多変量解析入門

永田 靖・棟近雅彦

輪読ゼミ第２回（2022/4/27）

教科書：p43～p60

担当者：B4 柳 智也

単回帰分析

単回帰分析とは？

→ある説明変数 x から目的変数 y を制御・予測すること

解析ストーリー

①最小二乗法による回帰式の推定

②寄与率・自由度調整済寄与率による回帰式の性能評価

③回帰係数の検定・区間推定

④残差・テコ比を用いた回帰式の妥当性の検討

⑤得られた回帰式による予測

①最小二乗法による回帰式の推定

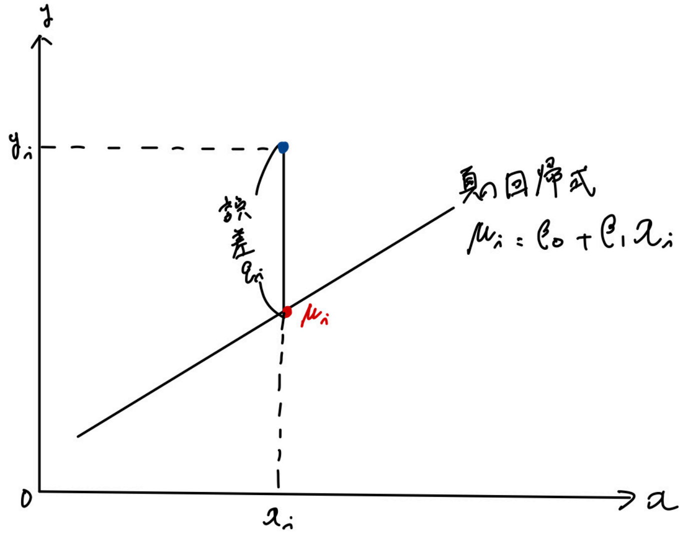
以下の単回帰モデルを想定

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \ \varepsilon_i \sim N(0, \sigma^2)$$

誤差 ε_i は独立に $N(0, \sigma^2)$ に従うと仮定し、回帰母数 β_0, β_1 を推定

解釈

- x の値を決めると母平均 μ_i が定まる
- 観測値 y_i はそれに誤差 ε_i が加わったもの
- 母平均に対して、$\mu_i = \beta_0 + \beta_1 x_i$ という構造を仮定



回帰母数の算出：最小2乗法

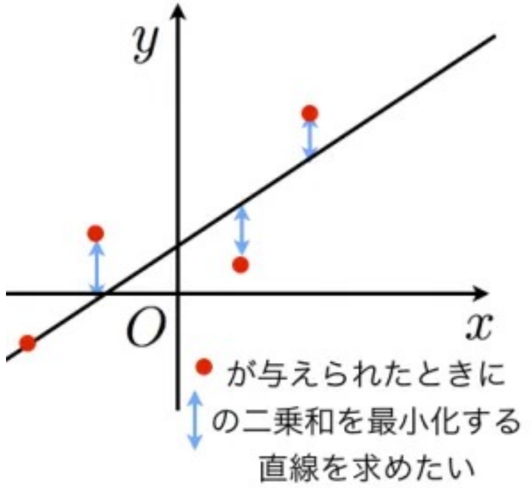
予測値を $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$ とし、これと実測値の差 e_i を考える

$$残差：e_i = y_i - \hat{y}_i = y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)$$

これを2乗し、全てのデータで足し合わせた残差平方和を最小化

$$残差平方和：S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)\}^2$$

このような回帰母数の算出方法を、最小2乗法と呼ぶ



残差平方和の最小化

S_e を β_0 と β_1 について偏微分し、0とする

$$\frac{\partial S_e}{\partial \beta_0} = -2\sum_{i=1}^{n}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i\right) = 0 \qquad \text{-①}$$

$$\frac{\partial S_e}{\partial \beta_1} = -2\sum_{i=1}^{n}x_i\left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i\right) = 0 \qquad \text{-②}$$

①、②を整理して正規方程式を得る

$$\widehat{\beta_0}n + \widehat{\beta_1}\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i \qquad \text{-①'}$$

$$\widehat{\beta_0}\sum_{i=1}^{n}x_i + \widehat{\beta_1}\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_iy_i \qquad \text{-②'}$$

①'より

$$\widehat{\beta_0} = \frac{\sum y_i}{n} - \widehat{\beta_1}\frac{\sum x_i}{n} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

これを②'に代入して

$$\left(\frac{\sum y_i}{n} - \widehat{\beta_1}\frac{\sum x_i}{n}\right)\sum x_i + \widehat{\beta_1}\sum x_i^2 = \sum x_iy_i$$

整理して

$$\widehat{\beta_1}\underline{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)} = \underline{\sum x_iy_i - \frac{(\sum x_i)(\sum y_i)}{n}} \qquad \text{-③}$$

残差平方和の最小化

ここで、x の平方和とx と y の偏差積和を考えると、

$$
\begin{aligned}
S_{xx} &= \sum(x_i - \bar{x})^2 = \sum\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\
&= \sum x_i^2 - n\bar{x}^2 \\
&= \underline{\sum x_i^2 - \left(\sum x_i\right)^2/n}
\end{aligned}
$$

$$
\begin{aligned}
S_{xy} &= \sum(x_i - \bar{x})(y_i - \bar{y}) \\
&= \sum(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) \\
&= \sum x_i y_i - n\bar{x}\bar{y} \\
&= \underline{\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n}
\end{aligned}
$$

よって、③は $\widehat{\beta_1}S_{xx} = S_{xy}$ と書け、

$$
\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}, \qquad \widehat{\beta_0} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}
$$

このとき、 S_e の最小値は、

$$
\begin{aligned}
S_e &= \sum\left\{y_i - \left(\widehat{\beta_0} + \widehat{\beta_1}x_i\right)\right\}^2 \\
&= \sum\left\{y_i - \bar{y} - \widehat{\beta_1}(x_i - \bar{x})\right\}^2 \\
&= \sum(y_i - \hat{y})^2 - 2\widehat{\beta_1}\sum(x_i - \bar{x})(y_i - \bar{y}) + \widehat{\beta_1}^2\sum(x_i - \bar{x})^2 \\
&= S_{yy} - 2\widehat{\beta_1}S_{xy} + \widehat{\beta_1}\frac{S_{xy}}{S_{xx}}S_{xx} \\
&= S_{yy} - \widehat{\beta_1}S_{xy}
\end{aligned}
$$

以上より、推定式は、

$$
\begin{aligned}
\hat{y} &= \widehat{\beta_0} + \widehat{\beta_1}x \\
&= \bar{y} - \widehat{\beta_1}\bar{x} + \widehat{\beta_1}x \quad \Longleftarrow \quad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} \text{ を代入} \\
&= \bar{y} + \widehat{\beta_1}(x - \bar{x})
\end{aligned}
$$

行列とベクトルによる表現

単回帰モデルを行列表記にすると、次のように変形できる

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \alpha_0 + \beta_1 (x_i - \bar{x}) + \varepsilon_i$$

$$\alpha_0 = \beta_0 + \beta_1 \bar{x}, \varepsilon_i \sim N(0, \sigma^2)$$

これは、ベクトルと行列を定義し、以下のように表現できる

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I_n)$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathrm{X} = \begin{bmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \alpha_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

行列とベクトルによる表現

残差ベクトル

$$\boldsymbol{e} = \begin{bmatrix} y_1 - \{\widehat{\alpha_0} + \widehat{\beta_1}(x_1 - \bar{x})\} \\ y_2 - \{\widehat{\alpha_0} + \widehat{\beta_1}(x_2 - \bar{x})\} \\ \vdots \\ y_n - \{\widehat{\alpha_0} + \widehat{\beta_1}(x_n - \bar{x})\} \end{bmatrix} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}$$

残差平方和

$$S_e = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$= [e_1, e_2, \ldots, e_n] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \boldsymbol{e}'\boldsymbol{e}$$

$$= \left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right)$$

$$= \left(\boldsymbol{y}' - \widehat{\boldsymbol{\beta}}'X'\right)\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right)$$

$$= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'X\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}'X'\boldsymbol{y} + \widehat{\boldsymbol{\beta}}'\mathrm{X}'\mathrm{X}\widehat{\boldsymbol{\beta}}$$

$$= \boldsymbol{y}'\boldsymbol{y} - 2\widehat{\boldsymbol{\beta}}X'\boldsymbol{y} + \widehat{\boldsymbol{\beta}}'\mathrm{X}'\mathrm{X}\widehat{\boldsymbol{\beta}} \quad (※)$$

※線形代数の復習

2つの行列AとBの掛け算と転置

$$\underline{(AB)' = B'A'}$$

行列とベクトルによる表現

残差平方和をベクトル $\hat{\boldsymbol{\beta}}$ で微分し $\mathbf{0}$ とおく

$$\frac{\partial S_e}{\partial \hat{\boldsymbol{\beta}}} = -2X'y + 2X'X\hat{\boldsymbol{\beta}} = \mathbf{0} \qquad (※)$$

※線形代数の復習（p39,40）

a ：定数ベクトル、A ：定数の対称行列

この時、ベクトル $\boldsymbol{x}$ による微分は

$$\frac{\partial \boldsymbol{x}'a}{\partial \boldsymbol{x}} = \boldsymbol{a}, \qquad \frac{\partial \boldsymbol{x}'A\boldsymbol{x}}{\partial \boldsymbol{x}} = 2A\boldsymbol{x}$$

これより、

$$X'X\hat{\boldsymbol{\beta}} = X'\boldsymbol{y} \quad \leftrightarrow \quad \hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$$

ここで、$X'X$ について考えると、

$$X'X = \begin{bmatrix} n & 0 \\ 0 & S_{xx} \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{nS_{xx}}\begin{bmatrix} S_{xx} & 0 \\ 0 & n \end{bmatrix} = \begin{bmatrix} 1/n & 0 \\ 0 & 1/S_{xx} \end{bmatrix}$$

$S_{xx} \neq 0$ なら、逆行列が存在する

平方和の分解

平方和の分解を行う

$$S_{yy} = \sum_{i=i}^{n}(y_i - \bar{y})^2$$

$$= \sum\left\{y_i - \left(\widehat{\beta_0} + \widehat{\beta_1}x_i\right) + \left(\widehat{\beta_0} + \widehat{\beta_1}x_i\right) - \bar{y}\right\}^2$$

$$= \sum\left\{y_i - \left(\widehat{\beta_0} + \widehat{\beta_1}x_i\right)\right\}^2 + \sum\left\{\left(\widehat{\beta_0} + \widehat{\beta_1}x_i\right) - \bar{y}\right\}^2 + 2\underline{\sum\left\{y_i - \left(\widehat{\beta_0} + \widehat{\beta_1}x_i\right)\right\}\left\{\left(\widehat{\beta_0} + \widehat{\beta_1}x_i\right) - \bar{y}\right\}}$$

ここで、

$$\sum\left\{y_i - \left(\widehat{\beta_0} + \widehat{\beta_1}x_i\right)\right\}\left\{\left(\widehat{\beta_0} + \widehat{\beta_1}x_i\right) - \bar{y}\right\} = \sum e_i\left\{\left(\widehat{\beta_0} + \widehat{\beta_1}x_i\right) - \bar{y}\right\} = \left(\widehat{\beta_0} - \bar{y}\right)\sum e_i + \widehat{\beta_1}\sum x_i e_i = 0$$

（p5の S_e の偏微分で導出した等式から、$\sum e_i = 0, \sum x_i e_i = 0$ が成立することに注意）

寄与率の定義

よって、

$$S_{yy} = \sum\left\{y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)\right\}^2 + \sum\left\{(\widehat{\beta_0} + \widehat{\beta_1} x_i) - \bar{y}\right\}^2$$

$$= S_e + S_R \quad \left(S_R = \sum\left\{(\widehat{\beta_0} + \widehat{\beta_1} x_i) - \bar{y}\right\}^2\right)$$

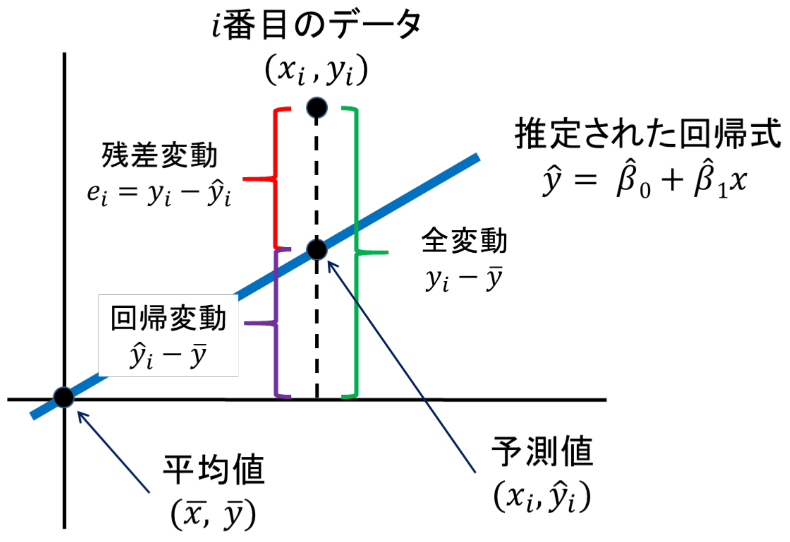
S_R は回帰式の予測値が平均とどれぐらい異なるかを表し、回帰による平方和と呼ぶ

ここから、回帰式の性能評価の指標を定義

$$R^2 = \frac{S_R}{S_{yy}} = 1 - \frac{S_e}{S_{yy}}$$

- R^2 は寄与率（決定係数）と呼ばれ、全変動のうち回帰によって説明できる変動の割合を表す
- S_R が大きいほど回帰で説明できる変動が大きくなるので、 R^2 は1に近づくほど良い

平方和の分解イメージ



自由度と自由度調整済み寄与率

各平方和の自由度は以下のように求められる

① S_{yy} がある値を取る時、n個のデータのうちn-1個のデータは自由な値を取ることができるので、 S_{yy} の自由度 $\phi_T = n - 1$

② S_e がある値を取る時、最小2乗法では $\sum e_i = 0, \sum x_i e_i = 0$ が成立するため、連立方程式と考えれば自由に決められる e_i はn-2個になり、 S_e の自由度 $\phi_e = n - 2$

③ S_R がある値を取る時、回帰式で自由に決められるのは $\widehat{\beta_0}$ と $\widehat{\beta_1}$ の2つだが、$\bar{y} = \widehat{\beta_0} + \widehat{\beta_1}\bar{x}$ が成立するため、片方が決まればもう一方が必然的に決まるので、 S_R の自由度 $\phi_R = 1$

寄与率を自由度で調整したものを自由度調整済寄与率と呼び、以下の式で表す。

$$R^{*2} = 1 - \frac{S_e/\phi_e}{S_{yy}/\phi_T}$$

統計量が従う分布を求める

単回帰モデル $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ から以下が成立

$$E(\boldsymbol{y}) = X\boldsymbol{\beta}$$

$$V(\boldsymbol{y}) = V(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{V}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$$

これより、$\hat{\boldsymbol{\beta}}$ の期待値と分散共分散行列は

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E\left((X'X)^{-1}X'\boldsymbol{y}\right) \\ &= (X'X)^{-1}X'E(\boldsymbol{y}) \\ &= (X'X)^{-1}X'X\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= V\left((X'X)^{-1}X'\boldsymbol{y}\right) \\ &= (X'X)^{-1}X'V(\boldsymbol{y})X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

これより、

$$\hat{\boldsymbol{\beta}} \left(= \begin{bmatrix} \widehat{\alpha_0} \\ \widehat{\beta_1} \end{bmatrix}\right) \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1})$$

具体的に計算して、

$$\sigma^2(X'X)^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/S_{xx} \end{bmatrix}$$

$$\widehat{\alpha_0} \sim N(\alpha_0, \sigma^2/n), \quad \widehat{\beta_1} \sim N(\beta_1, \sigma^2/S_{xx})$$

これを使って、回帰係数の検定と推定を行う

回帰係数の検定と推定

$$\widehat{\beta_1} \sim N(\beta_1, \sigma^2/S_{xx}) \text{ より、} u = \frac{\widehat{\beta_1}-\beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0, 1^2)$$

σ^2 が未知なので、推定量 $\hat{\sigma}^2 = V_e$ を代入すると、

$$t = \frac{\widehat{\beta_1} - \beta_1}{\sqrt{V_e/S_{xx}}} \sim t(\phi_e), \phi_e = n - 2$$

これを用いて、帰無仮説 $H_0\colon \beta_1 = 0$, 対立仮説 $H_1\colon \beta_1 \neq 0$ を検定することができ、

$t_0 = \frac{\widehat{\beta_1}}{\sqrt{V_e/S_{xx}}}$ を計算して $|t_0| \geq t(\phi_e, \alpha)$ なら有意水準 α で有意、帰無仮説を棄却する

また、β_1 の95%信頼区間は、$\widehat{\beta_1} \pm t(\phi_e, 0.05)\sqrt{V_e/S_{xx}}$ で求めることができる

残差とテコ比の検討

回帰分析の評価において、残差が異常に大きかったり、正規性を満たしていないことがある
→残差の検討の意味と内容を理解することが重要

①標準化残差
各サンプルにおいて、残差を標準化した値を用いて異常かどうかを判断する

$$e'_k = e_k/\sqrt{V_e} \sim N(0, 1^2)$$ とし、$|e'| \geq 3.0$ なら注意、$|e'| \geq 2.5$ なら留意

②残差の散布図
残差の散布図を描いて、残差に何らかの傾向があるか確認する

③残差の t 値
残差のt値を計算して、①と同様の基準で判定（h_{kk} はこの後紹介するテコ比）

$$t_k = e_k/\sqrt{(1-h_{kk})V_e}$$

テコ比

→各サンプルが予測値に対してどれだけ影響しているかを測る値

回帰係数の推定量を第 k サンプルの
予測値に代入すると、

$$
\begin{aligned}
\widehat{y_k} &= \widehat{\beta_0} + \widehat{\beta_1} x_k = \bar{y} + \widehat{\beta_1}(x_k - \bar{x}) \\
&= \bar{y} + \frac{S_{xy}(x_k - \bar{x})}{S_{xx}} \\
&= \bar{y} + \frac{(x_k - \bar{x})\sum(x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} \\
&= \frac{\sum y_i}{n} + \frac{(x_k - \bar{x})\sum(x_i - \bar{x})y_i}{S_{xx}} \\
&= h_{k1}y_1 + h_{k2}y_2 + \cdots + h_{kk}y_k + \cdots + h_{kn}y_n
\end{aligned}
$$

y_k の係数をテコ比と呼び、y_k が1単位変化
した時に $\widehat{y_k}$ が変化する量である。

$$h_{kk} = 1/n + (x_k - \bar{x})^2/S_{xx}$$

テコ比が大きすぎると、$\widehat{y_k}$ の値が y_k の値の
変動によって強く影響を受けるので良くない

テコ比は以下の基準を満たし、値の目安と
して 2.5×(テコ比の平均)が使われる

$$\sum_{k=1}^{n} h_{kk} = 1 + 1 = 2, \qquad \frac{1}{n} \leq h_{kk} \leq 1$$

次ページで証明

テコ比の値の範囲の証明

（証明）
$h_{kk} \geq 1/n$ は $(x_k - \bar{x})^2/S_{xx}$ が非負なことから明らかである
次に、$x_i - \bar{x} = X_i$ とおくと、$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ より、$-X_n = \sum_{i=1}^{n-1} X_i$ が成立する
このとき、 $k = n$ として

$$
\begin{aligned}
1 - h_{nn} &= 1 - \frac{1}{n} - \frac{(x_n - \bar{x})^2}{S_{xx}} = \frac{1}{nS_{xx}}\{(n-1)S_{xx} - n(x_n - \bar{x})^2\} \\
&= \frac{1}{nS_{xx}}\left\{(n-1)\sum_{i=1}^{n-1} X_i^2 - X_n^2\right\} = \frac{n-1}{nS_{xx}}\left(\sum_{i=1}^{n-1} X_i^2 - \frac{\left(\sum_{i=1}^{n-1} X_i\right)^2}{n-1}\right) \\
&\geq 0
\end{aligned}
$$

得られた回帰式による予測

$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$ を実際に計算すると、$\widehat{\alpha_0} = \bar{y}$ だとわかる

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1/n & 0 \\ 0 & 1/S_{xx} \end{bmatrix} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 - \bar{x} & x_2 - \bar{x} & \dots & x_n - \bar{x} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} 1/n & 1/n & \dots & 1/n \\ (x_1 - \bar{x})/S_{xx} & (x_2 - \bar{x})/S_{xx} & \dots & (x_n - \bar{x})/S_{xx} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} \bar{y} & 0 \\ 0 & \sum_{i=1}^{n}(x_i - \bar{x})y_i / S_{xx} \end{bmatrix}$$

$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$ より、

$\widehat{\beta_0} + \widehat{\beta_1}x = \widehat{\beta_0} + \widehat{\beta_1}\bar{x} - \widehat{\beta_1}\bar{x} + \widehat{\beta_1}x = \widehat{\alpha_0} + \widehat{\beta_1}(x - \bar{x})$

先ほど求めた統計量の分布

$\widehat{\alpha_0} \sim N(\alpha_0, \sigma^2/n)$, $\widehat{\beta_1} \sim N(\beta_1, \sigma^2/S_{xx})$

これを用いて、推定量の確率分布は、

$$\widehat{\beta_0} + \widehat{\beta_1}x \sim N\left(\beta_0 + \beta_1 x, \left\{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right\}\sigma^2\right)$$

この確率分布を応用して、x を任意の値 x_0 に設定し区間推定や予測区間を求める

区間推定と予測区間

求めた確率分布から、以下の2つが求められる

①母回帰 $\beta_0+\beta_1 x_0$ の信頼率95%信頼区間

$$\widehat{\beta_0}+\widehat{\beta_1}x_0 \pm t(\phi_e, 0.05)\sqrt{\left\{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right\}V_e}$$

①予測 $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon$ の信頼率95%信頼区間（誤差の変動が入ることに注意）

$$\widehat{\beta_0}+\widehat{\beta_1}x_0 \pm t(\phi_e, 0.05)\sqrt{\left\{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right\}V_e}$$

参考

・回帰分析における残差の自由度がn-kになる理由をk元一次方程式で説明してみる

https://qiita.com/anoiro/items/0ee717773b893450a721