データ解析

第2回：回帰分析

渡辺澄夫

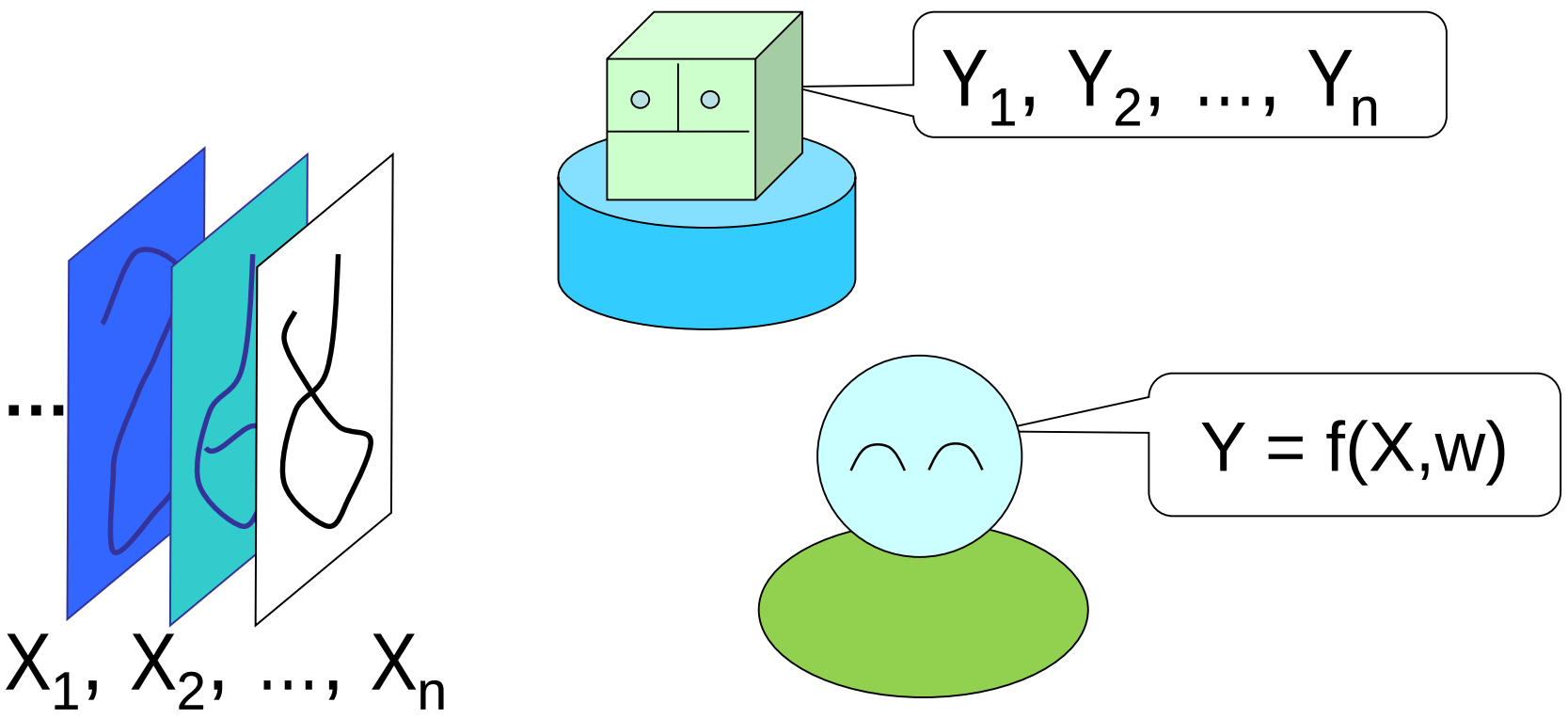
旅の
始まり

観測データ

回帰・判別分析

解析方法

回帰分析とは



確率変数 (X,Y) のデータが得られたとき、X=xが与えられたときの Y の平均値のことを回帰関数といい E[Y|x] と書く。これは x の関数である。データから回帰関数を推定する方法を考えよう。

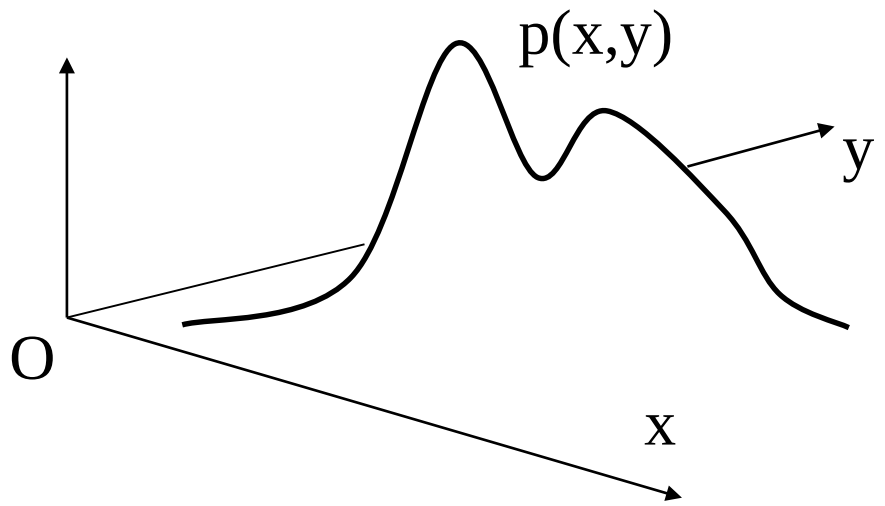
回帰関数の定義

二つの確率変数 (X,Y) の確率密度関数 q(x,y) が与えられたとき、

「x から y への回帰関数」を定義し、その数学的性質を調べます。

同時確率密度関数

定義.　(X,Y) を $R^M \times R^N$ に値をとる確率変数とし、同時確率密度関数 p(x,y) に従うものとする。ここで $x=(x_1,x_2,\ldots,x_M)$ $y=(x_1,x_2,\ldots,x_N)$ という表記を用いた。



復習. 上記の定義は、(X,Y) が集合 A の中に入る確率が

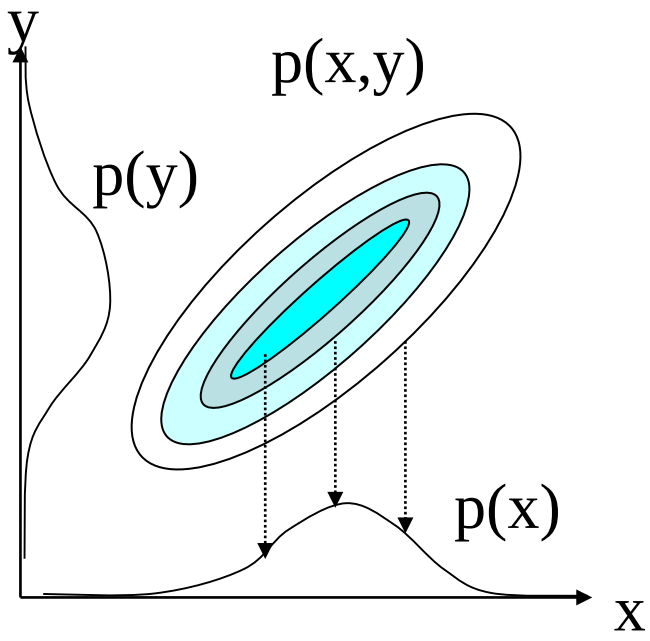
$P((X,Y)\in A) = \iint_A p(x,y)\,dxdy$ であるということである。

周辺確率密度関数

定義. 前ページの p(x,y) から定義される次の確率密度関数

$$p(x) = \int p(x,y)\, dy, \qquad p(y) = \int p(x,y)\, dx.$$

を、それぞれ X および Y の周辺密度関数という。



p(x,y)=p(x)p(y) が成り立つとき
X と Y は独立であるという。

一般には X と Y は独立ではない。

条件つき確率密度関数

定義. $p(x,y)$, $p(x)$, $p(y)$, をそれぞれ前ページまでのものとする。X が与えられたときの Y の条件つき確率 $p(y|x)$ および Y が与えられたときの X の条件つき確率 $p(x|y)$ をそれぞれ次式で定義する。

$$p(y|x) = p(x,y) / p(x),$$

$$p(x|y) = p(x,y) / p(y).$$

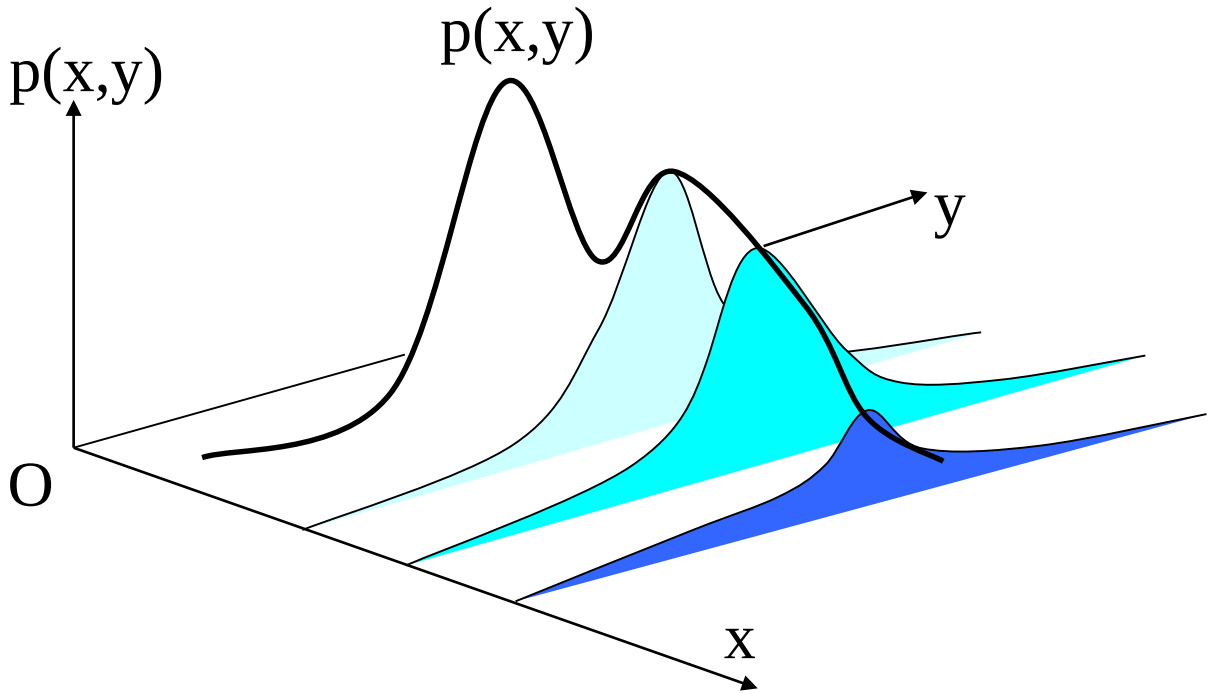
注意. $p(x)=0$ となる x に対しては $p(y|x)$ は定義されない。

定理.（ベイズの定理）$p(x,y)=p(y|x)p(x)=p(x|y)p(y)$.

注意 (この注は理解できなくても気にしなくていいです)。ユークリッド空間でない空間（例えば関数空間など）の上の確率測度に対しても条件つき確率は定義できるが、$p(x)=0$ となる集合の扱いが難しい。これは本質的に測度の割り算の問題であり、Radon-Nikodym の定理が必要になる。この講義では、そのように高度な問題は学びませんが、大切なので卒業までには勉強しましょう。

条件つき確率は推論を表す

$$p(y|x) = p(x,y) / p(x) = p(x,y) / \{ \int p(x,y')dy' \}$$

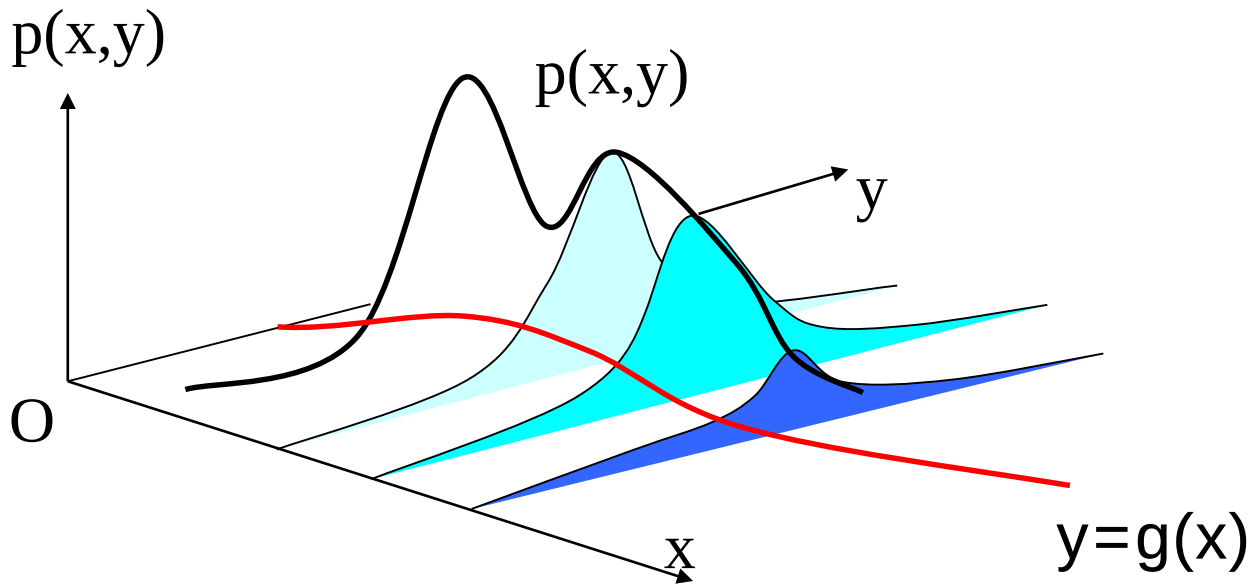


条件つき確率 p(y|x) は p(x,y) に比例していますが、y で積分したときに1になるように正規化したものになります。

回帰関数の定義

定義．p(x,y), p(x), p(y), p(y|x), p(x|y) をそれぞれ前ページまでのものとする。次の関数 g(x) を X が与えられたときの Y の回帰関数という。

$$g(x) = \int y\, p(y|x)\, dy = \int y\, p(x,y)\, dy \;/\; \{ \int p(x,y')\, dy' \}$$



例1

同時確率密度関数　$p(x,y) = (1/C) \exp(- 2x^2+2xy - y^2)$,

定数 C は　$C = \iint \exp(- 2x^2+2xy - y^2) \, dx \, dy = \pi$.

周辺確率密度関数は

$p(x) = (1/C) \int \exp(- 2x^2+2xy - y^2) \, dy = 1/\pi^{1/2} \exp(-x^2)$.

$p(y) = (1/C) \int \exp(- 2x^2+2xy - y^2) \, dx = 1/(2\pi)^{1/2} \exp(-y^2/2)$.

（復習）公式：　$a>0$ のとき $\int \exp(- a x^2) \, dx = (\pi/a)^{1/2}$

例1　つづき

条件つき確率は

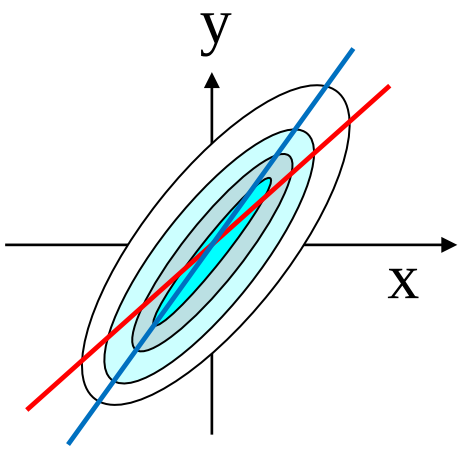
$$p(x|y) = p(x,y)/p(y) = 1/(\pi/2)^{1/2} \exp(-2(x-y/2)^2).$$

$$p(y|x) = p(x,y)/p(x) = 1/\pi^{1/2} \exp(-(y-x)^2).$$

回帰関数は

$$\int x\, p(x|y)\, dx = 1/(\pi/2)^{1/2} \int x \exp(-2(x-y/2)^2)\, dx = y/2$$

$$\int y\, p(y|x)\, dy = 1/\pi^{1/2} \int y \exp(-(y-x)^2)\, dy = x$$



注意:「XからYへの回帰関数」と「YからXへの回帰関数」は互いに逆関数ではありません。

回帰関数の性質

定理 (X,Y) を上記と同じ確率変数とする。連続関数 f の汎関数 E(f) を

$$E(f) = \iint \| y - f(x) \|^2 \, q(x,y) \, dxdy$$

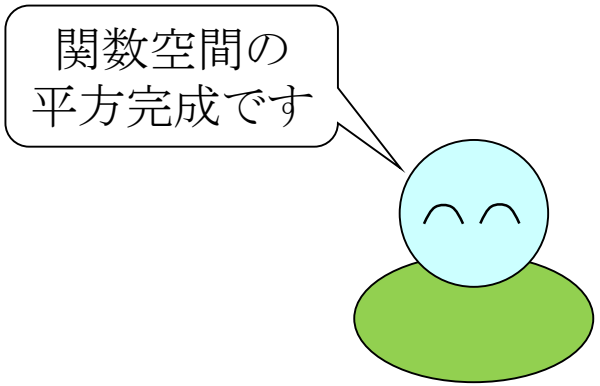
によって定義する。条件 q(x)>0 を仮定し、回帰関数 g(x) が連続関数であるとする。このとき E(f) は f(x)=g(x) のときに限り最小値

$$E(g) = \iint \| y - g(x) \|^2 \, q(x,y) \, dxdy$$

をとる。

注意.「XからYを推論したときの二乗誤差を最小にする関数は
何かと問うならば、それが回帰関数だ」ということです。

証明



汎関数 は定義から

$$E(f) = \iint \| y - f(x) \|^2 \, q(x,y) \, dxdy$$

$$= \int \{ \int \| y - f(x) \|^2 \, q(y|x) \, dy \} \, q(x) \, dx$$

$$= \int \{ \int \| y - g(x) + g(x) - f(x) \|^2 \, q(y|x) \, dy \} \, q(x) \, dx.$$

回帰関数の定義から $\int (y-g(x)) \, q(y|x) \, dy = 0$ が成り立つので

$$= \int \{ \int \| y - g(x) \|^2 \, q(y|x) dy + \| g(x) - f(x) \|^2 \} \, q(x) dx.$$

この汎関数が最小になるのは f(x)=g(x)のときだけである。（証明終）。

回帰関数の推定

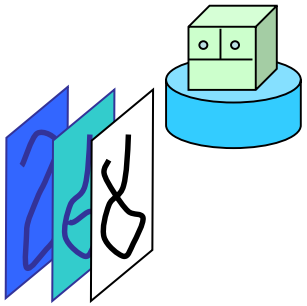
二つの確率変数 (X,Y) の確率密度関数 q(x,y) は実世界では不明です。回帰関数も分かりません。そこでパラメータ w を持つ統計モデル y=f(x,w) を用いてデータから回帰関数を推定します。統計モデルで回帰関数が実現できるとは限らないことに注意してください。

回帰分析の枠組み

データ

$X_1, X_2, \cdots, X_n$

$Y_1, Y_2, \cdots, Y_n$



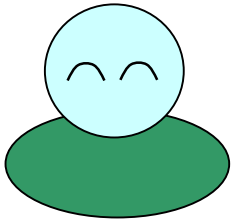
情報源

(X,Y) は $R^M \times R^N$ に値をとる確率変数

テストデータ

X

Y



統計モデル

$y=f(x,w)$

誤差関数

定義. 任意の f について E_{XY} と E_Dを次式で定義する。

$$E_{XY}[\ f(X,Y)\ ] = \iint f(x,y)\ q(x,y)\ dxdy$$

$$E_D[\ f(X,Y)\ ] = (1/n) \sum_{i=1}^{n} f(X_i,Y_i)$$

X か Y の一方だけのとき無いほうは省略

汎化誤差関数 $G(w) = E_{XY}[\ ||\ Y-f(X,w)\ ||^2\ ]$

経験誤差関数 $T(w) = E_D[\ ||\ Y-f(X,w)||^2\ ]$

注意. 同時確率密度関数 q(x,y) から定まる回帰関数を g(x) とする。
パラメータ w を持つ関数 f(x,w) はあくまでもモデルにすぎないので
一般には f(x,w) = g(x) となるようなパラメータ w が存在するとは限らない。

注意. 真の目的は汎化誤差 G(w) を小さくすることであるが実世界では
汎化誤差関数を直接に知ることはできない。代わりに T(w)を最小化する
ことでパラメータを用いるが・・・。

線形モデル

線形モデル　$f(x,w) = \sum_{k=1}^{m} w_k f_k(x)$　　($\{ w_k \}$ 実数パラメータ)

条件. $q(x)>0$ とする。$\{f_k(x)\}$ は M次元ベクトル x から N 次元空間への関数で線形独立であるものとする。

<u>定義</u>　行列F, F* とベクトル a, a* を次式で定義(j,k=1,2,..,m).

$$F_{jk} = E_X [f_j(X)^T f_k(X)],$$
$$F^*_{jk} = E_D [f_j(X)^T f_k(X)],$$
$$a_j = E_{XY} [Y^T f_j(X)],$$
$$a^*_j = E_D [Y^T f_j(X)].$$

補題.　行列 F が正定値　⇔　$\{f_k(x)\}$ は線形独立

線形モデルのパラメータ推定

定理.　線形モデルにおいて G(w), T(w) を最小にする w を w_0, w*と書くと

$$w_0 = F^{-1} a,$$
$$w^* = (F^*)^{-1} a^*.$$

注意. F* と a* はデータから計算できるので w* も計算できる。経験誤差 T(w) を最小にするパラメータ w* を二乗誤差最小推定量という。

補題. $w^* - w_0 = (F^*)^{-1} \times (a^* - F^* w_0)$

F^{-1}に確率収束　　平均0, 分散∝(1/n)

同じ分布からの独立なデータの平均なので

だから w* は w_0 に確率収束し、 $w^* - w_0$ は $(1/n^{1/2})$ オーダー。

注意. 推定された w* は w0 と同じではなく$(1/n^{1/2})$ オーダーだけずれている。

証明

まず G(w) について示す。

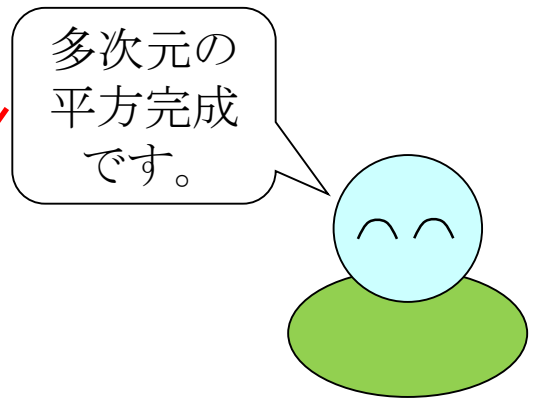
$$G(w) = E_{XY}\left[\| Y - \sum_{k=1}^{m} w_k f_k(X) \|^2 \right]$$

$$= E_Y[\| Y \|^2] - 2E_{XY}[\sum_k w_k Y^T f_k(X)] + E_{XY}[\sum_j \sum_k w_j w_k f_j(X)^T f_k(X)]$$

$$= E_Y[\| Y \|^2] - 2 \sum_k w_k a_k + \sum_j \sum_k w_j w_k F_{jk}(x)$$

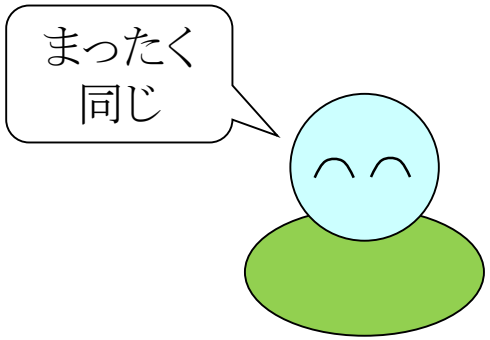
$$= (w, Fw) - 2 (a,w) + E_Y [\| Y \|^2]$$

$$= \| F^{1/2} (w - F^{-1}a) \|^2 - (a, F^{-1}a) + E_Y [\| Y \|^2]$$



F は正定値なので最小値は $w = F^{-1}a$ のときである。

証明つづき



次に T(w) について示す。計算は同様である。

$$T(w) = E_D [\| Y - \sum_{k=1}^{m} w_k f_k(X) \|^2]$$

$$= E_D [\| Y \|^2] - 2E_D [\sum_k w_k Y^T f_k(X)] + E_D [\sum_j \sum_k w_j w_k f_j(X)^T f_k(X)]$$

$$= E_D [\| Y \|^2] - 2 \sum_k w_k a^*_k + \sum_j \sum_k w_j w_k F^*_{jk}(X)$$

$$= (w, F^*w) - 2 (a^*,w) + E_D [\| Y \|^2]$$

$$= \| F^{*1/2} (w - (F^*)^{-1}a^*) \|^2 - (a^*, (F^*)^{-1}a^*) + E_D [\| Y \|^2]$$

F* は確率1で正定値である。最小値は確率1で $w^* = (F^*)^{-1}a^*$ のときである。

（証明終）

経験誤差と汎化誤差

定理. 線形モデル $f(x,w) = \sum_{k=1}^{m} w_k f_k(x)$ による推定を考える。

(X,Y)の同時確率密度関数 $q(x,y)=q(x)q(y|x)$ が $q(x)>0$ で，$q(y|x)$ が「$Y= f(X,w_0) + Z$」であると仮定する。ここで Z は平均0で分共分散行列が $(\sigma^2 I)$ の正規分布に従う確率変数 (Iは単位行列）であるとする。データ $\{(X_i,Y_i);i=1,2,..,n\}$ の出方についての平均を E で表すと n が大きいとき

$$E[\, G(w^*) \,] = N\sigma^2 + m\sigma^2/n + o(1/n)$$
$$E[\, T(w^*) \,] = N\sigma^2 - m\sigma^2/n + o(1/n)$$

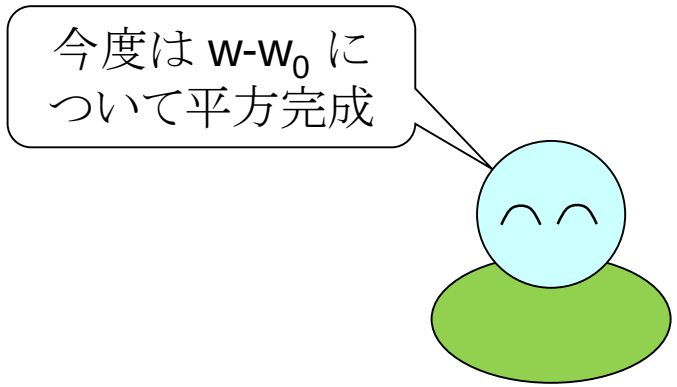
ここで N はYの次元.

注意. 二乗誤差最小推定量 w^* を用いると経験誤差は汎化誤差に収束する（n が無限大に近づくとき）。しかし n が有限であるならば両者の値は等しくない。汎化誤差は経験誤差よりも平均的に $2m\sigma^2/n$ ほど大きい。特に次式が成立。

$$E[\, G(w^*) \,] = (1+ 2m/(nN))\, E[\, T(w^*) \,] + o(1/n)$$

これより経験誤差平均より汎化誤差平均を求めることができる。

証明



経験誤差関数を計算すると

$$T(w) = E_D [\| Y - f(X,w) \|^2]$$

$$= E_D [\| Z + f(X,w_0) - f(X,w) \|^2]$$

$$= E_D [\| Z \|^2] + 2 E_D[Z^T \{ f(X,w_0) - f(X, w) \}] + E_D [\| f(X,w_0) - f(X,w) \|^2]$$

$$= E_D [\| Z \|^2] + 2 \Sigma_j b_j (w-w_0)_j + \Sigma_j \Sigma_k F^*_{jk} (w-w_0)_j (w-w_0)_k$$

ここで $b=(b_j)$ ただし $b_j=E_D [Z^T f_j (X)]$ とおいた ($f(X,w) = \Sigma_j w_j f_j(X)$)。

従って $T(w)$ を最小にする w^* は $(w^*-w_0) = - (F^*)^{-1} b$ を満たす。また

$$T(w^*) = E_D [\| Z \|^2] - \Sigma_j \Sigma_k (F^*)^{-1}_{jk} b_j b_k$$

$$= E_D [\| Z \|^2] - tr(F^* b b^T)$$

(注意. 任意のベクトルu,v と行列Aについて $(u,Av)=tr(Avu^T)$が成り立つ。)

証明つづき

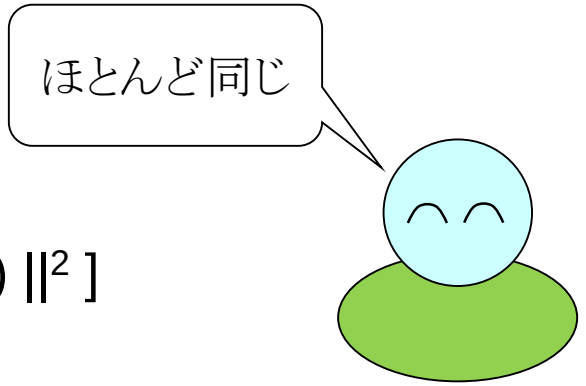
汎化誤差関数は

$$G(w) = E_{XY}[\,\| Y - f(X,w) \|^2\,]$$

$$= E_{XZ}[\,\| Z + f(X,w_0) - f(X,w) \|^2\,]$$

$$= E_Z[\| Z \|^2] + E_X[\,\| f(X,w_0) - f(X,w) \|^2\,]$$

$$= N\sigma^2 + \Sigma_j \Sigma_k F_{jk} (w-w_0)_j (w-w_0)_k$$



ここで $(w^*-w_0) = -(F^*)^{-1} b$ であるから

$$G(w^*) = N\sigma^2 + \Sigma_j \Sigma_k F_{jk} [(F^*)^{-1}b]_j [(F^*)^{-1}b]_k$$

$$= N\sigma^2 + \mathrm{tr}((F^*)^{-1} F (F^*)^{-1} bb^T)$$

$b=(b_j)$, $b_j=E_D[Z^T f_j(X)]$ とおいたので，仮定より bは平均0分散共分散が $(\sigma_2/n)F$ の正規分布に従う。F^* が F に平均収束し tr(単位行列)=m より

$$E[G(w^*)] = N\sigma^2 + m\sigma^2/n + o(1/n),$$

$$E[T(w^*)] = N\sigma^2 - m\sigma^2/n + o(1/n).$$ （証明終）

回帰分析の実際

昔々、惑星の挙動は天空の中の複雑な挙動を描くためとても
不思議なことだった。数学者ガウスは最小二乗法を用いて
その軌道を予測して人々を感動させたという。

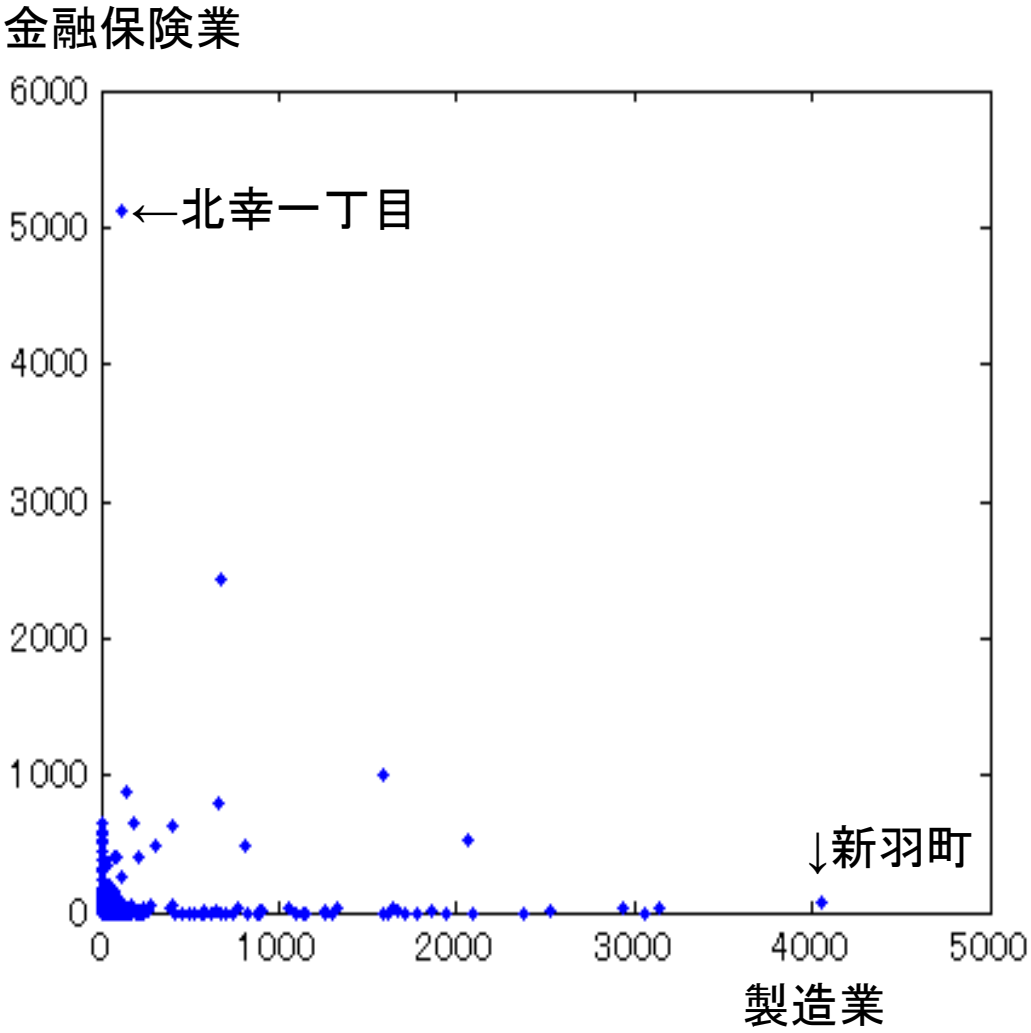
実際の例(1)

政府統計の総合窓口（e－Stat）を使わせて頂きました。
http://www.e-stat.go.jp/estat/html/spec.html
http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do

横浜市の n=1070 の「〇町〇丁目」について下記の業種で働く人の数を調べた。

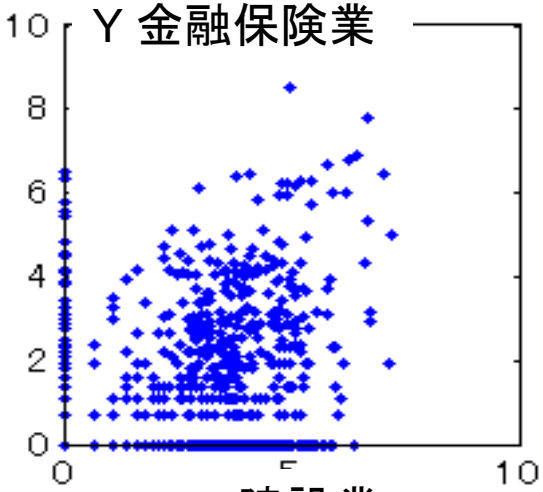
X1 建設業
X2 製造業
X3 運輸郵便業
X4 卸売小売業
X5 不動産業
X6 学術技術業
Y　金融保険業

町の規模によって働く人の数は大きく異なるのでlog(1+X_k) と log(1+Y) を用いることにした。
X=(X1,X2,X3,X4,X5,X6) から Y を推測してみよう。

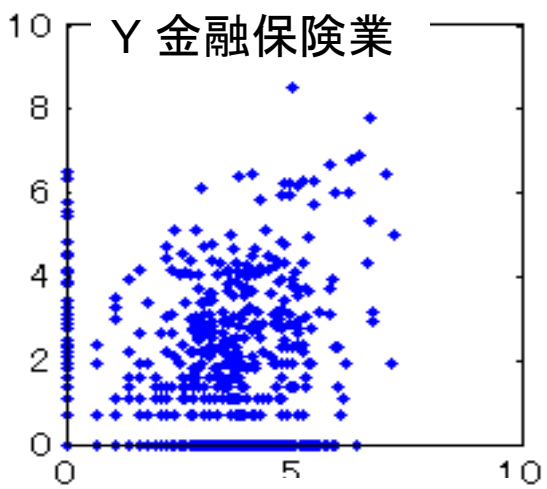


実際の例(2)

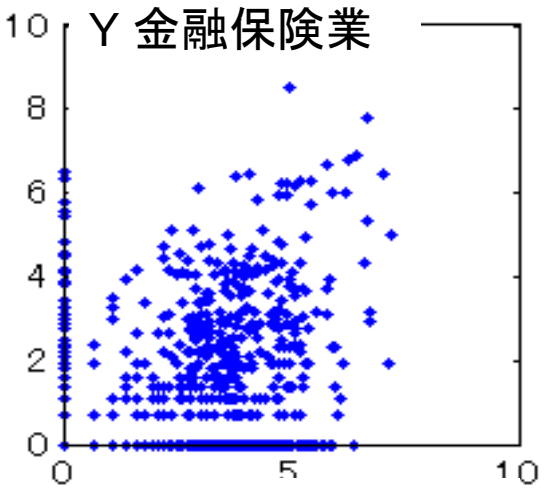
$\log(1+X_k)$ と $\log(1+Y)$ のデータ



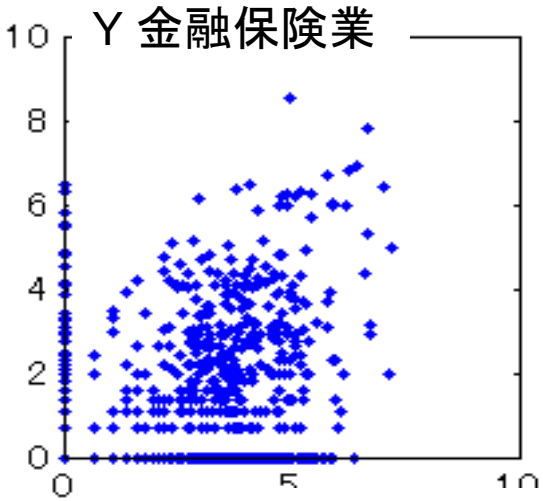
X1 建設業



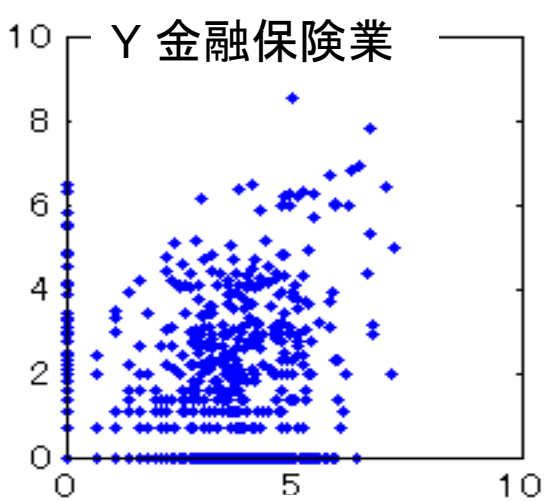
X2 製造業



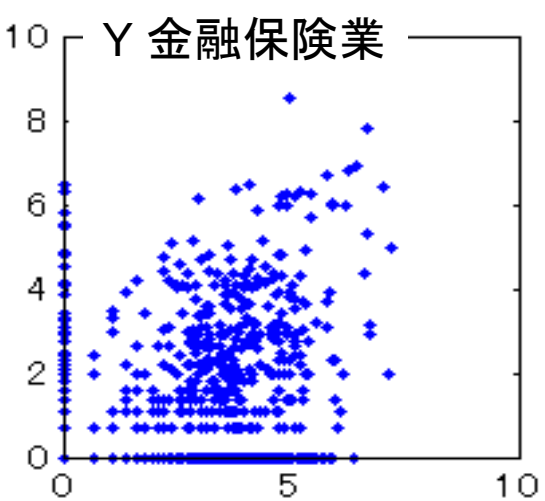
X3 運輸郵便業



X4 卸売小売業



X5 不動産業



X6 学術技術業

回帰分析の例

以下では $X_1=1$, $X_k=\log(1+X_{k+1})$, $Y=\log(1+Y)$ とおく。

データ i 番めを $X_i = (X_{ik})$, Y_i と書く。

モデル $Y= \sum_{k=1}^{K} w_k X_k$ を使って回帰分析をしてみた。

$$F^*_{jk} = (1/n) \sum_{i=1}^{n} X_{ij}\, X_{ik}$$

$$a^*_j = (1/n) \sum_{i=1}^{n} Y_i^T\, X_{ij}$$

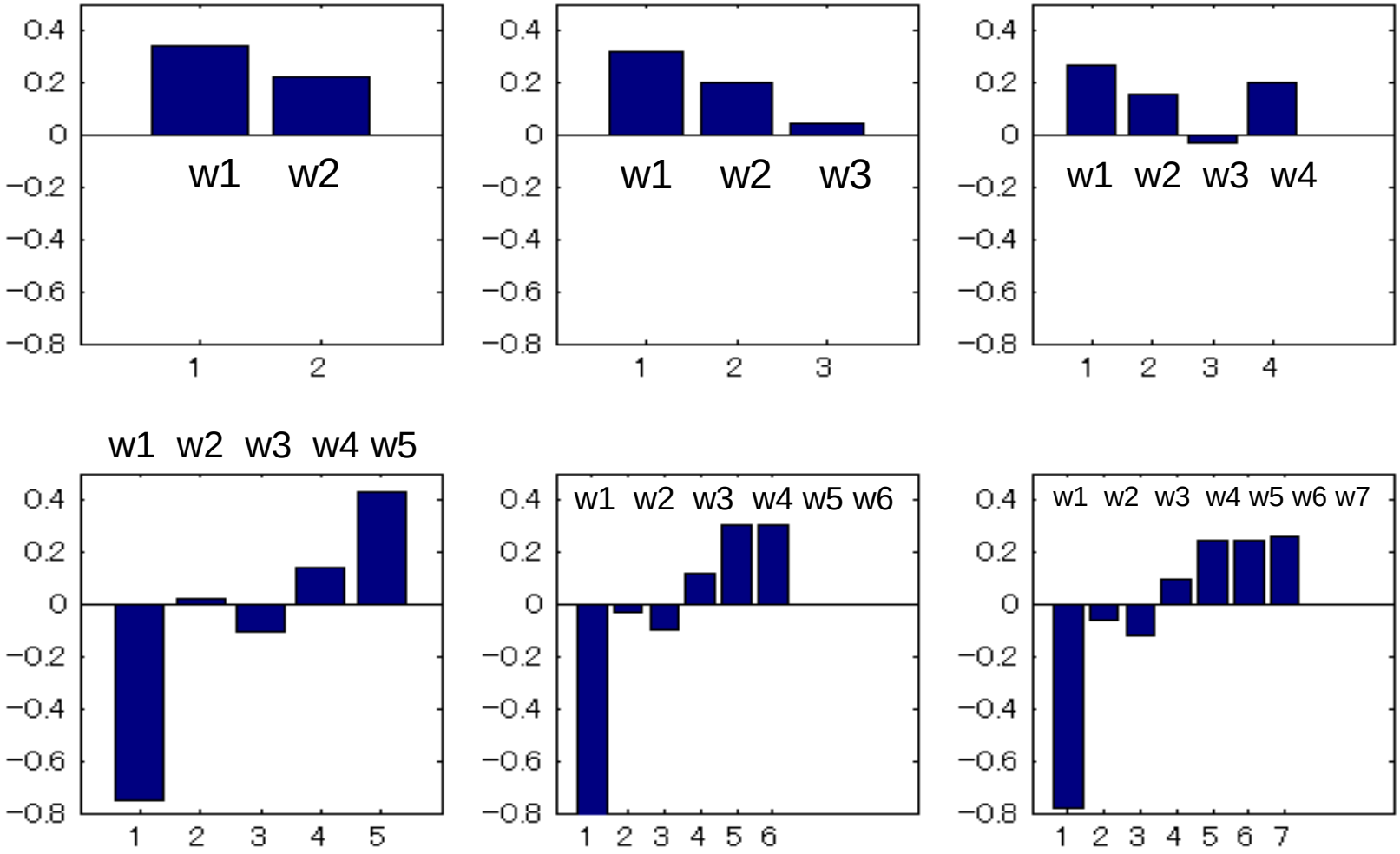
を用いて求めるパラメータは　$w^*=(F^*)^{-1}\, a^*$

K=1,2,3,4,5,6 の場合について w* を求めて比較した。

◎　w* の値からわかることがあるだろうか。

パラメータ

X1 定数　X2 建設業　X3 製造業　X4 運輸郵便業
X5 卸売小売業 X6 不動産業　X7学術技術業　Y 金融保険業



金融保険業に対して回帰係数がマイナス･･･建設業　製造業
金融保険業に対して回帰係数がプラス･･卸売小売業　不動産業　学術技術業

特徴量

(X,Y) の次元が小さい場合には、データを図示することができるので分析を始める前に図示をして全体像を把握する（基本中の基本）。

(X,Y) が河原で拾ってきた小石の表面積と体積だとしたら、
$(X^{1/2}, Y^{1/3})$ を使って回帰するのが良さそうである。

(X,Y) が天文学で現れる距離や大きさのときは (log X, log Y) にするといいかもしれない。

データ解析がうまくいくかどうかは、考察しているデータに対して適切な特徴量に気づくかどうかが最初のポイント。

今日、広く使われている深層学習では、人間が適切な特徴量に気づくことができなくても、モデル内に自動的に特徴量が生成されると言われています。将来、真に難しい問題に挑むとき使ってみてください。

因果関係と制御について

ここで想定されている (X,Y) は確率的な共起関係を表しているだけであり
どちらがどちらの原因と結果であるというわけではない。

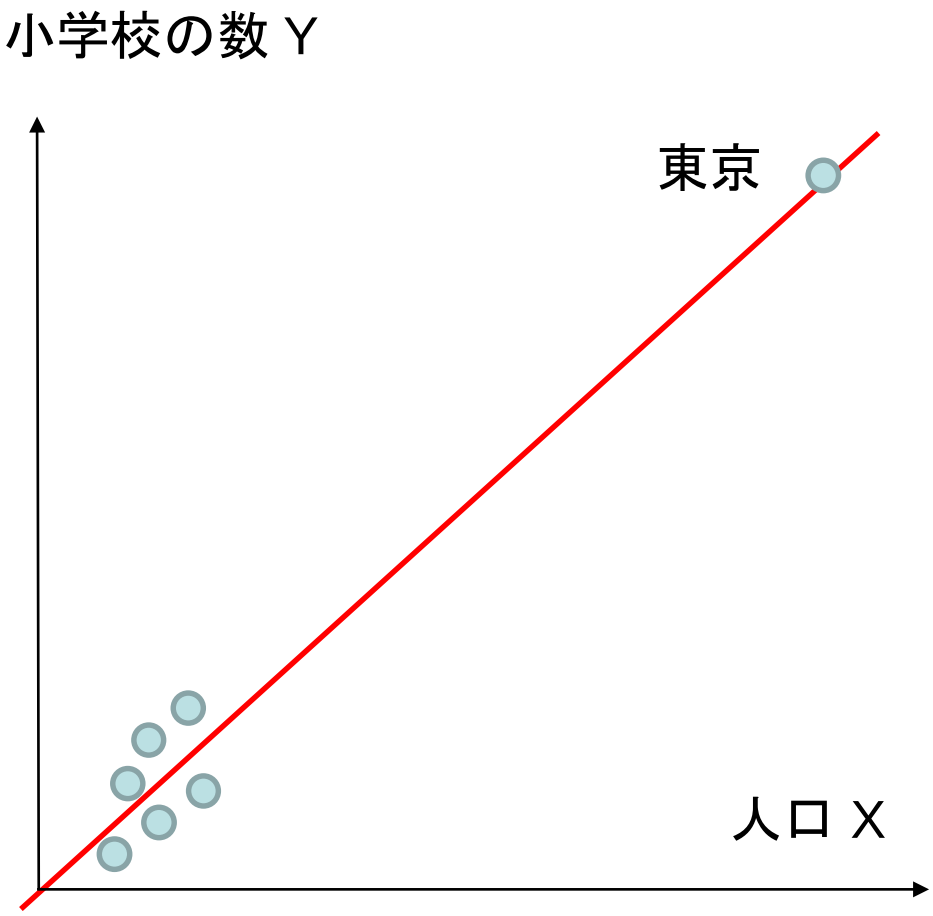
原則としてデータだけから因果関係を取り出すことはできない。
また、どちらか一方が原因で他方が結果であるということもない。

例.「春になると花が咲き（X）鳥が歌う（Y）」。XからYが推測できた
としてもＸが Ｙ の原因であると限らない。Ｘ と Ｙ の背後に共通の
原因があることもある。実世界は複雑です。

XからYへの条件つき確率あるいは回帰関数が得られたとき、XからYへの
確率的な推論が可能になる。しかし、Ｘ を変化させることで Ｙ を変化させる
（つまりXを操作してYを制御する）ことができるとは限らない。

例.「冬になる(X)、鍋物を食べる(Y)」においてＹ から Ｘ への条件つき
確率が計算できても、鍋物を食べて季節を変えることはできない。

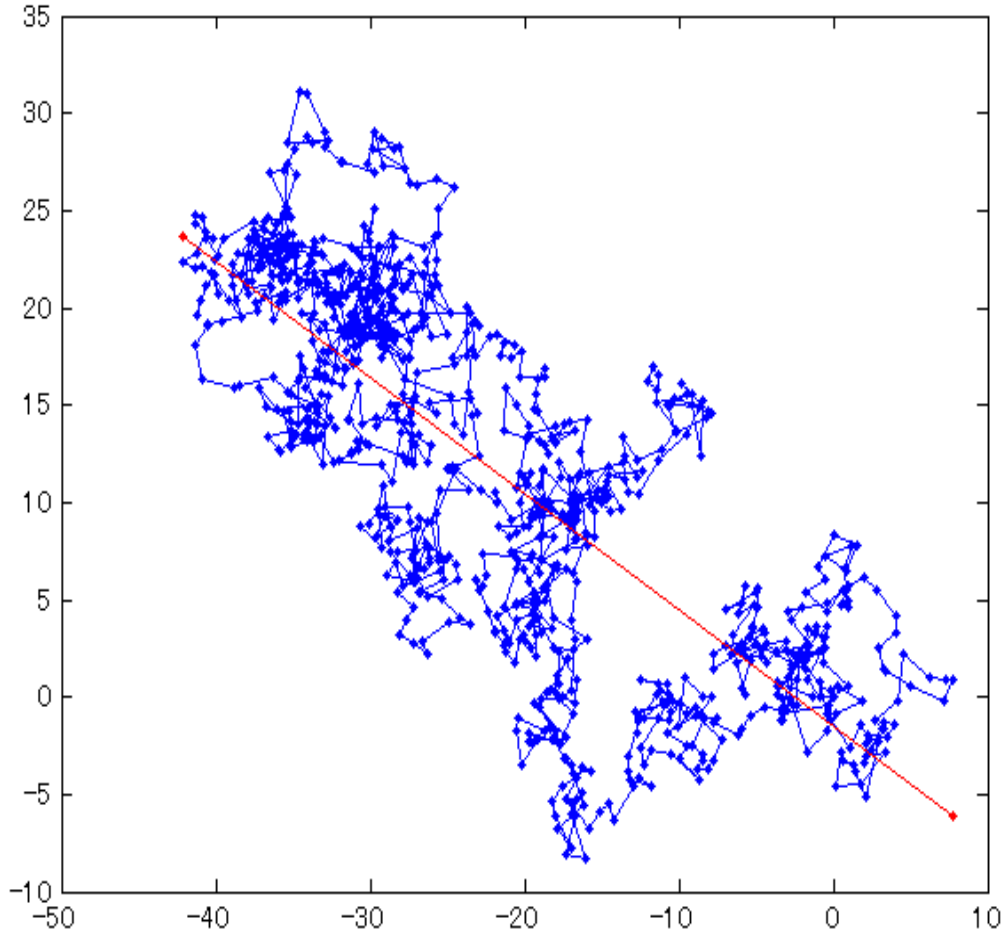
影響力の大きなデータ



ある都市の人口Xから小学校の数YをY=aX で推定するケースを考えよう。Xの例のほとんどが人口数千人くらいであるのに対して、ひとつのデータだけが東京（X=千万）のような場合、東京のデータの影響力が非常に大きく(leverage, てこ)、適切な推測ができないことがある。

Y=aXの推定精度をあげるためにX を工夫する方法は実験計画法あるいは能動学習法と呼ばれているがむやみに X を大きくして大丈夫かどうかは、問題ごとに考えることになる。

ランダムウォークに回帰？



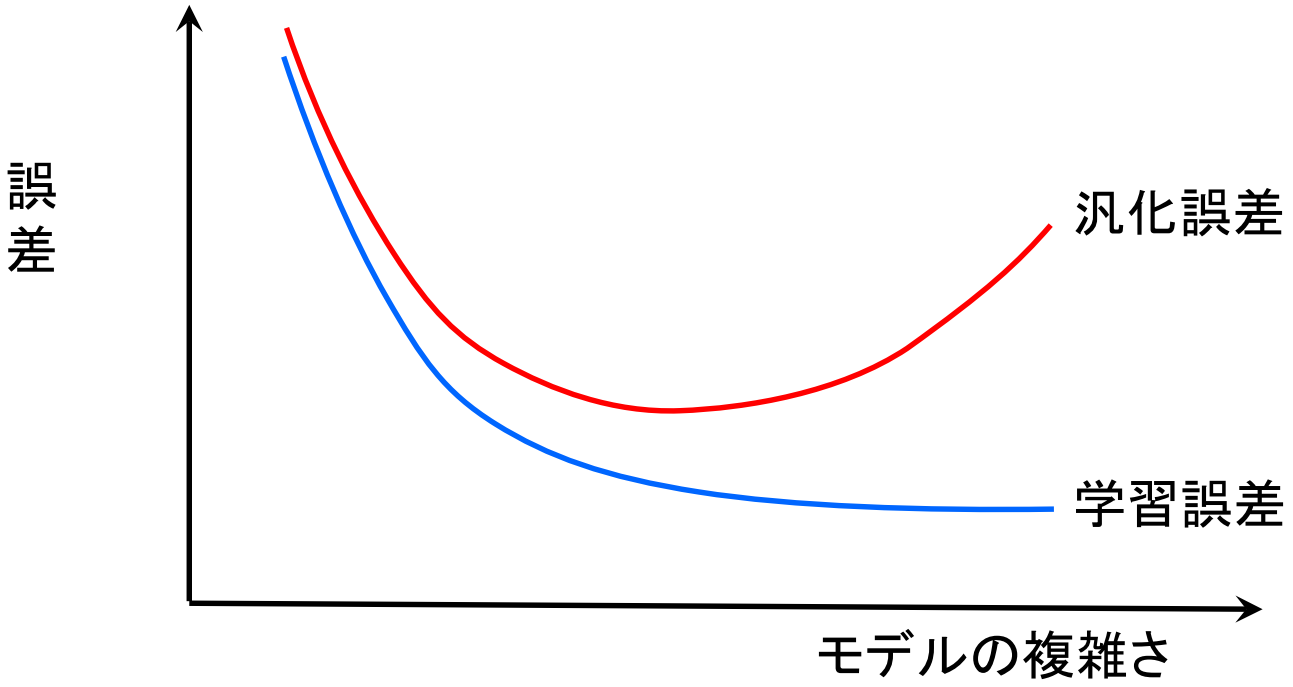
(X,Y) が独立なデータでなくても
二乗誤差を最小にするパラメータ
を求めることは計算上は可能で
あるが、(X,Y)が時系列であり
定常性を持たない場合、たとえば
ランダムウォークのような場合には
回帰曲線を求めても意味がない。

データの性質をあらかじめ調べず
に最小二乗法を適用すると
適切でない推測が行われる
場合もあるので注意しよう。

←は、ランダムウォークの結果に
直線をあてはめた例。たまたま
得られた軌跡に直線が当てはまる。

「最小二乗法＋モデル選択」

モデルが複雑であるほど「汎化誤差ー学習誤差」は大きくなる
（オーバーフィット）。



多数のパラメータを持つ複雑なモデルほど多様な関数が実現できるので
関数近似誤差は小さくなるが、オーバーフィットによる統計誤差は大きくなる。
前者をバイアスといい後者をバリアンスという。両方のバランスを取ることが
大切である。（バイアスーバリアンスの問題）。

正則化

経験誤差関数に正則化項を加えて最小化する(Tikhonov)と汎化誤差をちいさくできることがある。

$$H(w) = \sum_{i=1}^{n} ||Y_i - f(X_i, w)||^2 + R(w)$$

例　λ>0 (ハイパーパラメータ)として

Ridge 項　　$R(w) = \lambda \Sigma |w_j|^2$

Lasso 項　　$R(w) = \lambda \Sigma |w_j|$

汎化能力が向上するかどうかは、真の分布、学習モデル、正則化項、ハイパーパラメータによって異なる。どのようにλを最適化すると良いのだろうか。

◎　Ridge : 大昔からあるので最初の研究は不明。たぶんTikhonov。
◎　Lasso : 石川眞澄(1990)、Tibshirani (1996) など。