ベイズ統計

古谷知之

講義概要

- 線形回帰モデルのベイズ推定
  - 通常最小二乗法(OLS)との比較
  - 尤度関数と自然共役事前分布
  - 平均・分散未知の場合の階層ベイズ推定
- 無情報事前分布と非正則事前分布
- 偏差情報量基準
- R演習

線形回帰モデル（重回帰分析）

- 従属変数yとk個の独立変数$x_1, x_2, \ldots, x_k$に対する標本数がn個の重回帰モデルは以下のように記述できる$(i = 1, \ldots, n)$。

$$
\begin{gathered}
y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 = \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
\vdots \\
y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \\
\vdots \\
y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n
\end{gathered}
$$

線形回帰モデル（重回帰分析）

- 次のようなベクトルと行列を用いて、

$$
\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}
$$

- 次式のように簡略化できる

$$
\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}
$$

重回帰モデルの最小二乗推定量

- 誤差項$\boldsymbol{\varepsilon} = \boldsymbol{y} - X\boldsymbol{\beta}$の二乗和$Q$は、

$$Q = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T X^T X\boldsymbol{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$$

- 最小二乗法より、

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2X^T\boldsymbol{y} + 2X^T X\boldsymbol{\beta} = 0$$

- ここから以下の正規方程式を得る

$$X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}$$

- 両辺に左から$(X^T X)^{-1}$をかけると、回帰係数の推定量$\hat{\boldsymbol{\beta}}$を得る

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

重回帰モデルの統計量

- 回帰係数$\boldsymbol{\beta}$・誤差項$\boldsymbol{\varepsilon}$・従属変数$\boldsymbol{y}$の確率分布から、偏回帰係数$\hat{\boldsymbol{\beta}}$・予測値$\hat{\boldsymbol{y}}$・予測誤差$\boldsymbol{e}$の確率分布は以下のようになる

- 偏回帰係数

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1}X^T\boldsymbol{y} \sim N(\boldsymbol{\beta}, \sigma^2(X^T X)^{-1})$$

- 予測値

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = X(X^T X)^{-1}X^T\boldsymbol{y} = H\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 H)$$

$$H = X(X^T X)^{-1}X^T$$

Hはハット行列

- 予測誤差

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (I - H)\boldsymbol{y} \sim N\left(0, \sigma^2(I - H)\right)$$

線形回帰モデルの尤度関数

- データX、未知パラメータ$\boldsymbol{\beta}$ 、誤差項の分散$\boldsymbol{\sigma}^2$が与えられた条件下で、被説明変数$\boldsymbol{y}$が得られる条件付き確率を尤度関数という
- サンプルiの説明変数$x_i = (1, x_{i1}, \dots, x_{ik})$、被説明変数$y_i$とするとき、尤度関数は以下のような正規分布となる

$$p(y_i|x_i; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

尤度関数

- 尤度関数の平均と分散はそれぞれ以下のとおりとなる

- 平均

$$E(y_i|x_i; \boldsymbol{\beta}, \sigma^2) = x_i\boldsymbol{\beta}$$

- 分散

$$V(y_i|x_i; \boldsymbol{\beta}, \sigma^2) = \sigma^2$$

尤度関数

- 全てのサンプルiについての尤度関数は

$$
\begin{aligned}
p(y|X;\boldsymbol{\beta},\sigma^2) &= \prod_{i=1}^{n} p(y_i|x_i;\boldsymbol{\beta},\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right] \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[-\frac{\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right]
\end{aligned}
$$

対数尤度関数

- 尤度関数の自然対数をとると

$$
\ln[p(y|X;\boldsymbol{\beta},\sigma^2)] = \ln\left[\prod_{i=1}^{n} p(y_i|x_i;\boldsymbol{\beta},\sigma^2)\right]
$$

$$
= \ln\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[-\frac{\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right]\right]
$$

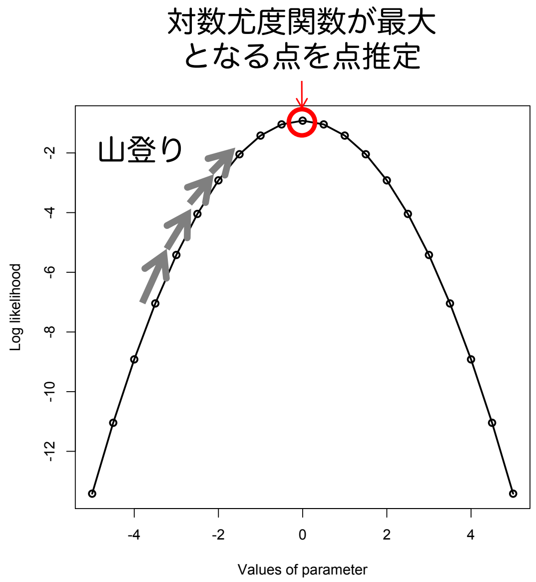
$$
= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2
$$

最尤法と最小二乗法

- 最尤法では(対数)尤度関数を最大化することで未知パラメータを得る⇒対数尤度関数は上に凸となる関数

- 対数尤度関数を最大化することは、次式を最小化することと同じ

$$\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2$$



最小二乗法による解

- 次式を最小化することにより得られる未知パラメータはそれぞれ以下のようになる

$$\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2$$

- 最小二乗解

$$\widehat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$\hat{\sigma}^2 = \frac{1}{n-(k+1)}\sum_{i=1}^{n}\left(y_i - x_i\widehat{\boldsymbol{\beta}}\right)^2$$

自由度: $n-(k+1)$

尤度関数(全データ)

- データX 、未知パラメータ$\boldsymbol{\beta}$ 、分散$\boldsymbol{\sigma}^2$が与えられた条件下で、被説明変数$\boldsymbol{y}$が得られる条件付き確率、すなわち尤度関数は、以下のような正規分布となる

$$p(y|X;\boldsymbol{\beta},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y - X\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

尤度関数(全データ)

- 尤度関数の平均と分散はそれぞれ以下のとおりとなる

- 平均

$$E(y|X;\boldsymbol{\beta},\sigma^2) = X\boldsymbol{\beta}$$

- 分散

$$V(y|X;\boldsymbol{\beta},\sigma^2) = \sigma^2$$

尤度関数(全データ)

- 全てのデータについての尤度関数は

$$p(y|X;\boldsymbol{\beta},\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{(y-X\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

$$=\frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{(y-X\boldsymbol{\beta})^T(y-X\boldsymbol{\beta})}{2\sigma^2}\right]$$

対数尤度関数(全データ)

- 全てのデータについての尤度関数は

$$\ln[p(y|X;\boldsymbol{\beta},\sigma^2)]=\ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{(y-X\boldsymbol{\beta})^T(y-X\boldsymbol{\beta})}{2\sigma^2}\right]\right]$$

$$=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{(y-X\boldsymbol{\beta})^T(y-X\boldsymbol{\beta})}{2\sigma^2}$$

最小二乗法による解(全データ)

- 対数尤度関数を最大化⇔最小二乗法による不偏推定量が得られる

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$$

$$\hat{\sigma}^2 = \frac{1}{\nu}\left(y - X\widehat{\boldsymbol{\beta}}\right)^T\left(y - X\widehat{\boldsymbol{\beta}}\right)$$

$$\nu = n - (k + 1)$$ …自由度

↑定数項を加えた変数の数

尤度関数

- 精度$\tau = 1/\sigma^2$とすると尤度関数は以下のように式変形できる

$$p(y|X;\boldsymbol{\beta},\tau) = \frac{\tau^{n/2}}{(2\pi)^{n/2}} \cdot \left\{exp\left[-\frac{\tau}{2}(y - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta})\right]\right\}$$

$$\propto \tau^{\nu/2} \cdot \left\{exp\left[-\frac{\nu\hat{\sigma}^2}{2}\tau\right]\right\}$$

自然共役事前分布と事後分布

- 正規分布の自然共役事前分布にガンマ分布があった
- 精度$\tau = 1/\sigma^2$が未知のとき、その事前分布にガンマ分布$Ga(\tau|a,b)$を乗じると事後分布(ガウス-ガンマ分布)が得られる

$$p(y|X;\boldsymbol{\beta},\tau)\cdot p(\tau|a,b)$$

$$\propto \tau^{\nu/2}\cdot exp\left[-\frac{\nu\hat{\sigma}^2}{2}\tau\right] \quad \cdots\text{ガウス分布}$$

$$\cdot\,\tau^{a-1}\cdot exp[-b\tau] \quad \cdots\text{ガンマ分布}$$

自然共役事前分布と事後分布

- 尤度関数(正規分布)に自然共役事前分布(ガンマ分布)を乗じて得られる事後分布(ガンマ分布)は以下のようになる（正規化係数のガンマ分布はキャンセル）

$$\begin{aligned}
&p(y|X;\boldsymbol{\beta},\tau)\cdot p(\tau|a,b)\\
&\propto \tau^{\nu/2}\cdot exp\left[-\frac{\nu\hat{\sigma}^2}{2}\tau\right]\cdot\tau^{a-1}\cdot exp[-b\tau]\\
&\propto \tau^{\nu/2}\cdot\tau^{a-1}\cdot exp\left[-b\tau-\frac{\nu\hat{\sigma}^2}{2}\tau\right]\\
&\propto \tau^{(a+\nu/2)-1}\cdot exp\left[-\left(b+\frac{\nu\hat{\sigma}^2}{2}\right)\tau\right]
\end{aligned}$$

自然共役事前分布と事後分布

- 事後分布は以下のようなガンマ分布となる

$$p(y|X;\boldsymbol{\beta},\tau)\cdot p(\tau|a,b)$$

$$\propto \tau^{(a+\nu/2)-1}\cdot exp\left[-\left(b+\frac{\nu\hat{\sigma}^2}{2}\right)\tau\right]$$

$$\propto Ga\left(a+\frac{\nu}{2},b+\frac{\nu\hat{\sigma}^2}{2}\right)$$

階層ベイズ

- 平均と分散(精度)が未知のとき、精度のガンマ分布と尤度関数の正規分布を乗じることで事後分布が得られる

精度τがガンマ分布に従う　p(τ)∝Γ( )

↓

精度τが与えられた条件下で
パラメータβが正規分布に従う　p(β|τ) ∝N( )

↓

パラメータβと精度τが与えられた
条件下で被説明変数yが得られる　p(y|β, τ) ∝N( )

線形回帰モデルのベイズ推定

- 線形回帰モデルは、未知パラメータ$\boldsymbol{\beta}$と精度τについて共役事前分布を与えることによりベイズ推定できる
- 自然共役事前分布

$$p(\boldsymbol{\beta}|\tau, y) = N\left((X^T X)^{-1} X^T y, (X^T X)^{-1}\right)$$

$$p(\tau|y) \propto Ga\left(a + \frac{\nu}{2}, b + \frac{\nu\hat{\sigma}^2}{2}\right) \propto Ga\left(\frac{\nu}{2}, \frac{\nu\hat{\sigma}^2}{2}\right)$$

線形回帰モデルのベイズ推定

- 共役事前分布の超（ハイパー）パラメータを導入することによりを次式のように簡略化する

$$p(\boldsymbol{\beta}|\tau, y) \sim N\left(b_0, \left(\frac{1}{\tau}\right) B_0\right)$$

$$p(\tau|y) \propto Ga\left(\frac{\nu_0}{2}, \frac{\nu_0 S_0}{2}\right)$$

- 超パラメータb_0, B_0, ν_0, S_0は適当に設定しても問題ない

線形回帰モデルのベイズ推定

- 超パラメータを以下のように置き換える

$$b_0 = (X^T X)^{-1} X^T y$$

$$B_0 = (X^T X)^{-1}$$

$$\nu_0 = \nu$$

$$S_0 = \hat{\sigma}^2 = \frac{1}{\nu} \left(y - X\widehat{\boldsymbol{\beta}}\right)^T \left(y - X\widehat{\boldsymbol{\beta}}\right)$$

線形回帰モデルのベイズ推定

- 以下のような事後分布が得られる

$$p(\boldsymbol{\beta}|\tau, y) \sim N(b_1, B_1)$$

$$p(\tau|y) \propto Ga\left(\frac{\nu_1}{2}, \frac{\nu_1 S_1}{2}\right)$$

- ここで、

$$b_1 = B_1(B_0^{-1} b_0 + \tau X^T y)$$
$$B_1^{-1} = B_0^{-1} + \tau X^T X$$
$$\nu_1 = \nu_0 + n$$
$$\nu_1 S_1 = \nu_0 S_0 + (y - X b_1)^T (y - X b_1)$$

線形回帰モデルのベイズ推定

- $\boldsymbol{\beta}$の周辺事後分布は以下のt分布に従う

$$p(\boldsymbol{\beta}|\tau, y) \sim t(b_1, \hat{\sigma}^2 S_1, \nu_1)$$

$$E(\boldsymbol{\beta}|\tau, y) = b_1$$

$$V(\boldsymbol{\beta}|\tau, y) = \frac{\nu_1 \hat{\sigma}^2}{\nu_1 - 2} S_1$$

線形回帰モデルのギブスサンプリング

- 繰り返し回数$s = 0, \dots, N$とし、初期値b_0, B_0, ν_0, S_0を設定する

1. $p\left(\boldsymbol{\beta}^{\boldsymbol{s+1}}|\tau^s, y\right) \sim N(b_1^s, (1/\tau)B_1^s)$から$\boldsymbol{\beta}^{\boldsymbol{s+1}}$を生成する
2. $p\left(\tau^{s+1}|\boldsymbol{\beta}^{\boldsymbol{s+1}}, y\right) \sim Ga\left(\frac{\nu_1^s}{2}, \frac{\nu_1^s S_1^s}{2}\right)$から$\tau^{s+1}$を生成する
3. $s < N$のとき1に戻る。$s = N$のとき計算終了

線形回帰モデルのベイズ推定

- パラメータの事後平均の点推定値は、ベイズ推定の事前情報とOLSの点推定値との重み付け平均

$$E(\boldsymbol{\beta}|\tau, y) = b_1 = B_1(B_0^{-1}b_0 + \tau X^T y)$$

- 事前情報B_0^{-1}とν_0が非常に小さいときには、事後情報はどうなるだろうか？

線形回帰モデルのベイズ推定

- 事前情報$B_0^{-1} = 0$及び$\nu_0 = 0$のとき、事後情報はそれぞれ以下のようになる

$$b_1 = (X^T X)^{-1} X^T y = \widehat{\boldsymbol{\beta}}$$
$$\nu_1 = \nu_0$$
$$S_1 = \frac{1}{\nu}\left(y - X\widehat{\boldsymbol{\beta}}\right)^T\left(y - X\widehat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2$$

- すなわち、最小二乗解と同じ結果となる

線形回帰モデルのベイズ推定

- 推測統計学による線形回帰モデルの推定は、ベイズ推定において事前情報を与えない場合と同じ結果である
- 事前情報なしでベイズ推定する場合の事前分布を無情報事前分布(non-informative prior)という
- ベイズ統計では未知パラメータ$\boldsymbol{\beta}$がランダム変数であると考えるが、推測統計学では$\hat{\boldsymbol{\beta}}$をランダム変数と考えているにほかならない

線形回帰モデルのベイズ推定

- 無情報事前分布は、積分しても1とならない(定数に収束しない)
- このような事前分布を非正則事前分布(improper prior)という
- 非正則事前分布の例として、一様分布が挙げられる
- 推測統計学による線形回帰モデルの不偏推定量は、事前情報に一様分布を与えた場合のベイズ推定とみなすこともできる

偏差情報量基準
(Deviance Information Criterion: DIC)

- モデルの当てはまりの良さを示す指標
- 未知パラメータθが得られた条件下でデータyが求められる確率$p(y|\theta)$⇒つまり尤度関数
- 対数尤度を用いて得られる次式をデヴィアンスとよぶ

$$D(\theta) = -2logp(y|\theta)$$

偏差情報量基準

- パラメータの点推定値に対するデヴィアンス

$$D(\bar{\theta}) = -2logp(y|\bar{\theta})$$

- デヴィアンスの平均値

$$\overline{D}(\theta) = E_{\theta|y}[-2logp(y|\bar{\theta})]$$

偏差情報量基準

- 有効なパラメータ数

$$pD = \overline{D}(\theta) - D(\bar{\theta})$$

- 偏差情報量基準(DIC)

$$DIC = \overline{D}(\theta) + pD = D(\bar{\theta}) + 2pD$$