回帰 (regression)

- 黒木玄

- 2022-07-13〜2022-07-13

このノートでは[Julia言語 (https://julialang.org/)](https://julialang.org/)を使用している:

- [Julia言語のインストールの仕方の一例 (https://nbviewer.org/github/genkuroki/msfd28/blob/master/install.ipynb)](https://nbviewer.org/github/genkuroki/msfd28/blob/master/install.ipynb)

自明な誤りを見つけたら, 自分で訂正して読んで欲しい. 大文字と小文字の混同や書き直しが不完全な場合や符号のミスは非常によくある.

このノートに書いてある式を文字通りにそのまま読んで正しいと思ってしまうとひどい目に会う可能性が高い. しかし, 数が使われている文献には大抵の場合に文字通りに読むと間違っている式や主張が書いてあるので, 内容を理解した上で訂正しながら読んで利用しなければいけない. 実践的に数学を使う状況では他人が書いた式をそのまま信じていけない.

このノートの内容よりもさらに詳しいノートを自分で作ると勉強になるだろう. 膨大な時間を取られることになるが, このノートの内容に関係することで飯を食っていく可能性がある人にはそのためにかけた時間は無駄にならないと思われる.

目次

In [1]:
```julia
ENV["LINES"], ENV["COLUMNS"] = 100, 100
using Base.Threads
using BenchmarkTools
using DataFrames
using Distributions
using LinearAlgebra
using Memoization
using Printf
using QuadGK
using RCall
using Random
Random.seed!(4649373)
using Roots
using SpecialFunctions
using StaticArrays
using StatsBase
using StatsFuns
using StatsPlots
default(fmt = :png, size = (400, 250),
    titlefontsize = 10, plot_titlefontsize = 12)
using SymPy
```

In [2]:

```julia
# Override the Base.show definition of SymPy.jl:
# https://github.com/JuliaPy/SymPy.jl/blob/29c5bfd1d10ac53014fa7fef468bc8deccadc2fc/src/types.

@eval SymPy function Base.show(io::IO, ::MIME"text/latex", x::SymbolicObject)
    print(io, as_markdown("\\displaystyle " *
            sympy.latex(x, mode="plain", fold_short_frac=false)))
end
@eval SymPy function Base.show(io::IO, ::MIME"text/latex", x::AbstractArray{Sym})
    function toeqnarray(x::Vector{Sym})
        a = join(["\\displaystyle " *
                sympy.latex(x[i]) for i in 1:length(x)], "\\\\")
        """\\left[ \\begin{array}{r}$a\\end{array} \\right]"""
    end
    function toeqnarray(x::AbstractArray{Sym,2})
        sz = size(x)
        a = join([join("\\displaystyle " .* map(sympy.latex, x[i,:]), "&")
                for i in 1:sz[1]], "\\\\")
        "\\left[ \\begin{array}{" * repeat("r",sz[2]) * "}" * a * "\\end{array}\\right]"
    end
    print(io, as_markdown(toeqnarray(x)))
end
```

In [3]:

```julia
safemul(x, y) = x == 0 ? x : isinf(x) ? typeof(x)(Inf) : x*y
safediv(x, y) = x == 0 ? x : isinf(y) ? zero(y) : x/y

x ≲ y = x < y || x ≈ y

mypdf(dist, x) = pdf(dist, x)
mypdf(dist::DiscreteUnivariateDistribution, x) = pdf(dist, round(Int, x))

distname(dist::Distribution) = replace(string(dist), r"{.*}" => "")
myskewness(dist) = skewness(dist)
mykurtosis(dist) = kurtosis(dist)
function standardized_moment(dist::ContinuousUnivariateDistribution, m)
    μ, σ = mean(dist), std(dist)
    quadgk(x -> (x - μ)^m * pdf(dist, x), extrema(dist)...)[1] / σ^m
end
myskewness(dist::MixtureModel{Univariate, Continuous}) =
    standardized_moment(dist, 3)
mykurtosis(dist::MixtureModel{Univariate, Continuous}) =
    standardized_moment(dist, 4) - 3
```

Out[3]: mykurtosis (generic function with 2 methods)

1 回帰 (regression)

このノートでは **回帰** (regression)について簡単に説明する.

1.1 回帰の超一般論

一般に, 回帰はデータの数値が $(x_1, y_1), \dots, (x_n, y_n)$ の形式で各々の y_i が数値 x_i に依存して決まるという場合を, パラメータ $x = (x_1, \dots, x_n)$, $\theta = (\theta_1, \dots, \theta_d)$ を持つ $y = (y_1, \dots, y_n)$ に関する確率密度函数(または確率質量函数)

$$p(y|x, \theta) \quad (\text{または } P(y|x, \theta))$$

でモデル化することによって行われるパラメータ θ の推定のことである.

比率の信頼区間の計算で使われる統計モデルは, 二項分布モデル

$$P(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k} \quad (k = 0, 1, \dots, n)$$

であった. この場合には y にあたる変数が k で, x にあたる変数はない.

平均の信頼区間の t 分布を使った信頼区間の計算で使われる統計モデルは, 正規分布の標本分布モデル

$$p(y|\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right)$$
$$(\mu, \sigma \in \mathbb{R},\ \sigma > 0,\ y = (y_1, \dots, y_n) \in \mathbb{R}^n)$$

であるとみなせる(実際には中心極限定理による近似がうまく行っているという弱い仮定のもとで, その平均の信頼区間は実用的に使用可能). この場合には上の y にあたる変数がこの場合の y で, x にあたる変数はない.

x にあたる変数がモデルのパラメータとして登場することが, 回帰(regression)の特徴である.

2 線形回帰

線形回帰一般の良い解説が見当たらなかったので, 以下の節で解説することにする. 以下の節のスタイルでの線形回帰の理解は機械学習一般の理解のためにも有用である.

2.1 データ

データは以下の形で得られると仮定する:

$$(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2.$$

これを2つのベクトルで表す:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

2.2 モデルの構成要素

$r < n$ であると仮定する.

r 個の函数 $f_1, \dots, f_r : \mathbb{R} \to \mathbb{R}$ が与えられているとし, その一次結合を決める **回帰係数** パラメータ $\beta_1, \dots, \beta_r \in \mathbb{R}$ が与えられているとし,

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, \quad f = \begin{bmatrix} f_1 \\ \vdots \\ f_r \end{bmatrix}, \quad f(x_i) = \begin{bmatrix} f_1(x_i) \\ \vdots \\ f_r(x_n) \end{bmatrix}$$

とおく. このとき,

$$\sum_{j=1}^r \beta_j f_j(x_i) = f(x_i)^T \beta.$$

ここで, $f(x_i)^T$ は縦ベクトル $f(x_i)$ を転置して得られる横ベクトルであるとする.

さらに, 分散パラメータ $\sigma^2 > 0$ が与えられているとする.

2.3 デザイン行列(計画行列, design matrix)の定義

デザイン行列 X を次のように定める:

$$X = \begin{bmatrix} f(x_1)^T \\ f(x_2)^T \\ \vdots \\ f(x_n)^T \end{bmatrix} = \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_r(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_r(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x_n) & f_2(x_n) & \cdots & f_r(x_n) \end{bmatrix}.$$

このとき,

$$X\beta = \left[\sum_{j=1}^r \beta_j f_j(x_i) \right]_{i=1}^n.$$

以下では簡単のため, デザイン行列 X のランクは可能な最大値の r であると仮定する.

このとき $\mathbb{R}^n$ の部分空間 $X\mathbb{R}^r$ の次元は r になり, 行列 X の定める線形写像 $X : \mathbb{R}^r \to \mathbb{R}^n$ は単射になる.

2.4 統計モデルの記述

y に関する次の確率密度函数を統計モデルとして採用する:

$$p(y|X,\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right).$$

ここで,

$$\|y - X\beta\|^2 = \sum_{i=1}^n \left(y_i - \sum_{j=1}^r \beta_j f_j(x_i)\right)^2.$$

$Y = [Y_i]_{i=1}^n$ をこの統計モデルに従う確率変数とすると,

$$Y = X\beta + \varepsilon.$$

ここで, $\varepsilon = [\varepsilon_i]_{i=1}^n$ と書くとき, $\varepsilon_1, \ldots, \varepsilon_n$ はそれぞれが平均 0, 分散 σ^2 の正規分布に従う独立な確率変数達になる. ゆえに,

$$E[Y] = X\beta, \quad E[\varepsilon\varepsilon^T] = \sigma^2 I.$$

ここで I は n 次の単位行列を表す.

このモデル内では, 仮想的なデータの数値 Y_i が $f_j(x_i)$ 達の一次結合 $\sum_{j=1}^r \beta_j f_j(x_i)$ で近似され, それらの差が平均 0, 分散 σ^2 の正規分布に従ってランダムに決まっている.

2.5 最尤法から最小二乗法による線形回帰が得られること

上の統計モデルのデータの数値 x, y に関する尤度函数 $(\beta, \sigma^2) \mapsto p(y|X,\beta,\sigma^2)$ を最小化するパラメータ β, σ^2 の値から, 最小二乗法が得られることを説明しよう.

尤度函数の対数の -2 倍は次の形になる:

$$L(\beta,\sigma^2) = -2\log p(y|X,\beta,\sigma^2) = \frac{1}{\sigma^2}\|y - X\beta\|^2 + n\log\sigma^2 + n\log(2\pi).$$

これを最小にする β, σ^2 を求めよう.

y と $X\beta$ の距離の2乗 $\|y - X\beta\|^2$ を最小にする $X\beta$ は $y \in \mathbb{R}^n$ の部分空間 $X\mathbb{R}^r$ への直交射影 $\hat{y}$ になる.

$X : \mathbb{R}^r \to \mathbb{R}^n$ は単射なので $\hat{y} = X\hat{\beta}$, $\hat{\beta} \in \mathbb{R}^r$ と一意に表される.

$\hat{\sigma}^2$ を $\hat{\sigma}^2 = \|y - \hat{y}\|^2/n = \|y - X\hat{\beta}\|^2/n$ と定める. このとき,

$$L(\hat{\beta},\sigma^2) = \frac{n\hat{\sigma}^2}{\sigma^2} + n\log\sigma^2 + n\log(2\pi).$$

を最小化する σ^2 は $\sigma^2 = \hat{\sigma}^2$ になることを確認できる. (一般に 函数 $\sigma^2 \mapsto a^2/\sigma^2 + \log\sigma^2$ は $\sigma^2 = a^2$ で最小になる.)

y の $X\mathbb{R}^r$ への直交射影 $\hat{y} = X\hat{\beta}$ の具体的な形を求めよう.

$y - \hat{y} = y - X\hat{\beta}$ は $X\mathbb{R}^r$ と直交するので,

$$X^T(y - X\hat{\beta}) = 0, \quad \text{すなわち} \quad X^T X\hat{\beta} = X^T y.$$

ゆえに,

$$\hat{\beta} = (X^T X)^{-1} X^T y, \quad \hat{y} = X(X^T X)^{-1} X^T y.$$

これと,

$$\hat{\sigma}^2 = \frac{1}{n}\|y - X\hat{\beta}\|^2$$

を合わせると, 尤度函数 $(\beta, \sigma^2) \mapsto p(y|X,\beta,\sigma^2)$ を最小化するパラメータ値が得られる.

x, y から $\hat{\beta} = (X^T X)^{-1} X^T y$ を求めて, $\beta = \hat{\beta}$ とおくと, 二乗和 $\|y - X\beta\|$ が最小化されるので, 回帰係数 β をその $\hat{\beta}$ として求める方法を **最小二乗法** と呼ぶ.

$y - \hat{y} = y - X\hat{\beta}$ を **残差** と呼ぶ.

2.6 βとσ²の不偏推定量

この節では, 前々節で記述した統計モデルに従う確率変数 Y とは別に, ベクトル値確率変数 $y=[y_i]_{i=1}^n$ で次の条件を満たすものを使用する:

$$y = X\beta + e, \quad E[e] = 0, \quad E[ee^T] = \sigma^2 I.$$

これはベクトル値($\mathbb{R}^n$ 値)の確率変数 $e = y - X\beta$ に関する仮定になっている. $E[ee^T] = \sigma^2 I$ という条件は e の分散共分散行列が $\sigma^2 I$ になることを意味している. このとき, y の平均と分散共分散行列はそれぞれ次のようになる:

$$E[y] = X\beta, \quad E[(y-X\beta)(y-X\beta)^T] = E[ee^T] = \sigma^2 I.$$

この設定における最小二乗法は次のように書ける:

$$\hat\beta = (X^TX)^{-1}X^Ty, \quad \hat\sigma^2 = \frac{1}{n}\|y - X\hat\beta\|^2.$$

このとき,

$$E[\hat\beta] = (X^TX)^{-1}X^TX\beta = \beta$$

なので, $\hat\beta$ は β の不偏推定量になっている.

以下では σ^2 の不偏推定量を構成する.

$\hat\beta$ の分散共分散行列は

$$\hat\beta - \beta = (X^TX)^{-1}X^T(X\beta + e) - \beta = (X^TX)^{-1}X^Te$$

より,

$$(\hat\beta-\beta)(\hat\beta-\beta)^T = (X^TX)^{-1}X^Tee^TX(X^TX)^{-1}$$

なので, 次のように計算される:

$$E[(\hat\beta-\beta)(\hat\beta-\beta)^T] = (X^TX)^{-1}X^TE[ee^T]X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.$$

2つ目の等号で $E[ee^T] = \sigma^2 I$ を使った.

$X\hat\beta$ の分散共分散行列は,

$$X\hat\beta - X\beta = X(\hat\beta - \beta), \quad (X\hat\beta - X\beta)(X\hat\beta - X\beta)^T = X(\hat\beta-\beta)(\hat\beta-\beta)^TX^T$$

より, 次のように計算される:

$$E[(X\hat\beta - X\beta)(X\hat\beta - X\beta)^T] = \sigma^2 X(X^TX)^{-1}X^T.$$

$X(X^TX)^{-1}X^T$ は $\mathbb{R}^n$ から $X\mathbb{R}^r$ への直交射影を与える行列であった. (対称行列になっていることにも注意せよ.)

残差 $y - X\hat\beta$ (これの期待値は 0 になる)の分散共分散行列は,

$$y - X\hat\beta = y - X(X^TX)^{-1}X^Ty = (I - X(X^TX)^{-1}X^T)y$$

より, 次のように計算される:

$$E[(y-X\hat\beta)(y-X\hat\beta)^T] = \sigma^2(I - X(X^TX)^{-1}X^T)(I - X(X^TX)^{-1}X^T)^T = \sigma^2(I - X(X^TX)^{-1}X^T).$$

ここで, $E[yy^T] = \sigma^2 I$ および, $X(X^TX)^{-1}X^T$ が $\mathbb{R}^n$ から $X\mathbb{R}^r$ への直交射影を与える行列であったことから, $I - X(X^TX)^{-1}X^T$ が $\mathbb{R}^n$ から $X\mathbb{R}^r$ の直交補空間への射影を与える行列であること(特に対称行列でかつ二乗しても不変なこと)を使った.

トレースの中で行列の順序を巡回的に回してもトレースの値は不変なので,

$$\operatorname{tr}(X(X^TX)^{-1}X^T) = \operatorname{tr}((X^TX)^{-1}X^TX) = \operatorname{tr}(I_r) = r.$$

ここで I_r は r 次の単位行列を表す. この結果と

$$\|y - X\hat\beta\|^2 = (y - X\hat\beta)^T(y - X\hat\beta) = \operatorname{tr}\Big((y-X\hat\beta)(y-X\hat\beta)^T\Big)$$

を使うと,

$$E[\|y - X\hat\beta\|^2] = \sigma^2\operatorname{tr}(I - X(X^TX)^{-1}X^T) = (n-r)\sigma^2.$$

ゆえに,

$$\hat{s}^2 = \frac{1}{n-r}\|y - X\hat{\beta}\|^2 = \frac{n}{n-r}\hat{\sigma}^2$$

とおくと,

$$E[\hat{s}^2] = \sigma^2.$$

すなわち, $\hat{s}^2 = \|y - X\hat{\beta}\|^2/(n-r)$ は σ^2 の不偏推定量である.

注意: 分散 σ^2 の不偏推定量を得るためには n ではなく, $n-r$ で割らなければいけない. この結果は $r=1$, $f_1(x)=1$ のとき(この場合には x_i に $f_1(x_i)=1$ に依存しないので x_i は無用になる), 独立同分布確率変数達 $y_1,\dots,y_n$ の不偏分散が $n-1$ で割ることによって構成されることを含んでいる. 次の節を見よ.

2.7 例: 平均の推定の場合

線形回帰は平均の推定を含む.

$r=1$ とし, β_i のインデックス i を 1 とせずに 0 と書くことにし,

$$f_0(x_i) = 1, \quad \beta_0 = \mu, \quad \hat{\beta}_0 = \hat{\mu}$$

の場合について考える. このとき,

$$X = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad X\beta = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix}.$$

ゆえに, $X^T X = n$, $X^T y = \sum_{i=1}^n y_i$ より,

$$\hat{\mu} = (X^T X)^{-1} X^T y = \frac{1}{n}\sum_{i=1}^n y_i =: \bar{y}.$$

さらに, $y - X\hat{\beta} = [y_i - \hat{\mu}]_{i=1}^n = [y_i - \bar{y}]_{i=1}^n$ より,

$$\hat{\sigma}^2 = \frac{1}{n}\|y - X\hat{\beta}\| = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2.$$

これは, σ^2 の不偏推定量になるように $n-1$ で割らずに, n で割った場合の標本分散になっている. 前節で示したように σ^2 の不偏推定量 $\hat{s}^2$ は, $n-1$ で割ることによって次のようにして得られる:

$$\hat{s}^2 = \frac{1}{n-1}\|y - X\hat{\beta}\| = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2.$$

2.8 例: 単回帰の場合

$r=2$ とし, β_i のインデックス i を $1,2$ とせずに $0,1$ とすることにし,

$$f_0(x_i) = 1, \quad f_1(x_i) = x_i$$

の場合を考える. さらに, 以下のようにおく:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i,$$
$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2, \quad \overline{y^2} = \frac{1}{n}\sum_{i=1}^n y_i^2, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^n x_i y_i.$$

このとき, 偏補正をしていない x_i 達と y_i 達の標本分散と標本共分散はそれぞれ次のように書ける:

$$\hat{\sigma}_x^2 = \overline{x^2} - \bar{x}^2, \quad \hat{\sigma}_y^2 = \overline{y^2} - \bar{y}^2, \quad \hat{\sigma}_{xy} = \overline{xy} - \bar{x}\bar{y}. \qquad (*)$$

デザイン行列 X については, 次が成立する:

$$
X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad X^T X = n \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix},
$$

$$
X^T y = n \begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix}, \quad (X^T X)^{-1} = \frac{1}{n(\overline{x^2} - \bar{x}^2)} \begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.
$$

ゆえに,

$$
\overline{x^2}\bar{y} - \bar{x}\overline{xy} = \overline{x^2}\bar{y} - \bar{x}^2\bar{y} + \bar{x}^2\bar{y} - \bar{x}\overline{xy} = \hat{\sigma}_x^2\bar{y} - \hat{\sigma}_{xy}\bar{x}
$$

が成立していることに注意すれば,

$$
\hat{\beta} = [\hat{\beta}_j]_{j=0}^1 = (X^T X)^{-1} X^T y = \begin{bmatrix} \bar{y} - (\hat{\sigma}_{xy}/\hat{\sigma}_x^2)\bar{x} \\ \hat{\sigma}_{xy}/\hat{\sigma}_x^2 \end{bmatrix}.
$$

すなわち, 次の回帰直線の公式が得られた:

$$
\hat{\beta}_0 + \hat{\beta}_1 x_* = \bar{y} + \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}(x_* - \bar{x}).
$$

$\hat{\sigma}^2$ の公式も求めよう.

$$
y - X\hat{\beta} = (I - X(X^T X)^{-1} X^T) y
$$

より,

$$
\begin{aligned}
\|y - X\hat{\beta}\|^2 &= y^T (I - X(X^T X)^{-1} X^T)^T (I - X(X^T X)^{-1} X^T) y \\
&= y^T (I - X(X^T X)^{-1} X^T) y \\
&= y^T y - (X^T y)^T (X^T X)^{-1} X^T y.
\end{aligned}
$$

そして,

$$
\begin{aligned}
& y^T y = n\overline{y^2}, \\
& (X^T y)^T (X^T X)^{-1} X^T y = \frac{n}{\hat{\sigma}_x^2}\left(\overline{x^2} - 2\bar{x}\bar{y}\overline{xy} + \overline{xy}^2\right)
\end{aligned}
$$

および上の(*)を使って整理すると,

$$
\hat{\sigma}^2 = \frac{1}{n}\|y - X\hat{\beta}\|^2 = \frac{\hat{\sigma}_x^2\hat{\sigma}_y^2 - \hat{\sigma}_{xy}^2}{\hat{\sigma}_x^2}.
$$

前々節の結果より, σ^2 の不偏推定量 $\hat{s}^2$ は n ではなく, $n - r = n - 2$ で割ることによって次のようにして得られる:

$$
\hat{s}^2 = \frac{1}{n-2}\|y - X\hat{\beta}\|^2 = \frac{n}{n-2}\frac{\hat{\sigma}_x^2\hat{\sigma}_y^2 - \hat{\sigma}_{xy}^2}{\hat{\sigma}_x^2}.
$$

x_i 達と y_i 達の不偏分散と不偏共分散を

$$
s_x^2 = \frac{n}{n-1}\hat{\sigma}_x^2, \quad s_y^2 = \frac{n}{n-1}\hat{\sigma}_y^2, \quad s_{xy} = \frac{n}{n-1}\hat{\sigma}_{xy}
$$

と書くと,

$$
\begin{aligned}
& \beta_0 = \bar{y} - \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}\bar{x} = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}, \quad \beta_1 = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \frac{s_{xy}}{s_x^2}, \\
& \hat{\sigma}^2 = \frac{n-1}{n}\frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2}, \quad \hat{s}^2 = \frac{n-1}{n-2}\frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2}.
\end{aligned}
$$

以上の公式は

- [「標本分布について」のノート (https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/04%20Distribution%20of%20samples.ipynb)](https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/04%20Distribution%20of%20samples.ipynb)

の「最小二乗法による線形回帰」の節で得た公式に一致する.

2.9 多変量正規分布の定義

$\mu \in \mathbb{R}^n$ と固有値がすべて正の n 次の実対称行列 Σ に対して, 多変量正規分布

$$\mathrm{MvNormal}(\mu, \Sigma)$$

の確率密度函数が

$$p(y|\mu,\Sigma) = \frac{1}{\det(2\pi\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right) \quad (y \in \mathbb{R}^n)$$

と定義される.

例えば, $\mu = (m, m, \dots, m)$, $\Sigma = \sigma^2 I$ (I は n 次の単位行列で $\sigma^2 > 0$)のとき,

$$p(y|\mu,\Sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - m)^2\right).$$

これは正規分布の標本分布 $\mathrm{Normal}(m, \sigma^2)^n$ の密度函数に等しい.

2.10 問題: χ²分布

前節の n 変量正規分布 $\mathrm{MvNormal}(\mu, \Sigma)$ を考える:

$$p(y|\mu,\Sigma) = \frac{1}{\det(2\pi\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right) \quad (y \in \mathbb{R}^n).$$

この分布に従う n 次元ベクトル値確率変数を Y と書き, 確率変数 χ^2 を

$$\chi^2 = (Y-\mu)^T\Sigma^{-1}(Y-\mu)$$

と定めると, χ^2 は自由度 n のχ²分布に従うことを示せ.

解答例: 実対称行列 Σ は直交行列 U で対角化できる.

$$\Sigma = UD^2U^{-1} = UD^2U^T, \quad D = \mathrm{diag}(\sigma_1, \dots, \sigma_n), \quad \sigma_i > 0.$$

このとき,

$$\det(2\pi\Sigma)^{1/2} = \prod_{i=1}^n (2\pi\sigma_i^2)^{1/2} = (2\pi)^{n/2}\sigma_1\cdots\sigma_n.$$

さらに, $x = [x_i]_{i=1}^n = D^{-1}U^T(y-\mu)$ とおくと, $\Sigma^{-1} = UD^{-2}U^T$ より

$$(y-\mu)^T\Sigma^{-1}(y-\mu) = x^Tx$$

となり, 直交行列による変換が体積を保つことより,

$$|dy_1\cdots dy_n| = \sigma_1\cdots\sigma_n\,|dx_1\cdots dx_n|$$

となるので,

$$\begin{aligned}
p(y|\mu,\Sigma)\,|dy_1\cdots dy_n| &= \frac{1}{(2\pi)^{n/2}\sigma_1\cdots\sigma_n}\exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right)|dy_1\cdots dy_n| \\
&= \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}x^Tx\right)|dx_1\cdots dx_n| \\
&= \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right)|dx_1\cdots dx_n|.
\end{aligned}$$

ゆえに, 変数 x に対応するベクトル値確率変数を $X = [X_i]_{i=1}^n$ と書くと, $X_1, \dots, X_n$ はそれぞれが標準正規分布に従う独立な確率変数達になり,

$$\chi^2 = (Y-\mu)^T\Sigma^{-1}(Y-\mu) = X^TX = \sum_{i=1}^n X_i^2$$

となるので, χ^2 は自由度 n のχ²分布に従う.

解答終

2.11 真の回帰函数と推定された回帰函数

「βとσ²の不偏推定量」の節の仮定の下で, $x_* \in \mathbb{R}$ の函数

$$f(x_*)^T\beta = \sum_{j=1}^r \beta_j f_j(x_*)$$

を **真の回帰函数** と呼び,

$$f(x_*)^T\hat{\beta} = \sum_{j=1}^r \hat{\beta}_j f_j(x_*)$$

を **推定された回帰函数** と呼ぶことにする.

注意: 真の回帰函数の「真の」の意味は「現実における真の」という意味ではない.

2.12 信頼区間 (標準正規分布版)

この節では, 「βとσ²の不偏推定量」の節で仮定した条件に加えて, $\hat{\beta} = (X^TX)^{-1}X^Ty$ が近似的に多変量正規分布に従い, $\hat{s}^2 \approx \sigma^2$ となっていると仮定する.

このとき, βとσ²の不偏推定量」の節で示した結果

$$E[\hat{\beta}] = \beta, \quad E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T] = \sigma^2(X^TX)^{-1}$$

より, 次の近似が成立している:

$$\hat{\beta} \sim \mathrm{MvNormal}(\beta,\ \hat{s}^2(X^TX)^{-1}),\ \text{approximately}.$$

ゆえに, $x_* \in \mathbb{R}$ について,

$$f(x_*)^T\hat{\beta} \sim \mathrm{Normal}\Big(f(x_*)^T\beta,\ \widehat{\mathrm{SE}}_{f(x_*)^T\hat{\beta}}\Big),\ \text{approximately}.$$

ここで,

$$\widehat{\mathrm{SE}}_{f(x_*)^T\hat{\beta}} = \hat{s}\sqrt{f(x_*)^T(X^TX)^{-1}f(x_*)}$$

これより, 真の回帰函数の $x_* \in \mathbb{R}$ の値 $f(x_*)^T\beta$ に関する仮説 $f(x_*)^T\beta = f(x_*)^T\beta_0$ のP値と $f(x_*)^T\beta$ の値の信頼区間を定義できる.

仮説 $f(x_*)^T\beta = f(x_*)^T\beta_0$ のP値の定義: $t(x_*, \beta_0)$ を

$$t(x_*, \beta_0) = \frac{f(x_*)^T\hat{\beta} - f(x_*)^T\beta_0}{\widehat{\mathrm{SE}}_{f(x_*)^T\hat{\beta}}} = \frac{f(x_*)^T\hat{\beta} - f(x_*)^T\beta_0}{\hat{s}\sqrt{f(x_*)^T(X^TX)^{-1}f(x_*)}}$$

と定め, P値を次のように定める:

$$\mathrm{pvalue}_{\mathrm{Normal}}(y|X, f(x_*)^T\beta = f(x_*)\beta_0) = 2(1 - \mathrm{cdf}(\mathrm{Normal}(0,1), |t(x_*, \beta_0)|)).$$

真の回帰函数上の値 $f(x_*)^T\beta$ の信頼度 $1-\alpha$ の信頼区間の定義: $z_{\alpha/2}$ を

$$z_{\alpha/2} = \mathrm{quantile}(\mathrm{Normal}(0,1), 1-\alpha/2).$$

と定め, 信頼区間を次のように定める:

$$\mathrm{confint}^{f(x_*)^T\beta}_{\mathrm{Normal}}(y|X) = \Big[f(x_*)^T\hat{\beta} - z_{\alpha/2}\widehat{\mathrm{SE}}_{f(x_*)^T\hat{\beta}},\ f(x_*)^T\hat{\beta} + z_{\alpha/2}\widehat{\mathrm{SE}}_{f(x_*)^T\hat{\beta}}\Big].$$

2.13 信頼区間 (t分布版)

前節の標準正規分布を使って定義されたP値と信頼区間の t 分布を使った補正を構成しよう.

そのために, 前節までに仮定した条件よりもさらに強い以下の条件を仮定する:

$$e = [e_i]_{i=1}^n \sim \mathrm{Normal}(0, \sigma)^n = \mathrm{MvNormal}(0, \sigma^2 I).$$

このとき,

$$\hat\beta \sim \operatorname{MvNormal}(\beta,\ \sigma^2(X^TX)^{-1}).$$

ゆえに, $x_* \in \mathbb{R}$ について,

$$f(x_*)^T\hat\beta \sim \operatorname{Normal}\Big(f(x_*)^T\beta,\ \operatorname{SE}_{f(x_*)^T\hat\beta}\Big).$$

ここで,

$$\operatorname{SE}_{f(x_*)^T\hat\beta} = \sigma\sqrt{f(x_*)^T(X^TX)^{-1}f(x_*)}.$$

そして,

$$\frac{(n-r)\hat s^2}{\sigma} = \|y - X\hat\beta\|^2 \sim \operatorname{Chisq}(n-r)$$

となることを示せる.

$\hat y = X\hat\beta$ と $y - \hat y = y - X\hat\beta$ が独立であることより, $f(x_*)^T\hat\beta$ と $(n-r)\hat s^2/\sigma = \|y - X\hat\beta\|^2$ は独立になるので,

$$t(x_*, \beta) = \frac{f(x_*)^T\hat\beta - f(x_*)^T\beta}{\widehat{\operatorname{SE}}_{f(x_*)^T\hat\beta}} = \frac{f(x_*)^T\hat\beta - f(x_*)^T\beta}{\hat s\sqrt{f(x_*)^T(X^TX)^{-1}f(x_*)}} \sim \operatorname{TDist}(n-r).$$

これを使うと t 分布を使って補正したP値と信頼区間を以下のように定義できる:

仮説 $f(x_*)^T\beta = f(x_*)^T\beta_0$ のP値の定義: $t(x_*, \beta_0)$ を

$$t(x_*, \beta_0) = \frac{f(x_*)^T\hat\beta - f(x_*)^T\beta_0}{\widehat{\operatorname{SE}}_{f(x_*)^T\hat\beta}} = \frac{f(x_*)^T\hat\beta - f(x_*)^T\beta_0}{\hat s\sqrt{f(x_*)^T(X^TX)^{-1}f(x_*)}}$$

と定め, P値を次のように定める:

$$\operatorname{pvalue}_{\mathrm{TDist}}(y|X, f(x_*)^T\beta = f(x_*)\beta_0) = 2(1 - \operatorname{cdf}(\operatorname{TDist}(n-r), |t(x_*, \beta_0)|)).$$

真の回帰函数上の値 $f(x_*)^T\beta$ の信頼度 $1-\alpha$ の信頼区間の定義: $t_{\nu,\alpha/2}$ を

$$t_{\nu,\alpha/2} = \operatorname{quantile}(\operatorname{TDist}(\nu), 1 - \alpha/2).$$

と定め, 信頼区間を次のように定める:

$$\operatorname{confint}^{f(x_*)^T\beta}_{\mathrm{TDist}}(y|X) = \left[f(x_*)^T\hat\beta - t_{n-r,\alpha/2}\widehat{\operatorname{SE}}_{f(x_*)^T\hat\beta},\ f(x_*)^T\hat\beta + t_{n-r,\alpha/2}\widehat{\operatorname{SE}}_{f(x_*)^T\hat\beta}\right].$$

2.14 予測区間

以上の仮定に加えて、さらに

$$\begin{bmatrix} e \\ e_* \end{bmatrix} \sim \operatorname{Normal}(0, \sigma)^{n+1}$$

と仮定し, $x_* \in \mathbb{R}$ を取り,

$$y_* = f(x_*)^T\beta + e_*$$

とおく. このとき,

$$y_* - f(x_*)^T\hat\beta = f(x_*)^T\beta - f(x_*)^T\hat\beta + e_* \sim \operatorname{Normal}\Big(0, \operatorname{SE}_{y_* - f(x_*)^T\hat\beta}\Big).$$

ここで,

$$\operatorname{SE}_{y_* - f(x_*)^T\hat\beta} = \sigma\sqrt{1 + f(x_*)^T(X^TX)^{-1}f(x_*)}.$$

前節の $\operatorname{SE}_{f(x_*)^T\hat\beta}$ との違いは平方根の中に e_* の分散から出て来た 1 が含まれていることである.

これより,

$$\widehat{\operatorname{SE}}_{y_* - f(x_*)^T\hat\beta} = \hat s\sqrt{1 + f(x_*)^T(X^TX)^{-1}f(x_*)}.$$

とおくと.

$$\frac{y_* - f(x_*)^T\hat{\beta}}{\widehat{\mathrm{SE}}_{y_* - f(x_*)^T\hat{\beta}}} \sim \mathrm{TDist}(n-r).$$

これを使って, y_* の **予測区間** (prediction interval)を次のように定義できる:

$$\mathrm{predint}^{y_*}_{\mathrm{TDist}}(y|X) = \left[f(x_*)^T\hat{\beta} - t_{n-r,\alpha/2}\widehat{\mathrm{SE}}_{y_*-f(x_*)^T\hat{\beta}},\ f(x_*)^T\hat{\beta} + t_{n-r,\alpha/2}\widehat{\mathrm{SE}}_{y_*-f(x_*)^T\hat{\beta}}\right].$$

前節の $\widehat{\mathrm{SE}}_{f(x_*)^T\hat{\beta}}$ と $\widehat{\mathrm{SE}}_{y_*-f(x_*)^T\hat{\beta}}$ の違いは, 後者の定義式の平方根中に 1 が含まれていることである. だから, 予測区間は信頼区間よりも必ず広くなる. そうなる理由は y_* の定義を見れば明らかで, y_* の定義にはノイズの項 e_* が含まれている. その分だけ区間の幅が広くなる.

3 線形回帰の計算例

4 ロジスティック回帰

In [ ]:

```

```