線形回帰モデルの推定

樋口さぶろお　https://hig3.net

龍谷大学 先端理工学部 数理・情報科学課程

確率統計 I　L12(2021-06-30 Wed)

最終更新: Time-stamp: "2021-06-30 Wed 18:23 JST hig"

今日の目標

- 回帰分析を確率変数の言葉で説明できる
- 隠されたパラメタを最尤推定できる
- 回帰分析で回帰係数を推定できる 岩薙林 確率・統計 §9
- 多次元の確率変数の推定ができる

L11-Q1

Quiz 解答:カイ二乗分布の確率と $\chi^2_k(\alpha)$

1. $z(()0.025) = 1.960.$
2. 標準正規分布の確率密度関数は偶関数なので,
   $z(1 - 0.025) = -z(0.025) = -1.960.$
3. $\chi^2_1(0.05) = 3.841.$ 別解. $0.05 = P(W > w_0) = P(Z^2 > w_0) = P(Z < -\sqrt{w_0}$ or $Z > +\sqrt{w_0}) = 2 \times P(Z > \sqrt{w_0})$. よって,
   $\sqrt{w_0} = 1.960.$
4. $\chi^2_1(1 - 0.05) = 0.00393.$

L11-Q2

Quiz 解答:母分散の区間推定

標本サイズは $n = 9$, 自由度は $9 - 1$, 母分散 σ^2 の信頼係数 0.95 の信頼区間は,

$$\frac{n-1}{\chi^2_{n-1}(\alpha/2)} \times S^2 < \sigma^2 < \frac{n-1}{\chi^2_{n-1}(1-\alpha/2)} \times S^2$$

$$\frac{8}{17.53} \times 72 < \sigma^2 < \frac{8}{2.180} \times 72$$

$$32.85 < \sigma^2 < 264.2$$

L11-Q3

Quiz 解答:t 分布の確率と $t_k(\alpha)$

1. $z(0.025) = 1.960.$
2. 標準正規分布の確率密度関数は偶関数なので,
   $z(1 - 0.025) = -z(0.025) = -1.960.$

3. $t_{40}(0.025) = 2.021.$
4. t 分布の確率密度関数は偶関数なので,
   $t_{40}(1 - 0.025) = -t_{40}(0.025) = -2.021.$

L11-Q4

Quiz 解答:母平均値の区間推定 (母分散未知)

1. 重さの標本平均値は $\overline{X} = 50\mathrm{g}$. 不偏標本分散は $S^2 = \frac{1}{4-1} \cdot 14\mathrm{g}^2$. t 分布表から 自由度 $k = n - 1 = 3$ の $t_3(0.05/2)$ を参照して, 信頼係数 0.95 の信頼区間は

$$50 - 3.182 \times \sqrt{\tfrac{14}{3}/4} < \mu < 50 + 3.182 \times \sqrt{\tfrac{14}{3}/4}.$$

2. 同様に, t 分布表から 自由度 $k = n - 1 = 3$ の $t_3(0.01/2)$ を参照して,

$$50 - 5.841 \times \sqrt{\tfrac{14}{3}/4} < \mu < 50 + 5.841 \times \sqrt{\tfrac{14}{3}/4}.$$

ここまで来たよ

11 母分散, 母平均値の区間推定

12 線形回帰モデルの推定
- 最尤推定
- 線形モデルとしての回帰分析
- 多次元の確率変数の母期待値の推定
- 標本共分散, 標本相関係数

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

11 母分散, 母平均値の区間推定

12 線形回帰モデルの推定
- 最尤推定
- 線形モデルとしての回帰分析
- 多次元の確率変数の母期待値の推定
- 標本共分散, 標本相関係数

(確率変数でない) 変数 x に依存する確率変数 Y

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 $(i = 1, \ldots, n)$ で製造したときの重さが y_i $(i = 1, \ldots, n)$ である確率密度は, 独立分布なので積.

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

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,

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

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

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

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

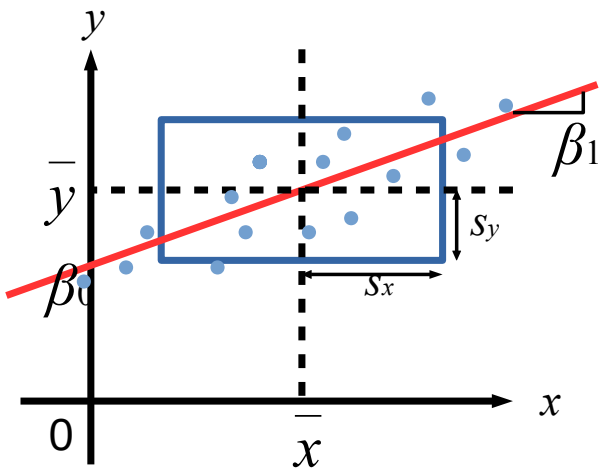
$$Y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \epsilon$$

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.

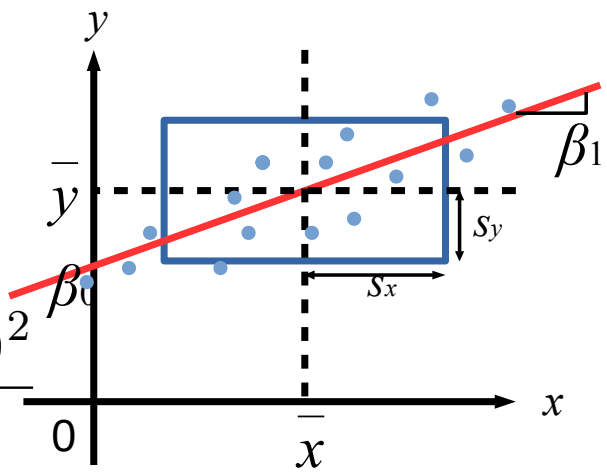
決定係数

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

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



$0 \leq R^2 \leq 1$ で, 1 に近いほど, あてはまりがよい.

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

岩薙林 確率・統計 例題 9.2, 9.3, §9 問題 3,4,5, §9 練習問題 1

L12-Q1

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

11 母分散, 母平均値の区間推定

12 線形回帰モデルの推定
- 最尤推定
- 線形モデルとしての回帰分析
- 多次元の確率変数の母期待値の推定
- 標本共分散, 標本相関係数

(復習) 多次元の確率変数

確率統計 I(2021)L04

確率分布 (母分布, 母集団)

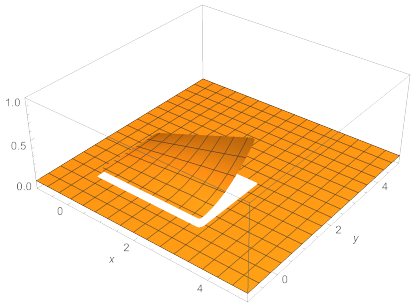
離散型

確率関数 $p(x,y) =$

$y \backslash x$	7	8	9	計
0	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	
1	0	0	$\frac{1}{3}$	
計				

連続型

確率密度関数 $f(x,y) =$



標本 (サイズ n)

離散型

データ番号 i	x	y
1	8	0
2	9	1
$\vdots$	$\vdots$	$\vdots$
n	8	0

連続型

データ番号 i	x	y
1	1.2	0.95
2	2.5	1.24
$\vdots$	$\vdots$	$\vdots$
n	0.9	0.04

多次元の母期待値の推定は 1 次元と同じのりで

同時分布が与えられたときの母期待値

岩薩林 確率・統計 (3.16)p.60 | 確率統計 I(2021)L04

$$\text{連続型} \quad \mathrm{E}[g(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x,y)\cdot g(x,y)\mathrm{d}x\,\mathrm{d}y$$

1 次元 確率統計 I(2021)L10 $\rightsquigarrow$ 多次元

標本期待値

$$g(X,Y) \text{ の標本期待値 } \overline{g(X,Y)} = \frac{1}{n}[g(X_1,Y_1)+\cdots+g(X_n,Y_n)]$$

が, $\mathrm{E}[g(X,Y)]$ の ‘よい’ 推定量になっている.

$g(X,Y)$ が X のみによる場合 $\mathrm{E}[g(X)]$, 例えば $\mathrm{E}[X]$ だったら, 周辺分布 $f_{\mathrm{X}}(x) = \int f(x,y)\mathrm{d}y$ で考えられば, 確率統計 I(2021)L11 そのままじゃん.

L12-Q2

Quiz(多次元の確率変数の母期待値母共分散の推定)

以下は, 2 次元の確率変数 X, Y のサイズ $n = 5$ の標本である

X	Y
1	5
3	15
4	14
5	11
7	20

1. 母平均値 $\mathrm{E}[X]$ を推定しよう.
2. 母期待値 $\mathrm{E}[X^2Y]$ を推定しよう.
3. 母分散 $\mathrm{V}[X]$ を推定しよう.
4. 母共分散 $\mathrm{Cov}[X, Y]$ を推定しよう.
5. 母相関係数 $\rho[X, Y]$ を推定しよう.

ここまで来たよ

11 母分散, 母平均値の区間推定

12 線形回帰モデルの推定
- 最尤推定
- 線形モデルとしての回帰分析
- 多次元の確率変数の母期待値の推定
- 標本共分散, 標本相関係数

母分散の推定値

確率統計 I(2021)L04

不偏標本分散 確率統計 I(2021)L10

$$\text{不偏標本分散 } {S_x}^2 = S_{xx} = \frac{1}{n-1}[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2]$$
$$= \frac{n}{n-1}\left[\frac{1}{n}\sum_i X_i^2 - (\overline{X})^2\right]$$

が, 母分散 $\mathrm{V}[X]$ の ‘よい’ 推定値になっている.

$\mathrm{V}[Y]$ の推定値は, $S_y^2 = S_{yy}$ と書く.

母共分散の推定

母共分散 covariance

岩薩林 確率・統計 (3.19)p.61,(3.20)p.62

X, Y が確率変数で, $\mu_{\mathrm{X}} = \mathrm{E}[X], \mu_{\mathrm{Y}} = \mathrm{E}[Y]$ とおいたとき,

$$\text{母共分散}\ \mathrm{Cov}[X,Y] \overset{\text{定義}}{=} \mathrm{E}[(X-\mu_{\mathrm{X}})(Y-\mu_{\mathrm{Y}})] \tag{3.19}$$

$$= \boxed{\text{岩薩林 確率・統計 (3.20)}} \cdots = \mathrm{E}[XY] - \mathrm{E}[X] \times \mathrm{E}[Y]. \tag{C1,(3.20))}$$

不偏標本共分散

岩薩林 確率・統計 なし

不偏標本共分散

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{x}\cdot\overline{y}\right]$$

が母共分散 $\mathrm{Cov}[X,Y]$ の 'よい' 推定値になっている.

母相関係数の推定

母相関係数 correlation

岩薩林 確率・統計 (3.22)p.64

$$\text{母相関係数}\rho[X,Y] \overset{\text{定義}}{=} \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{V}[X]}\sqrt{\mathrm{V}[Y]}} \tag{3.22}$$

標本相関係数

岩薩林 確率・統計 §1.3(1.19)p.22

標本相関係数

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_x^2}\sqrt{S_y^2}}$$

が母相関係数 $\rho[X,Y]$ に対応する.

(n や $n-1$ は分子分母でキャンセルしているから, 分子分母一斉に変えてもよい)

フィッシャーの z-変換

フィッシャーの z-変換

(X, Y) を 2 次元正規分布にしたがう確率変数とするとき,
サイズ n の標本の標本相関係数を確率変数 R と考える.
確率変数

$$Z = \frac{1}{2} \log \frac{1+R}{1-R}$$

は, 近似的に $\mathrm{N}(\mathrm{Cov}[X, Y], \frac{1}{n-3})$ にしたがう.

逆変換は $R = \tanh Z$.
これで, 相関係数の区間推定や, 検定 (典型的には, 帰無仮説 $r = 0$) ができると.

L12-Q2

Quiz(多次元の確率変数の母期待値母共分散の推定)

以下は, 2 次元の確率変数 X, Y のサイズ $n = 5$ の標本である

X	Y
1	5
3	15
4	14
5	11
7	20

1. 母平均値 $\mathrm{E}[X]$ を推定しよう.
2. 母期待値 $\mathrm{E}[X^2Y]$ を推定しよう.
3. 母分散 $\mathrm{V}[X]$ を推定しよう.
4. 母共分散 $\mathrm{Cov}[X, Y]$ を推定しよう.
5. 母相関係数 $\rho[X, Y]$ を推定しよう.