母平均値の区間推定・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　L13(2022-07-04 Mon)

最終更新: Time-stamp: "2022-07-04 Mon 06:49 JST hig"

今日の目標

- 母平均値を区間推定できる　岩薙林 確率・統計 §7.3
- 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

L12-Q1

Quiz 解答: 母平均値の区間推定 (母分散既知)

1. 標本平均値は $\overline{X} = 50$. よって, 信頼係数 0.95 の信頼区間は

$$50 - 1.96 \times \sqrt{\frac{3^2}{4}} < \mu < 50 + 1.96 \times \sqrt{\frac{3^2}{4}}.$$

すなわち, $47.06 < \mu < 52.94$.

2. 同様に,

$$50 - 2.58 \times \sqrt{\frac{3^2}{4}} < \mu < 50 + 2.58 \times \sqrt{\frac{3^2}{4}}.$$

すなわち, $46.13 < \mu < 53.87$.

scipy.stats.norm(loc=0,scale=1).ppf(1-0.05/2) で 1.96
scipy.stats.norm(loc=0,scale=1).ppf(1-0.01/2) で 2.58 が得られる.

L12-Q2

Quiz 解答: 区間推定の性質

1

L12-Q3

Quiz 解答: 母比率の区間推定

A 候補に投票したを $X = 1$, しなかったを $X = 0$ とする.

1. 標本比率は $\hat{p} = \frac{35}{50} = 0.7$. 母比率 p を 0.7 と推定する.

2. 母比率 p の信頼係数 $1-\alpha=0.95$ の信頼区間は,

$$\frac{7}{10}-1.96\times\sqrt{\frac{1}{50}\cdot\frac{7}{10}\cdot\left(1-\frac{7}{10}\right)}<p<\frac{7}{10}+1.96\times\sqrt{\frac{1}{50}\cdot\frac{7}{10}\cdot\left(1-\frac{7}{10}\right)}$$
$$\frac{7}{10}-0.13<p<\frac{7}{10}+0.13$$
$$0.57<p<0.83$$

信頼係数 0.95 では当選ってことですね (放送用語「当選確実」で, 多くの選挙区で判定したとき, 後で社長があやまらなきゃいけない確率は 0.05).

3. 母比率 p の信頼係数 0.99 の信頼区間は,

$$\frac{7}{10}-2.58\times\sqrt{\frac{1}{50}\cdot\frac{7}{10}\cdot\left(1-\frac{7}{10}\right)}<p<\frac{7}{10}+2.58\times\sqrt{\frac{1}{50}\cdot\frac{7}{10}\cdot\left(1-\frac{7}{10}\right)}$$
$$\frac{7}{10}-0.17<p<\frac{7}{10}+0.17$$
$$0.53<p<0.87$$

信頼係数 0.99 のほうが慎重な判断基準ですが, それでも当選ってことですね.

L12-Q4

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. $\chi^2_1(1 - 0.05) = 0.00393.$

L12-Q5

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 $1 - \alpha = 0.95$ の信頼区間は,

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

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

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

scipy.stats.chi2(df=8).ppf(1-0.05/2) で 17.53
scipy.stats.chi2(df=8).ppf(0.05/2) で 2.180 が得られる.

ここまで来たよ

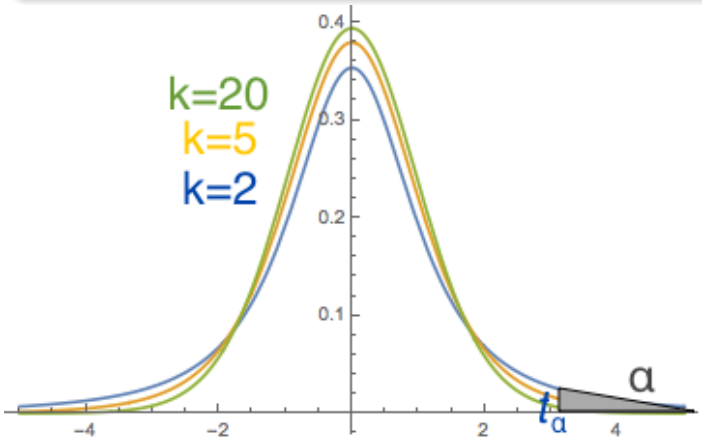
13 母比率・母分散の区間推定
- 母平均値の区間推定 (正規母集団, 母分散未知)

13 母平均値の区間推定・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

t 分布

t 分布

$Z \sim \mathrm{N}(0, 1^2)$, $W \sim \chi^2_k$, Z と W が独立,
のとき連続型確率変数 $T = \dfrac{Z}{\sqrt{W/k}}$ のしたがう分布を自由度 k の (スチューデントの, またはゴセットの)t 分布 t_k という.



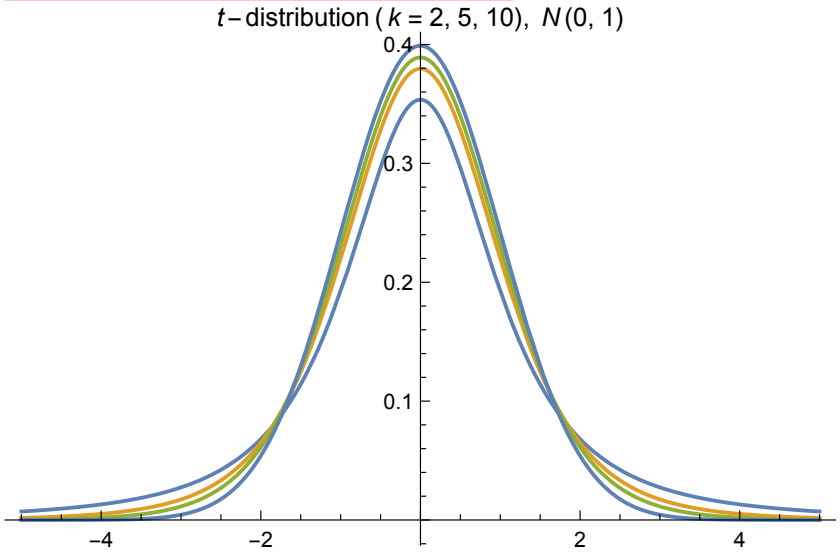
自由度 k が小さいとき, $\mathrm{N}(0, 1^2)$ より低く広い.
自由度 $k \to +\infty$ で $\mathrm{N}(0, 1^2)$ に一致する.

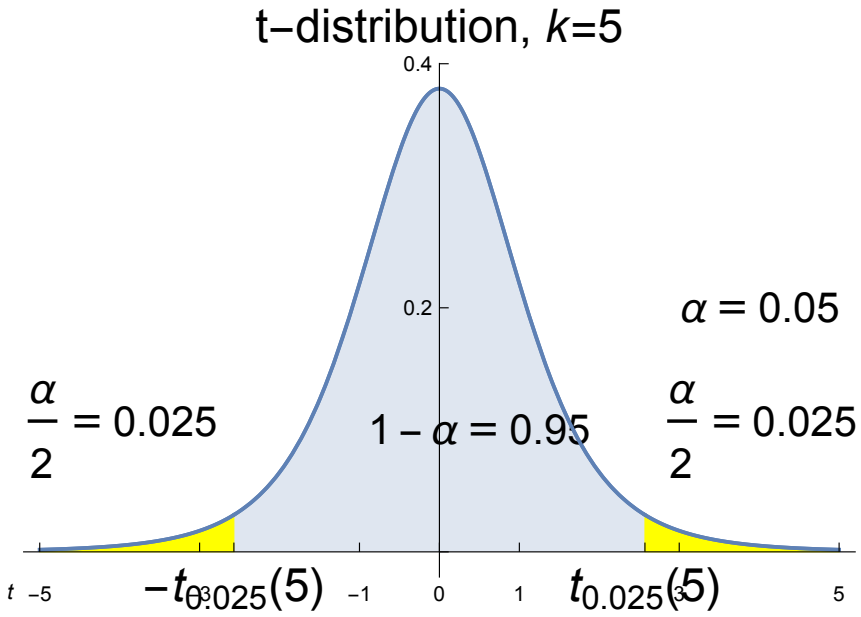
t 分布の確率密度関数と累積分布関数

岩薩林 確率・統計 付表 2

Python scipy.stats.t(df=k)
Excel t.dist(x,k,TRUE/FALSE=累積分布/確率密度), t.inv(x,k)=$F^{-1}(x)$.
t 分布の確率密度関数は偶関数なので, $t_k(1-\alpha) = -t_k(\alpha)$.
岩薩林 確率・統計 付表 2 には上側確率 $P(T > t_k(\alpha)) = \alpha$ となる $t_k(\alpha)$ を定義が載ってる. .ppf(1-α).

岩薩林 確率・統計 図 5.8(p.128)





L13-Q1

Quiz(t 分布の確率と $t_k(\alpha)$)

標準正規分布にしたがう確率変数 $Z \sim \mathrm{N}(0, 1^2)$ と, 自由度 $k = 40$ の t 分布にしたがう確率変数 $T \sim \mathrm{t}_{40}$ を考える. 各分布の累積分布関数 F, F^{-1} を使って答え, さらに Python, Excel, 数表を使って数値にしよう.

1. 確率 $P(Z > z_0) = 0.025$ となる $z_0 = z(0.025)$ を求めよう.
2. 確率 $P(Z > z_0) = 1 - 0.025$ となる $z_0 = z(1 - 0.025)$ を求めよう.
3. 確率 $P(T > w_0) = 0.025$ となる $w_0 = t_{40}(0.025)$ を求めよう.
4. 確率 $P(T > w_0) = 1 - 0.025$ となる $w_0 = t_{40}(1 - 0.025)$ を求めよう.

母平均値の区間推定 (正規母集団, 母分散未知) 岩薙林 確率・統計 §7.1

母分散既知の区間推定 確率統計 I(2022)L12 では, 標本平均値を標準化した. $Z = \frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$.
ふつう, μ がわからないときは σ^2 もわかってない.
σ^2 のかわりに不偏標本分散 S^2 (確率変数) を使った, 標準化もどき

$T = \frac{\overline{X}_{(n)}-\mu}{\sqrt{S^2/n}} = \frac{\frac{\overline{X}_{(n)}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}$ でやっちゃいたい. 分子 $\sim \mathrm{N}(0,1^2)$, 分母
$= \sqrt{W/(n-1)}, W \sim \chi^2_{n-1}$ より, $T \sim \mathrm{t}_{n-1}$.

T の分布

母集団 $\mathrm{N}(\mu,\sigma^2)$ から, サイズ n の標本 $X_1,\cdots,X_n$ を取り出したとき,

$$T = \frac{\overline{X}_{(n)} - \mu}{\sqrt{S^2/n}}$$

は, 自由度 $n-1$ の Student の t 分布にしたがう.

母集団が厳密に正規分布にしたがわなくても近似的に正しいことが多い.

母平均値の信頼区間 (母分散未知) 岩薩林 確率・統計 定理 7.1(7.3)

(母分散未知の) 正規分布 $\mathrm{N}(\mu, ?^2)$ にしたがう母集団から, サイズ n の標本を得たとき, 母平均値 μ の 信頼係数 $1-\alpha$ の信頼区間は

$$\overline{X} - F^{-1}(1-\tfrac{\alpha}{2}) \times \sqrt{\frac{S^2}{n}} < \mu < \overline{X} - F^{-1}(\tfrac{\alpha}{2}) \times \sqrt{\frac{S^2}{n}}.$$

$$\overline{X} - t_{n-1}(\tfrac{\alpha}{2}) \times \sqrt{\frac{S^2}{n}} < \mu < \overline{X} + t_{n-1}(\tfrac{\alpha}{2}) \times \sqrt{\frac{S^2}{n}}.$$

ただし, $\overline{X}$: 標本平均値, S^2: 不偏標本分散, n: 標本サイズ, $t_{n-1}(\frac{\alpha}{2})$: 自由度 $n-1$ の t 分布の上側確率が $\frac{\alpha}{2}$ となる点.

母分散既知 確率統計 I(2022)L12 と比べて,

$\sigma^2 \rightsquigarrow S^2$

$z(\frac{\alpha}{2}) \rightsquigarrow t_{n-1}(\frac{\alpha}{2})$

L13-Q2

Quiz(母平均値の区間推定 (母分散未知))

あるドーナツ製造マシンが製造するドーナツの重さ $X\mathrm{g}$ は, 正規分布にしたがう確率変数である
製造された 4 個のドーナツの重さを測定したところ, 次のようだった.
$51\mathrm{g}, 52\mathrm{g}, 47\mathrm{g}, 50\mathrm{g}$.

1. 母平均値 $\mu = \mathrm{E}[X]$ を, 信頼係数 $1-\alpha = 0.95$ で区間推定しよう. まず, $t_k(\alpha)$ または F^{-1} で書き, 小数に直そう.
2. 母平均値 $\mu = \mathrm{E}[X]$ を, 信頼係数 $1-\alpha = 0.99$ で区間推定しよう.

岩薩林 確率・統計 例題 7.1(p.158), 問題 1(p.158) 第 7 章練習問題 1(2)

区間推定のまとめ 岩薩林 確率・統計 p.xiv

- 母比率 (母集団がベルヌーイ分布) 岩薩林 確率・統計 定理 7.5(p.170)
  - 入力 信頼係数 α, 標本サイズ n, 該当するデータの個数 k
  - 係数 $z(\alpha/2)$ を求める分布: 標準正規分布 岩薩林 確率・統計 付表 1 下
- 母分散 (母集団が正規分布) 岩薩林 確率・統計 定理 7.3(p.163)
  - 入力 信頼係数 α, 標本サイズ n, 不偏標本分散 S^2
  - 係数 $\chi^2_{n-1}(\alpha/2)$ を求める分布 自由度 $n-1$ のカイ二乗分布
    岩薩林 確率・統計 付表 3
- 母平均値 (母集団が正規分布) 岩薩林 確率・統計 定理 7.1(p.157)
  - 入力 信頼係数 α, 標本サイズ n, 不偏標本分散 S^2, 標本平均値 $\overline{X}$
  - 係数 $t_{n-1}(\alpha/2)$ を求める分布 自由度 $n-1$ の t 分布 岩薩林 確率・統計 付表 2
- 「母分散がわかってるときの母平均値」という実用上あまりないケース 岩薩林 確率・統計 定理 6.1(p.145) 岩薩林 確率・統計 付表 1 下

ここまで来たよ

13 母比率・母分散の区間推定
- 母平均値の区間推定 (正規母集団, 母分散未知)

13 母平均値の区間推定・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

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

13 母比率・母分散の区間推定
- 母平均値の区間推定 (正規母集団, 母分散未知)

13 母平均値の区間推定・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

(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,

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

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

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

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

$Y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \epsilon$ ($\hat{\beta}_0, \hat{\beta}_1$: 回帰係数) に代入.

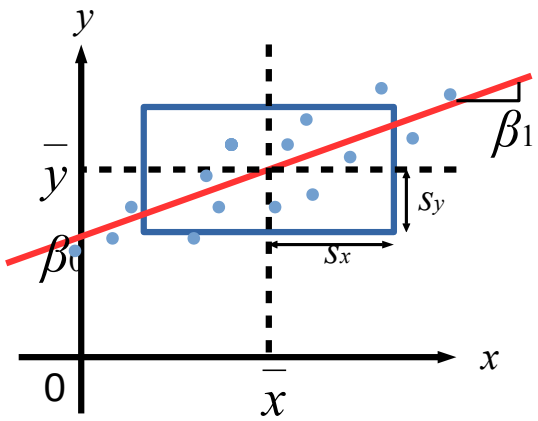
回帰直線

岩薩林 確率・統計 §9.1

推定結果 $\hat{\beta}_0, \hat{\beta}_1$ を係数とする xy 平面の直線

データ分析

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



回帰係数, 予測値の信頼区間

確率統計 II,III