L03 2 次元正規分布・単回帰

樋口さぶろお

龍谷大学 先端理工学部 数理・情報科学課程

理論物理学特論　L03(2021-10-16 Sat)

最終更新: Time-stamp: "2021-10-12 Tue 16:20 JST hig"

今日の目標

- 指定の母平均値と母 (共) 分散の 2 次元正規分布の確率密度関数が書ける

再生性 reproductive property

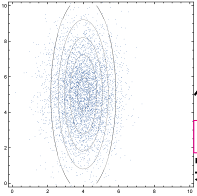
正規分布の再生性 岩薩林 確率・統計定理 4.3

独立な連続型確率変数 $X \sim \mathrm{N}(\mu_\mathrm{X}, \sigma_\mathrm{X}^2)$, $Y \sim \mathrm{N}(\mu_\mathrm{Y}, \sigma_\mathrm{Y}^2)$ に対して,

$$S = aX + bY \sim \mathrm{N}(a\mu_\mathrm{X} + b\mu_\mathrm{Y}, a^2\sigma_\mathrm{X}^2 + b^2\sigma_\mathrm{Y}^2)$$

特に,

$$S = X + Y \sim \mathrm{N}(\mu_\mathrm{X} + \mu_\mathrm{Y}, \sigma_\mathrm{X}^2 + \sigma_\mathrm{Y}^2)$$

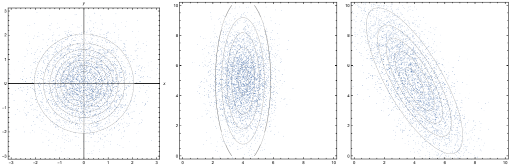


再生性とは, 独立な確率変数 X, Y の和 $S = X + Y$ が (パラメタは違うけど) 同じ分布 (今なら正規分布) にしたがうという, 稀なよい性質.
他に, 二項分布, ポアソン分布など.
岩薩林 確率・統計例題 4.14 特定の a, b, μ, σ での証明
実は, 一般の (独立とかぎらない)2 次元正規分布に対しても $aX + bY$ は 1 次元正規分布にしたがうことがわかる).

ここまで来たよ

2. 2 次元の正規分布
   - 2 次元正規分布 (X, Y が独立でない)
3. 線形回帰モデル (単回帰)
   - 最尤推定
   - 線形回帰モデル
   - Jupyter Notebook で実行

等高線が傾いた楕円の分布も作りたい!



考え方 1
等高線が円 $D = x^2 + y^2$
等高線が傾いた楕円 $D = x^2 + cxy + 2y^2$　　2 次形式の標準化 (線形代数 I,II)
考え方 2 独立じゃなくしたいから, 指数関数の引数に xy の項でもいれておけば?

一般の 2 次元正規分布

$$f(x,y) = \text{定数}' \times \mathrm{e}^{-ax^2-by^2+cxy+px+qy}$$

$\rightsquigarrow$

一般の 2 次元正規分布の同時確率密度関数

2 次元正規分布 $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ の同時確率密度関数は, 確率変数を $\boldsymbol{X} = {}^{\mathrm{t}}(X, Y)$ とするとき,

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{2/2}\sqrt{\det\Sigma}}\mathrm{e}^{-\frac{1}{2}{}^{\mathrm{t}}(\boldsymbol{x}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$\text{母平均値 (ベクトル)}\boldsymbol{\mu} = \begin{pmatrix} \mu_{\mathrm{X}} \\ \mu_{\mathrm{Y}} \end{pmatrix} \overset{\text{要証明}}{=} \begin{pmatrix} \mathrm{E}[X] \\ \mathrm{E}[Y] \end{pmatrix}$$

$$\text{母共分散行列}\Sigma = \begin{pmatrix} \sigma_{\mathrm{X}}^2 & C_{\mathrm{XY}} \\ C_{\mathrm{XY}} & \sigma_{\mathrm{Y}}^2 \end{pmatrix} \overset{\text{要証明}}{=} \begin{pmatrix} \mathrm{V}[X] & \mathrm{Cov}[X,Y] \\ \mathrm{Cov}[X,Y] & \mathrm{V}[Y] \end{pmatrix}$$

L03-Q1

Quiz(2 次元正規分布の確率密度関数)

2 次元正規分布の同時確率密度関数で,

$\boldsymbol{\mu} = \begin{pmatrix} -4 \\ 5 \end{pmatrix}$,

$\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$

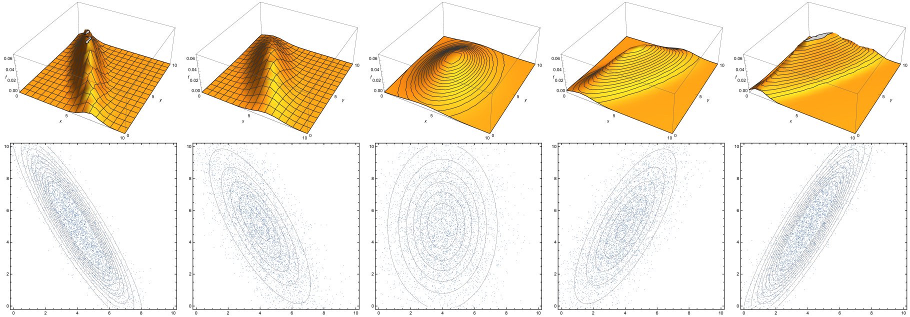
のとき,

1. 母平均値, 母 (共) 分散を答えよう.
2. 同時確率密度関数 f を (行列やベクトルを使わずに) 具体的に書こう.

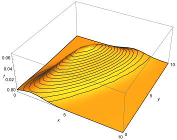
母相関係数 r, 共分散行列 Σ. $c = \frac{9\sqrt{21}}{10}$

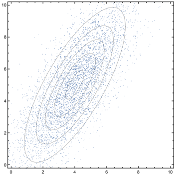
$r = -0.90$　　-0.76　　0　　$+0.76$　　$+0.90$

$$\Sigma = \begin{pmatrix} 3 & -c \\ -c & 7 \end{pmatrix}, \begin{pmatrix} 3 & -2\sqrt{3} \\ -2\sqrt{3} & 7 \end{pmatrix}, \begin{pmatrix} 3 & 0 \\ 0 & 7 \end{pmatrix}, \begin{pmatrix} 3 & +2\sqrt{3} \\ +2\sqrt{3} & 7 \end{pmatrix}, \begin{pmatrix} 3 & +c \\ +c & 7 \end{pmatrix}$$



2 次元正規分布の周辺分布は 1 次元正規分布





このほか, $X = x_0$ の条件付き分布, $S = aX + bY$ の分布も 1 次元正規分布 $\rightarrow$ 再生性

確率密度関数から母ナントカ I

L03-Q2

Quiz(2 次元正規分布)

次の 2 変数確率密度関数は 2 次元正規分布を定める.

$$f(x,y) = C \cdot \mathrm{e}^{-2x^2+6xy-7y^2+26x-54y-7}.$$

1. 母平均値, 母共分散行列を求めよう.
2. $\mathrm{E}[1] = 1$ が満たされるように定数 C を定めよう.

ここまで来たよ

2 2 次元の正規分布
- 2 次元正規分布 (X, Y が独立でない)

3 線形回帰モデル (単回帰)
- 最尤推定
- 線形回帰モデル
- Jupyter Notebook で実行

線形モデル (統計モデルのある一族)

あるドーナツ製造機の作るドーナツの重さ Y は次のモデルに従う.

$$Y = \beta_0 + \epsilon, \qquad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

Y, ϵ: 連続型確率定数, $\sigma > 0, \beta_0$:定数=パラメタ=母数

ϵ: 誤差, ノイズ. 小文字だけど確率変数.
(隠された) パラメタ母数 μ, σ^2 を, n 個のドーナツの重さのデータから推定したい.

$$\begin{aligned} \mathrm{E}[Y] =& \beta_0 + \mathrm{E}[\epsilon] = \beta_0, \\ \mathrm{V}[Y] =& \qquad \mathrm{V}[\epsilon] = \sigma^2. \end{aligned}$$

正規分布と限定した以外は, ここしばらくやってた, 母平均値, 母分散の推定の言い換えに過ぎない.
だけど, 多数のパラメタを含む一般的なモデルにも使える考え方をする.

尤度 likelihood

$\epsilon = Y - \beta_0 \sim \mathrm{N}(0, \sigma^2)$ より，ドーナツの重さ y を得る確率密度は，

$$f(y|\beta_0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y-\beta_0)^2}{2\sigma^2}}$$

サイズ n の標本が $y_1, \cdots, y_n$ である確率密度は，独立同分布なので積で，

$$\begin{aligned} f(y_1, y_2, \cdots, y_n|\beta_0, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y_i-\beta_0)^2}{2\sigma^2}} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \mathrm{e}^{-\sum_{i=1}^{n} \frac{(y_i-\beta_0)^2}{2\sigma^2}} \end{aligned}$$

n 次元正規分布

多変量解析及び演習

この f を，y_i の確率密度関数と思わず，が測定済データ $y_1, \ldots, y_n$ が定数，β_0, σ が変数と思ったとき，尤度 (ゆうど) 関数 $L(\beta_0, \sigma)$ という．

$$L(\beta_0, \sigma) = f(y_1, y_2, \cdots, y_n|\beta_0, \sigma)$$

最尤推定 岩薩林 確率・統計なし

最尤推定

β_0, σ の推定値として $L(\beta_0, \sigma)$ が最大になる値を選ぶ

2 変数関数の最大値 $\rightsquigarrow$ 偏微分　　微積分 II

$$0 = \frac{\partial L}{\partial \beta_0}(\beta_0, \sigma) = \frac{\partial L}{\partial \sigma}(\beta_0, \sigma)$$

ここでは最初の等式だけ解く．合成微分.

$$0 = (\text{定数})^{-n/2} \times \sum_{i=1}^{n} \frac{1}{\sigma^2}(y_i - \beta_0) \times \mathrm{e}^{\text{同じ}}$$

$$0 = \sum_{i=1}^{n}(y_i - \beta_0)$$

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$$

ここでは最初の式だけ.
合成微分.
β_0(母平均値) の最尤推定 (MLE=maximum likelihood estimation) 値 $\hat{\beta}_0$ は標本平均値じゃん

ここまで来たよ

2. 2 次元の正規分布
   - 2 次元正規分布 (X, Y が独立でない)
3. 線形回帰モデル (単回帰)
   - 最尤推定
   - 線形回帰モデル
   - Jupyter Notebook で実行

機械学習としての線形回帰

このドーナツ製造機の作るドーナツの重さ Y は，温度 x によるらしい．
次の線形回帰モデルを仮定する．

$$Y = \beta_0 + \beta_1 \cdot x + \epsilon, \qquad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

Y, ϵ: 連続型確率定数，β_0, β_1:回帰係数，$\sigma > 0$:定数=パラメタ=母数

Y: 目的変数 (従属変数) ここでは確率変数
x: 説明変数 (独立変数) ここでは確率変数でない
これは教師あり学習の一種
出力が連続的な値，予測器が 1 次関数．
(x,y) の n 個の訓練データ $\overset{\text{学習}}{\to}$ パラメタの推定値 $\hat{\beta}_0, \hat{\beta}_1$ $\to$ 予測器
$y = f(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
性能評価
機械学習のりなら，訓練データと分けておいたテストデータを使うところだが，統計学のりでは，数学的仮定に基づき，決定係数 R^2 で評価 (訓練

母平均値の推定と似ている/違う

(確率変数でない) 変数 x に依存する確率変数 Y

永田棟方 多変量解析法入門 §4

このドーナツ製造機の作るドーナツの重さ Y は, 温度 x によるらしい. 次の線形回帰モデルを仮定する.

$$Y = \beta_0 + \beta_1 \cdot x + \epsilon, \qquad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

Y, ϵ: 連続型確率定数, β_0, β_1:回帰係数, $\sigma > 0$:定数=パラメタ=母数

Y: 目的変数 (従属変数) ここでは確率変数
x: 説明変数 (独立変数) ここでは確率変数でない

ノイズ・誤差 $\epsilon = Y - \beta_0 + \beta_1 \cdot x \sim \mathrm{N}(0, \sigma^2)$.
$\epsilon = Y - \beta_0 - \beta_1 x \sim \mathrm{N}(0, \sigma^2)$ より, ドーナツの重さ y を得る確率密度は,

$$f(y|x, \beta_0, \beta_1, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y-\beta_0-\beta_1 \cdot x)^2}{2\sigma^2}}$$

(正確に) 指定した温度 x_i $(i=1,\dots,n)$ で製造したときの重さが y_i $(i=1,\dots,n)$ である確率密度は, 独立分布なので積.

$$
\begin{aligned}
f(y_1, y_2, \cdots, y_n | x_1, \dots, x_n, \beta_0, \beta_1, \sigma) =& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y_i-\beta_0-\beta_1\cdot x)^2}{2\sigma^2}} \\
=& \left(2\pi\sigma^2\right)^{-n/2} \mathrm{e}^{-\sum_{i=1}^{n}\frac{(y_i-\beta_0-\beta_1\cdot x_i)^2}{2\sigma^2}} \\
=& L(\beta_0, \beta_1, \sigma)
\end{aligned}
$$

最尤推定

測定済のデータ x_i, y_i $i = 1, \cdots, n$ を定数と思ったときの, 3 変数関数
$L(\beta_0, \beta_1, \sigma) = f(y_1, y_2, \cdots, y_n | x_1, \ldots, x_n, \beta_0, \beta_1, \sigma)$ の最大値は? $\rightsquigarrow$ 偏微分 微積分 II

$$0 = \frac{\partial L}{\partial \beta_0}(\beta_0, \beta_1, \sigma) = \frac{\partial L}{\partial \beta_1}(\beta_0, \beta_1, \sigma) = \frac{\partial L}{\partial \sigma}(\beta_0, \beta_1, \sigma)$$

ここでは最初の 2 つの等式だけ解く.

$$0 = \frac{\partial L}{\partial \beta_0} = (\text{定数}) \sum_i \frac{1}{\sigma^2}(y_i - \beta_0 - \beta_1 x_i) \times \mathrm{e}^{\text{同じ}}$$

$$0 = \frac{\partial L}{\partial \beta_0} = (\text{定数}) \sum_i \frac{1}{\sigma^2} x_i (y_i - \beta_0 - \beta_1 x_i) \times \mathrm{e}^{\text{同じ}}$$

しょせん, β_0, β_1 の連立 1 次方程式

正規方程式

岩薩林 確率・統計 §9.2

$$n\beta_0 + (\sum_i x_i)\beta_1 = \sum_i y_i$$

$$(\sum_i x_i)\beta_0 + (\sum_i x_i^2)\beta_1 = \sum_i x_i y_i$$

加減法 $\rightsquigarrow$

$$\hat\beta_1 = \frac{s_{xy}}{s_{xx}}, \qquad y = \frac{s_{xy}}{s_{xx}}x + \overline{y} - \frac{s_{xy}}{s_{xx}}\overline{x}$$

$$\hat\beta_0 = \overline{y} - \frac{s_{xy}}{s_{xx}}\overline{x} \qquad y - \overline{y} = \frac{s_{xy}}{s_{xx}}(x - \overline{x})$$

ここで,

$$\overline{x} = \frac{1}{n}\sum_i x_i$$ 平均値っぽい形

$$\overline{y} = \frac{1}{n}\sum_i y_i$$ 平均値っぽい形

$$s_{xy} = \frac{1}{n}\sum_i x_i y_i - \overline{x}\cdot\overline{y} \overset{\boxed{\text{岩薩林 確率・統計定理 1.5}}}{=} \frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y})$$ 共分散ぽい

$$s_{xx} = \frac{1}{n}\sum_{i=1}^n x_i^2 - \overline{x}^2 \overset{\boxed{\text{岩薩林 確率・統計定理 1.2}}}{=} \frac{1}{n}\sum_i (x_i - \overline{x})^2$$ 分散ぽい

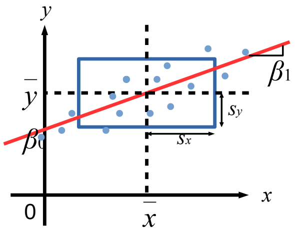
$$Y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \epsilon$$

回帰直線 岩薩林 確率・統計 §9.1

推定結果 β_0, β_1 を係数とする xy 平面の直線

$$y - \overline{y} = \frac{s_{xy}}{s_{xx}}(x - \overline{x})$$

β_0, β_1 回帰係数

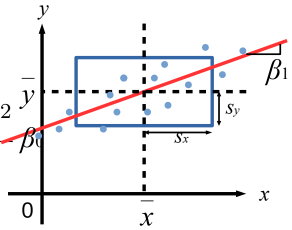


回帰係数, 予測値の信頼区間

Y_i は確率変数.
こうやって得られる推定結果を $\hat{\beta}_0, \hat{\beta}_1$ と書くと, これらも確率変数.
$\hat{\beta}_0, \hat{\beta}_1$ にも信頼区間などが考えられる. さらに, (自分が新たに設定した温度 x に対する) 予測値 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ にも信頼区間などが考えられる.

決定係数 岩薩林 確率・統計 §9.2

残差　直線 (予測値) からの上下方向のずれ,
$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.



決定係数

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2} = 1 - \frac{\sum_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\sum_i (y_i - \overline{y})^2}$$

$0 \le R^2 \le 1$ で, $1 - R^2$ が 0 に近いほど, あてはまりがよい.
$1 - R^2$ は, $\frac{\text{予測からのデータのずれ}}{\text{単一平均値 } \overline{y} \text{ からのデータのずれ}}$, つまり, $\frac{\text{回帰直線からの誤差}}{\text{平均値からの誤差}}$.
実は R^1 は, 2 次元データ (x_i, y_i) の相関係数 岩薩林 確率・統計 p.22 　データ分析

岩薩林 確率・統計例題 9.2, 9.3, §9 問題 3,4,5, §9 練習問題 1

L03-Q3

Quiz(回帰係数と回帰直線)

x を説明変数, y を目的変数とする線形モデル $Y = \beta_0 + \beta_1 \cdot x + \epsilon$, $\epsilon \sim \mathrm{N}(0, \sigma^2)$ において, x を $x = x_1, \ldots, x_n$ と変えて Y を測定したところ, $y_1, \ldots, y_n$ となった. それをまとめたのが下である.

(x, y) のデータの個数 n	16
$\overline{x} = \frac{1}{n}\sum_i x_i$	9
$\overline{y} = \frac{1}{n}\sum_i y_i$	-4
$\frac{1}{n}\sum_i x_i^2 - \overline{x}^2$	49
$\frac{1}{n}\sum_i y_i^2 - \overline{y}^2$	36
$\frac{1}{n}\sum_i x_i y_i - \overline{x} \cdot \overline{y}$	-25

このとき, 回帰直線の式を, x, y で書こう. 整理しなくてよい.

ここまで来たよ

2 2 次元の正規分布
- 2 次元正規分布 (X, Y が独立でない)

3 線形回帰モデル (単回帰)
- 最尤推定
- 線形回帰モデル
- Jupyter Notebook で実行

Jupyter Notebook で実行

```
[ ]
formula='height ~ weight'  # formula 表記. ~ の左辺を目的変数, 右辺を説明変数として線形回
result=smf.ols(formula, body).fit() # ols = 普通の最小二乗法で fit せよ
result.summary() # result に蓄えられた結果を取り出す
```

目的変数は height

決定係数R
(Adjusted 自由度調整済)

OLS Regression Results

Dep. Variable:	height	**R-squared:**	0.905
Model:	OLS	**Adj. R-squared:**	0.904
Method:	Least Squares	**F-statistic:**	935.3
Date:	Sat, 09 Oct 2021	**Prob (F-statistic):**	6.31e-52
Time:	00:59:55	**Log-Likelihood:**	-255.30
No. Observations:	100	**AIC:**	514.6
Df Residuals:	98	**BIC:**	519.8
Df Model:	1		
Covariance Type:	nonrobust		

データの個数
自由度

F統計量
F検定のp値
(有意確率)

値　標準誤差　t検定のp値 (有意確率)　95%信頼区間

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	129.4555	1.122	115.412	0.000	127.230	131.681
weight	0.8233	0.027	30.582	0.000	0.770	0.877

切片β0
説明変数weightの係数β1

Omnibus:	1.147	**Durbin-Watson:**	2.218
Prob(Omnibus):	0.564	**Jarque-Bera (JB):**	1.053
Skew:	-0.063	**Prob(JB):**	0.591
Kurtosis:	2.514	**Cond. No.**	149.