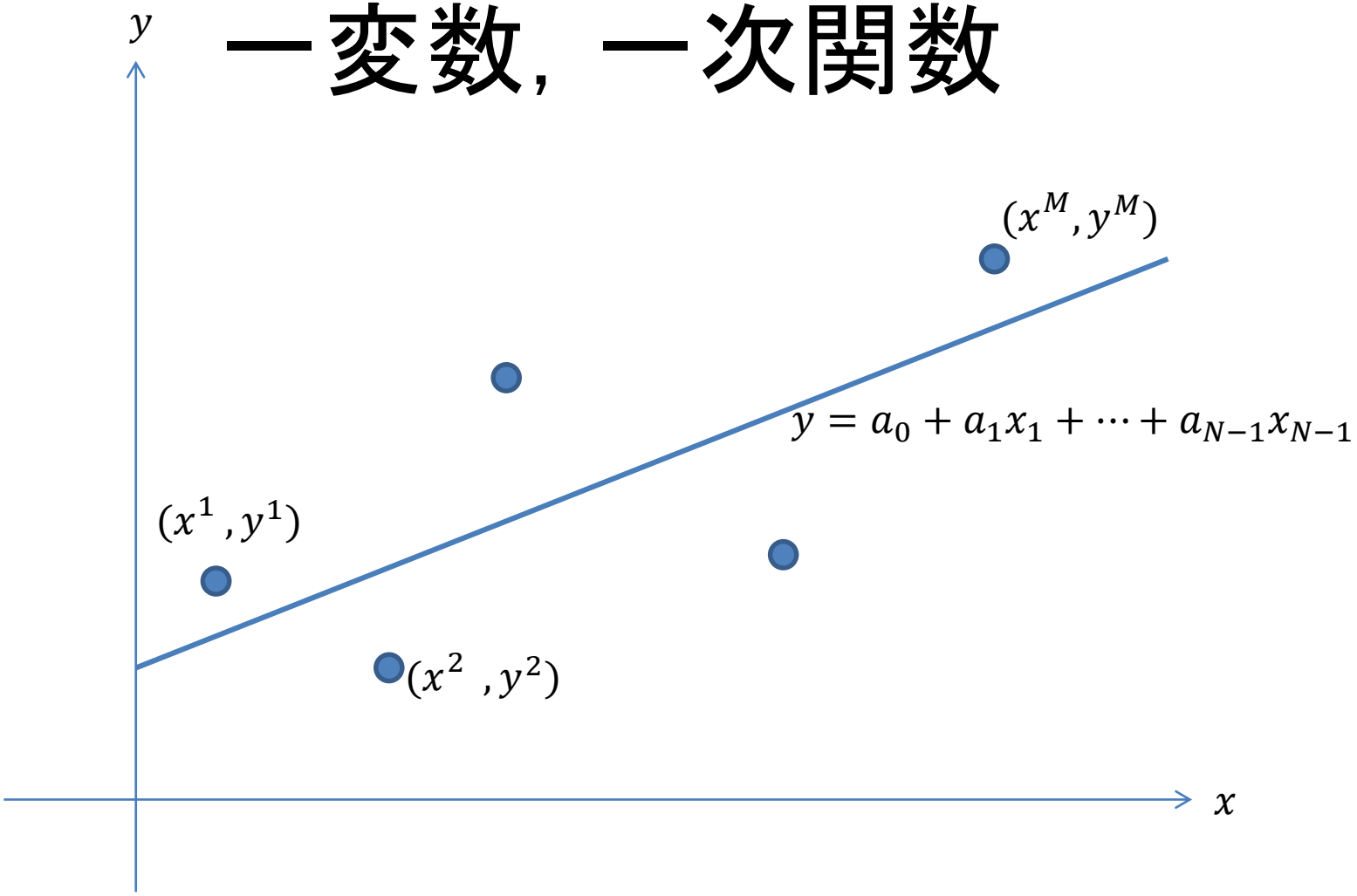
$$y = \frac{1}{1+e^{-(a_0+a_1x_1+a_2x_2+\cdots)}}$$

最小二乗近似


y
(x^M, y^M)
$y = a_0 + a_1 x$
(x^i, y^i)
(x^1, y^1)
x


y
(x_M , y_M)
$y = a_0 + a_1 x$
(x_i , y_i)
$y = a_0 + a_1 x + a_2 x^2 + \cdots$
$= a_0 + a_1 x + a_2 x_2 + \cdots$
(x_1 , y_1)
x

一変数，一次関数



パラメータ決定

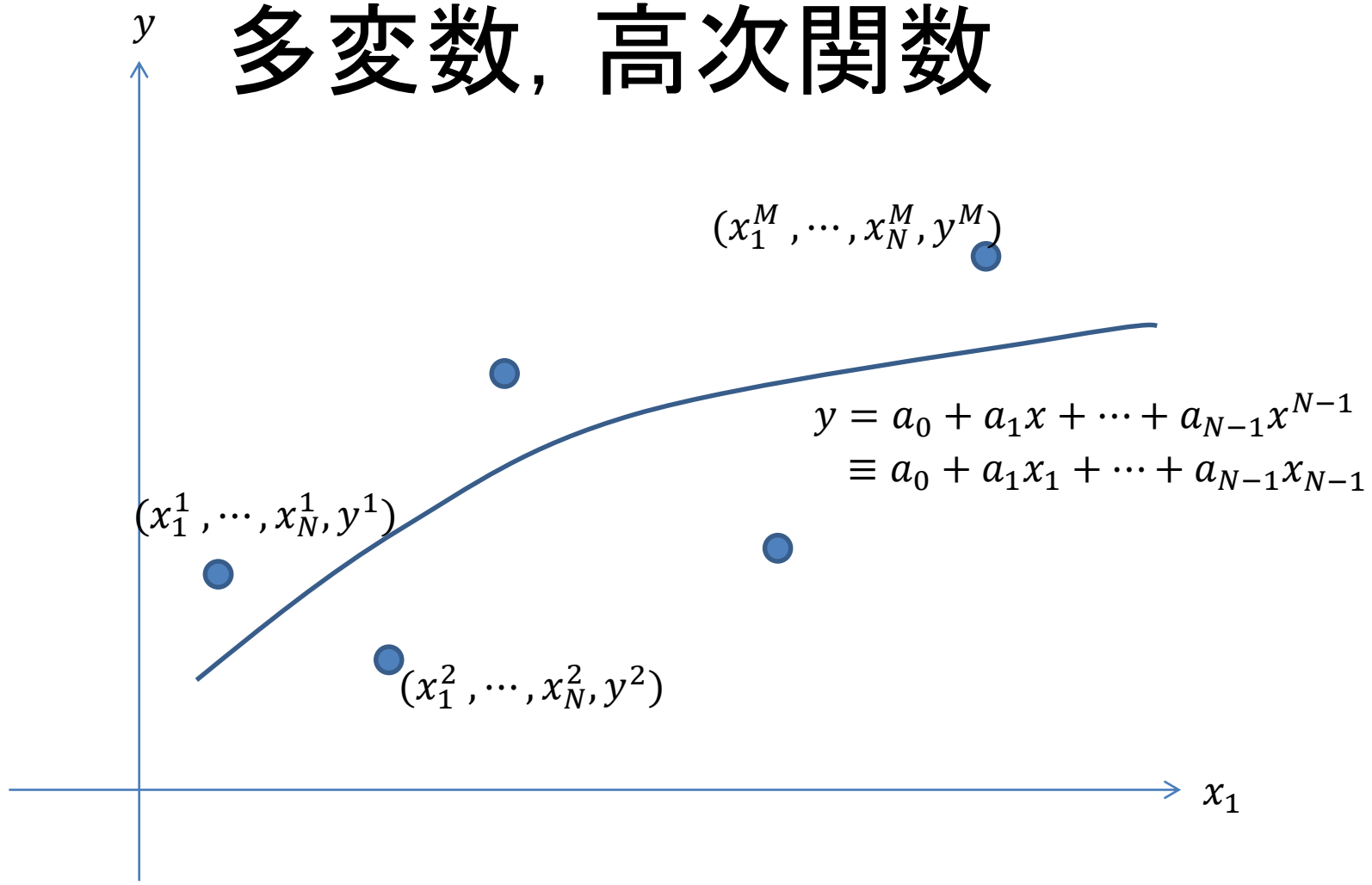
1変数の一次関数を用いると

$$y = a_0 + a_1 x$$

つまり

$$y_1 = a_0 + a_1 x_1$$
$$y_2 = a_0 + a_1 x_2$$
$$\cdots$$
$$y_N = a_0 + a_1 x_N$$

多変数，高次関数


y
$(x_1^M, \cdots, x_N^M, y^M)$
$y = a_0 + a_1 x + \cdots + a_{N-1} x^{N-1}$
$\equiv a_0 + a_1 x_1 + \cdots + a_{N-1} x_{N-1}$
$(x_1^1, \cdots, x_N^1, y^1)$
$(x_1^2, \cdots, x_N^2, y^2)$
x_1

相関係数

- R−2値
  - 相関係数．2変数の線形な関係性を示す．

変数x, yのデータがn組あるとする．つまり，

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

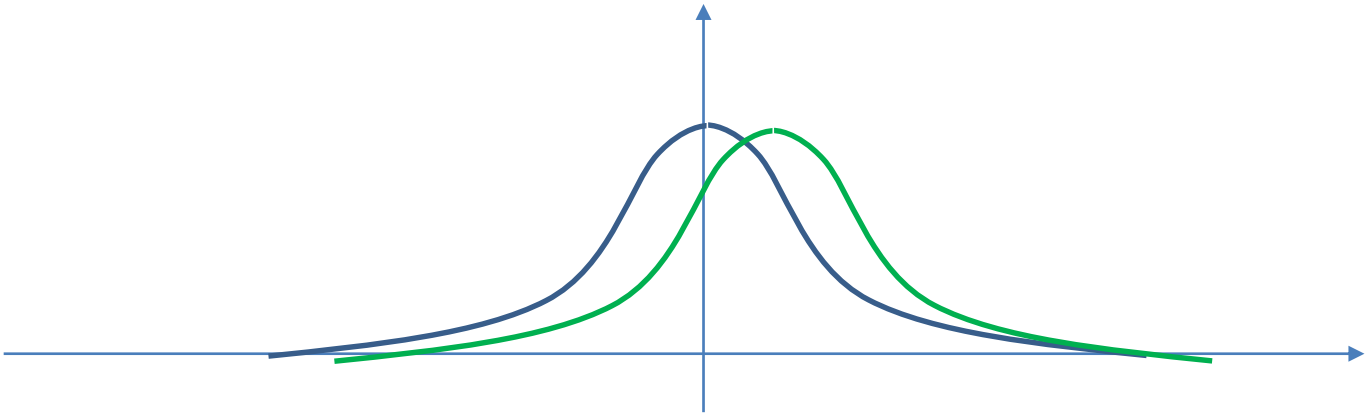
このとき，両者の相関係数は次式で与えられる．

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - y)^2}}$$

ここで，$\bar{x}, \bar{y}$はそれぞれの平均値を示す．

T検定（t値）

- T検定（t値）
  - ２つのサンプルの平均値に有意な差があるか？を示す.



T検定（t値）

- 回帰分析においては，
  - 説明変数が目的変数に与える影響の大きさ.
  - 回帰直線の勾配が0と有意に異なるかどうかを検定する.
  - この値の絶対値が2以上であれば，係数は説明変数として認めることができる.

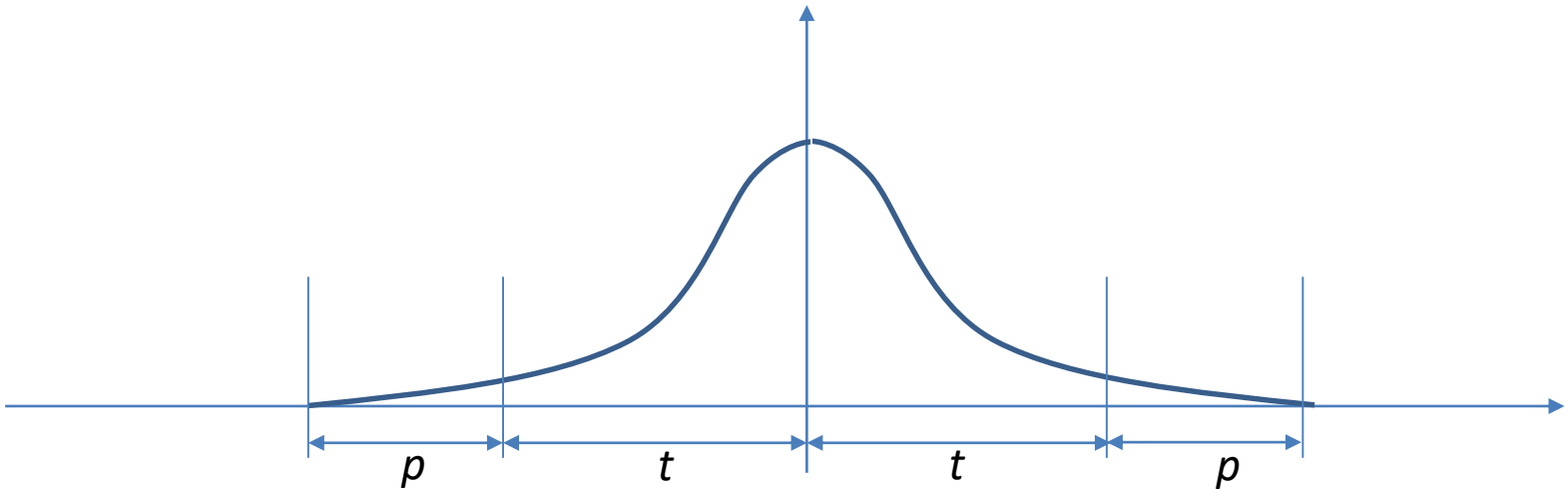
- 定義

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

相関係数：r ，サンプル数：n

P–値

- データから計算された統計量よりも極端な統計量が観測される確率.
- ある数値(相関係数等)において，p値が大きい場合は，”たまたま”その値であるだけ.
- 有意水準として1%有意，5%有意，10%有意がよく用いられる.（通常，0.05以内であればよい.）



F検定

- F検定
  - モデル式自体の有効度合を判定する
  - F分布を利用して分散の比の検定を行うもので，等分散性の検定に用いられる.
  - 実際には，t検定の前提条件である等分散性の検定に用いられることが多い.

時系列データの分析

時系列データ

時系列データ = トレンド + 周期性 + ランダム性

- トレンド
  - 潮流，流行．傾向変動（経済学）．
- 周期性
  - 特定の現象が一定期間ごとに現れる
- ランダム
  - 予測できない動き．ホワイトノイズで表現する．

自己回帰モデル

- 経済指標予測や気象予測など，時系列データの予測にもっとも一般的に用いられる方法が自己回帰モデルである.
- 最も簡単な線形自己回帰は次式で与えられる.
- 線形自己回帰モデルでは，目的変数は目的変数の過去値を説明変数とする.

自己回帰モデル

自己回帰(AutoRegressive)モデル：AR(p)

$$x_t = a_0 + \sum^{p} a_i x_{t-i} + \epsilon_t$$

ここで，誤差項

$$\epsilon_t = \sigma Z$$

$$Z \in N(0,1)$$

- 過去の値の一次関数と誤差（ランダム）との和で表現する.

ボラティリティ変動モデル

前提

時系列データが次のように表現できる.

$$x_t = \tilde{x}_t + \epsilon_t$$

ここで，$\tilde{x}_t$ は予測値，ϵ_t は予測値からの乖離幅.

乖離幅を次式で表現する.

$$\epsilon_t = \sigma_t N_t$$

σ_t:ボラティリティ

N_t:正規乱数

分散自己回帰モデル

分散自己回帰(Autoregressive conditional heteroscedasticity)モデル: ARCH(q)

- ボラティリティを次式から求める.

$$\sigma_t^2 = \alpha_0 + \sum^{q} \alpha_i \epsilon_{t-i}^2$$

自己回帰モデル：演習

自己回帰モデル

- Auto Regressive (AR) Model では，目的変数は目的変数の過去値を説明変数とする.

$$x_t = a_0 + \sum_{i=1}^{p} a_i x_{t-i} + u_t$$

$$u_t \in \sigma^2 N(0,1)$$

- 係数a_iは最尤推定を用いて決定する.
- ここでは簡単のため重回帰分析を用いることにする.

AR(1)モデル(1)

AR(1) ($p = 1$の場合)

$$x_t = a_0 + a_1 x_{t-1} + u_t$$

誤差項u_tを無視すると

$$x_t = a_0 + a_1 x_{t-1}$$

この係数a_0, a_1を求める.

方法

・x_tとx_{t-1}で線形回帰分析を行う.

AR(1)モデル(2)

AR(1)

$$x_t = a_0 + a_1 x_{t-1} + u_t$$

を変形すると，誤差項は

$$u_t = x_t - (a_0 + a_1 x_{t-1})$$

上式の誤差の平均と分散を計算する.

方法

1. 誤差を求める.
2. 誤差の平均と標準偏差を求める.
3. AR(1)モデルで予測をする.

AR(2)モデル(1)

AR(2) ($p = 2$の場合)

$$x_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + u_t$$

誤差項u_tを無視すると

$$x_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2}$$

この係数a_0, a_1, a_2を求める.

方法

- x_t, x_{t-1}, x_{t-2}で重回帰分析を行う.

AR(2)モデル(2)

AR(2)

$$r_t = a_0 + a_1 r_{t-1} + a_2 r_{t-2} + u_t$$

を変形すると，誤差項は

$$u_t = r_t - (a_0 + a_1 r_{t-1} + a_2 r_{t-2})$$

上式の誤差の平均と分散を計算する.

方法

1. 誤差を求める.
2. 誤差の平均と標準偏差を求める.
3. AR(2)モデルで予測をする.

演習問題

- 実際の株価データで同様のことを行う.

誤差 = |予測値 - 実測値|