線型推測論

第06回 確率分布の仮定と線形仮説の検定(1)

2017/5/18

千葉大学大学院医学研究院

長島 健悟

今回のお話

- 前回まで
  - 基本的には誤差の三条件の仮定のみ
  - 期待値と分散だけに条件を置いて, 具体的な確率分布を仮定しなかった
- 今回
  - 仮説検定と信頼区間を議論するため, 誤差の三条件に加え, 正規性の仮定を置く

正規性の仮定

- 正規性の仮定

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

- 系6-1
  - $Y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$

ゴール

- 一般線型仮説の検定 (General linear hypothesis tests)
- 帰無仮説：$H_0\colon \mathbf{C\beta} = \mathbf{0}$
- 対立仮説：$H_1\colon \mathbf{C\beta} \neq \mathbf{0}$
  - $\mathbf{C}$：$m \times p$次元の対比係数行列
    - $\mathbf{C} = (\mathbf{c}_1' \quad \mathbf{c}_2' \quad \cdots \quad \mathbf{c}_m')'$
  - $(m \times p) \times (p \times 1) = (m \times 1)$なので, m個の仮説を同時に検定できる

単一仮説の検定

- $m = 1$の場合, 単一仮説の検定が得られる
- $H_0 : \mathbf{C\beta} = \mathbf{c}_1'\boldsymbol{\beta} = 0$ 仮説検定になる

$$\mathbf{c}_1'\boldsymbol{\beta} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$= \sum_{j=1}^{p} c_{1j}\beta_j$$

複数仮説の検定

- 一般のmの場合がこれにあたる
- H_0: $\mathbf{C\beta} = \begin{pmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \\ \vdots \\ \mathbf{c}_m' \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \mathbf{c}_1'\boldsymbol{\beta} \\ \mathbf{c}_2'\boldsymbol{\beta} \\ \vdots \\ \mathbf{c}_m'\boldsymbol{\beta} \end{pmatrix} = \mathbf{0}$
- m個の仮説を"同時に"評価できる方法が必要
- なんでF検定が現れるのか？ F検定の意味は？

単一仮説の検定の場合

- ロジックはとても単純である
  - $\mathbf{c}_1'\boldsymbol{\beta} = 0$ を検定したい
  - 帰無仮説のもとで$\dfrac{\mathbf{c}_1'\hat{\boldsymbol{\beta}}}{\sqrt{\widehat{\mathrm{Var}}(\mathbf{c}_1'\hat{\boldsymbol{\beta}})}} \sim t(n-p)$に基づく仮説検定が考えられる
    - $\widehat{\mathrm{Var}}\left(\mathbf{c}_1'\hat{\boldsymbol{\beta}}\right) = s^2\mathbf{c}_1'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_1$

t統計量の導出

- Zが標準正規分布$N(0,1)$に従い，VがZとは独立に自由度νのχ^2分布に従うとき，
$$Y = \frac{Z}{\sqrt{V/\nu}},$$
は自由度νのt分布に従う

t統計量の導出：一標本t検定

- 一標本t検定の場合
  - $Y_i \sim N(\mu, \sigma^2)$
  - $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
  - $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - $\frac{\bar{Y}-\mu}{\sqrt{s^2/n}} \sim t(n-1)$

t統計量の導出：一標本t検定

- $Y = \frac{Z}{\sqrt{V/\nu}}$の形をしている
  - $Z = \frac{\bar{Y}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$
  - $(n-1)\frac{s^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 \sim \chi^2(n-1)$
  - $\frac{Z}{\sqrt{(n-1)\frac{s^2}{\sigma^2}/(n-1)}} = \frac{\bar{Y}-\mu}{\sqrt{s^2/n}} \sim t(n-1)$

カイ二乗分布の復習

- $Y_i \sim N(\mu, \sigma^2)$, $Z_i \sim N(0,1)$のとき
  - 定義：$Z_i^2 \sim \chi^2(1)$, $\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$, $\sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$, $\frac{Y_i-\mu}{\sigma} \sim N(0,1)$
- 偏差平方和の分布 (直感的な説明)
  - $\sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2 - \frac{(\bar{Y}-\mu)^2}{\sigma^2/n} \Leftrightarrow$ $\sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 + \frac{(\bar{Y}-\mu)^2}{\sigma^2/n} = \sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2$
  - $\sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 \sim \chi^2(n-1)$

二次形式の分布

- 本講義では行列による表記を学習する
- $\mathbf{Y}'\mathbf{A}\mathbf{Y}$の形を二次形式と呼ぶのであった
  - $\mathbf{Y}'\mathbf{A}\mathbf{Y} = \sum_i \sum_j a_{ij} y_i y_j$
- 二次形式を用いれば二乗和の形を行列で表現できる
  - $\sum_i y_i^2 = \mathbf{Y}'\mathbf{I}\mathbf{Y}$
  - $\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 = \left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)' \mathbf{I} \left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)$

二次形式の分布

- 一般線型モデルでは二次形式の分布が重要である
- 定理6-1
  - $\mathbf{A}$が冪等行列であり, $r = \mathrm{rank}(\mathbf{A})$であるとする
  - $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$のとき$\mathbf{Z}'\mathbf{A}\mathbf{Z} \sim \chi^2(r)$

冪等行列 (idempotent matrix)

- $\mathbf{A} = \mathbf{A}\mathbf{A}$をみたす正方行列のことを冪等行列とよぶ
  - 冪等行列$\mathbf{A}$については

$$\mathrm{rank}(\mathbf{A}) = \mathrm{tr}(\mathbf{A})$$

  がなりたつ

- 冪等行列の例

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

二次形式の分布

- より一般の場合
- $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$のとき
  - $\frac{\mathbf{Y}'\mathbf{A}\mathbf{Y}}{\sigma^2} \sim \chi^2\left(r, \frac{1}{2\sigma^2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\right)$
    - 非心カイ二乗分布
  - $\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)'\mathbf{A}\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right) \sim \chi^2(r)$

二次形式による偏差平方和の分布

- $\sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2=\sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2-\frac{(\bar{Y}-\mu)^2}{\sigma^2/n}$
- $\sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2=\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)'\mathbf{I}\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)$

二次形式による偏差平方和の分布

- $\mathbf{j} = (1 \quad \cdots \quad 1)', \mathbf{J} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$
- $\frac{(\bar{Y}-\mu)^2}{\sigma^2/n} = n\left\{\frac{1}{n}\mathbf{j}'\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)\right\}^2$

$$= n\left\{\frac{1}{n}\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)\mathbf{j}\right\}\left\{\frac{1}{n}\mathbf{j}'\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)\right\}$$

$$= \left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)'\left(\frac{1}{n}\mathbf{J}\right)\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)$$

二次形式による偏差平方和の分布

- $\sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 = \left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)'\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)$

二次形式による偏差平方和の分布

- $\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)'\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)$の分布を求めよう
  - $\mathbf{I}-\frac{1}{n}\mathbf{J}$は冪等行列である
  - $\mathrm{rank}\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)=\mathrm{tr}\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)=n-1$
  - $\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)'\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)\sim\chi^2(n-1)$

練習問題

- $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I}), \mathbf{X} : n \times p,\ \mathrm{rank}(\mathbf{A}) = 5$
  - $\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)' \mathbf{A} \left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right) \sim \chi^2(5)?$
  - $\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)' \left(\frac{1}{n}\mathbf{J}\right) \left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right) \sim \chi^2(1)?$
  - $\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)' \{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\} \left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right) \sim \chi^2(p)?$
  - $\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)' \{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\} \left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right) \sim \chi^2(n-p)?$

t統計量の導出：練習問題

- $\mathbf{c}_1'\widehat{\boldsymbol{\beta}} = \mathbf{c}_1'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim N(\mathbf{c}_1'\boldsymbol{\beta}, \sigma^2\mathbf{c}_1'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_1)$
  - $\mathrm{E}[\mathbf{c}_1'\widehat{\boldsymbol{\beta}}] = \mathbf{c}_1'\mathrm{E}[\widehat{\boldsymbol{\beta}}]$, $\mathrm{Var}[\mathbf{c}_1'\widehat{\boldsymbol{\beta}}] = \mathbf{c}_1'\mathrm{Cov}[\widehat{\boldsymbol{\beta}}]\mathbf{c}_1$
- したがって，

$$Z = \frac{\mathbf{c}_1'\widehat{\boldsymbol{\beta}} - \mathbf{c}_1'\boldsymbol{\beta}}{\sqrt{\sigma^2\mathbf{c}_1'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_1}} \sim N(0,1)$$

t統計量の導出：練習問題

- $V = (n-p)\frac{s^2}{\sigma^2} \sim \chi^2(n-p)$
- $\frac{Z}{\sqrt{V/(n-p)}} = \frac{\mathbf{c}_1'\hat{\boldsymbol{\beta}} - \mathbf{c}_1'\boldsymbol{\beta}}{\sqrt{\sigma^2\mathbf{c}_1'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_1}} \frac{1}{\sqrt{(n-p)\frac{s^2}{\sigma^2}/(n-p)}} =$
  $\frac{\mathbf{c}_1'\hat{\boldsymbol{\beta}} - \mathbf{c}_1'\boldsymbol{\beta}}{\sqrt{s^2\mathbf{c}_1'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_1}} \sim t(n-p)$
- 特にH_0のもとでは$\mathbf{c}_1'\boldsymbol{\beta} = 0$
  - $\frac{\mathbf{c}_1'\hat{\boldsymbol{\beta}}}{\sqrt{s^2\mathbf{c}_1'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_1}} \sim t(n-p)$

$(n-p)s^2/\sigma^2 \sim \chi^2(n-p)$

- 冪等行列$\mathbf{A}$を用いs^2は二次形式で書ける
  - $s^2 = \frac{1}{n-p}\mathbf{Y}'\mathbf{A}\mathbf{Y}, \mathbf{A} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
  - $\mathrm{rank}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}) = n - p$
- 帰無仮説のもとで$Y_i \sim N(0, \sigma^2)$ $(\boldsymbol{\beta} = \mathbf{0})$
- したがって
  - $(n-p)\frac{s^2}{\sigma^2} = \left(\frac{\mathbf{Y}}{\sigma}\right)' \mathbf{A} \left(\frac{\mathbf{Y}}{\sigma}\right) \sim \chi^2(n-p)$

具体例：練習問題

- $\mathbf{c}_1' = (1 \quad 0 \quad \cdots \quad 0)$
- $\boldsymbol{\beta}' = (\beta_1 \quad \beta_2 \quad \cdots \quad \beta_p)$とする
  - $\mathbf{c}_1'\boldsymbol{\beta} =?$
    - $\mathbf{c}_1'\boldsymbol{\beta} = \beta_1, H_0\colon \beta_1 = 0$
  - $\dfrac{\mathbf{c}_1'\hat{\boldsymbol{\beta}}}{\sqrt{s^2\mathbf{c}_1'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_1}} =?$
    - $\dfrac{\hat{\beta}_1}{\sqrt{s^2 g_1}}, g_1 = (\mathbf{X}'\mathbf{X})^{-1}$の(1,1)要素

冪等行列の性質

- $\mathbf{A}$が冪等行列ならば, その固有値は0または1のみである
  - 固有値・固有ベクトルの定義から：$\mathbf{Ax} = \lambda\mathbf{x}$
  - $\mathbf{Ax} = \mathbf{AAx} = \mathbf{A}\lambda\mathbf{x} = \lambda^2\mathbf{x}$
  - $\lambda\mathbf{x} = \lambda^2\mathbf{x} \Leftrightarrow \lambda(1-\lambda)\mathbf{x} \Leftrightarrow \lambda = 0,1$

冪等行列の性質

- $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots)$：対角要素が各固有値
- $\mathbf{X} = (\mathbf{x}_1, \dots)$：各固有値ベクトルならべた行列 (正則行列, $\mathbf{X}'\mathbf{X} = \mathbf{I}$)
- 固有値分解：$\mathbf{AX} = \mathbf{X\Lambda} \Leftrightarrow \mathbf{X}'\mathbf{AX} = \mathbf{\Lambda}$
  - $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{X}'\mathbf{AX}) = \mathrm{rank}(\mathbf{\Lambda})$
    正則行列をかけてもランクは不変
  - $\mathrm{rank}(\mathbf{\Lambda}) = \mathrm{tr}(\mathbf{\Lambda}) = \mathrm{tr}(\mathbf{\Lambda X}'\mathbf{X}) = \mathrm{tr}(\mathbf{A})$
    冪等行列の固有値は0または1のみ