

線型推測論

第06回 確率分布の仮定と線形仮説の検定(1)

2022/5/19

慶應義塾大学病院

長島 健悟

今回のお話

- 前回まで (点推定のみ)
 - 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 : $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$
- 対立仮説 : $H_1: \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}$
- $\mathbf{C}$: $m \times p$ 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
- $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{c}'_1\boldsymbol{\beta} = 0$ 仮説検定になる

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

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

複数仮説の検定

- 一般の m の場合

- $H_0: \mathbf{C}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{c}'_1 \\ \mathbf{c}'_2 \\ \vdots \\ \mathbf{c}'_m \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \mathbf{c}'_1\boldsymbol{\beta} \\ \mathbf{c}'_2\boldsymbol{\beta} \\ \vdots \\ \mathbf{c}'_m\boldsymbol{\beta} \end{pmatrix} = \mathbf{0}$

- m 個の仮説を“同時に”評価する方法
 - これは F 検定になる
- なぜ F 検定が現れるのか？ F 検定の意味は？

仮説検定の基本

- 仮説検定の作り方
 - 例 : $H_0: \beta = \beta_0, H_1: \beta \neq 0$ ($\beta = \beta$) の検定
 - H_1 のもとでのパラメータ推定量 $\hat{\beta}$ の(漸近)分布 $F(\hat{\beta}; \beta)$ を求める
 - 推定量 $\hat{\beta}$ の値が帰無仮説のもとでの分布 $F(\hat{\beta}; \beta = \beta_0)$ で極端な値になっているかどうかを評価
 - 特に $E[\hat{\beta}] = \beta$ なら, $G(\hat{\beta} - \beta)$ を使うと簡単

単一仮説の検定と仮説検定の基本

- 知りたい仮説の統計量 (推定量) の分布が分かれば, 仮説検定を構成できる
- $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{c}'_1\boldsymbol{\beta} = 0$ の単一仮説の検定
 - $\mathbf{c}'_1\boldsymbol{\beta}$ が 0 かどうかの検定で, 推定量 $\mathbf{c}'_1\hat{\boldsymbol{\beta}}$ をもとに統計量を構成する
 - $E[\mathbf{c}'_1\hat{\boldsymbol{\beta}}] = \mathbf{c}'_1\boldsymbol{\beta}$ である
 - $\mathbf{c}'_1\hat{\boldsymbol{\beta}} - \mathbf{c}'_1\boldsymbol{\beta} = \mathbf{c}'_1\hat{\boldsymbol{\epsilon}}$ の分布が必要

一番単純な単一仮説の検定の場合

- 一番単純な場合を考える

- $\mathbf{c}'_1 = (1), p = 1, \mathbf{X} = \mathbf{1} = (1, 1, \dots, 1)'$

$$H_0: \mathbf{c}'_1 \boldsymbol{\beta} = \beta_1 = 0$$

$$\mathbf{Y} \sim N(\mathbf{1}\beta_1, \sigma^2 \mathbf{I})$$

- $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

- $\mathbf{c}'_1 \hat{\boldsymbol{\beta}} = 1(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{Y} = n^{-1} \sum_{i=1}^n Y_i = \bar{Y}$

- この分布が分かればよい？

一番単純な単一仮説の検定の場合

- $\mathbf{c}_1' \hat{\boldsymbol{\beta}}$ の分布は何になる？
 - 正規分布の和なので, 正規分布に従う
 - 仮定より, $Y_i \sim N(\beta_1, \sigma^2)$
 - $Y_i/n \sim N(\beta_1/n, \sigma^2/n^2)$
 - $\sum_{i=1}^n Y_i /n \sim N(\beta_1, \sigma^2/n)$

帰無仮説のもとでの分布

$$\mathbf{c}'_1 \hat{\boldsymbol{\beta}} = \sum_{i=1}^n Y_i / n \sim N(0, \sigma^2 / n)$$

- これで統計量の分布が完全にわかった？
- × : σ^2 は未知パラメータ, これもなんとかしたい
- 若干の準備をしておく

$$\frac{\mathbf{c}'_1 \hat{\boldsymbol{\beta}}}{\sqrt{\sigma^2 / n}} \sim N(0, 1)$$

t 統計量

- 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,

$$T = \frac{Z}{\sqrt{V/\nu}},$$

は自由度 ν の t 分布に従う

- χ^2 分布に従う確率変数を見つけてあげればいい

$(n-1)s^2/\sigma^2$ の分布

- 今までにもう一つの推定量が登場していま

した：
$$s^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-p}$$

- 一番簡単なとき

- $$s^2 = \frac{(\mathbf{Y} - \mathbf{1}\hat{\beta})'(\mathbf{Y} - \mathbf{1}\hat{\beta})}{n-1} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$$
$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{\sigma} \right)^2 \sim ?$$

- 標準正規分布の二乗和の形になっている

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)
 - $\sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{Y_i - \mu}{\sigma}\right)^2 - \frac{(\bar{Y} - \mu)^2}{\sigma^2/n} \Leftrightarrow$
 $\sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2 + \frac{(\bar{Y} - \mu)^2}{\sigma^2/n} = \sum_{i=1}^n \left(\frac{Y_i - \mu}{\sigma}\right)^2$
 - $\sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2 \sim \chi^2(n - 1)$

t 統計量 (再掲)

- 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,

$$T = \frac{Z}{\sqrt{V/\nu}},$$

は自由度 ν の t 分布に従う

t 統計量の導出：一番単純な場合

- $Z \sim N(0, 1), V \sim \chi^2(\nu)$ なら $T = \frac{Z}{\sqrt{V/\nu}} \sim t(\nu)$

- 一番単純な場合

- $Z : \frac{\mathbf{c}_1' \hat{\boldsymbol{\beta}}}{\sqrt{\sigma^2/n}} \sim N(0, 1)$

- $V : \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

- $T : \frac{\mathbf{c}_1' \hat{\boldsymbol{\beta}}}{\sqrt{\sigma^2/n}} \left(\frac{(n-1)s^2}{(n-1)\sigma^2} \right)^{-1/2} = \frac{\mathbf{c}_1' \hat{\boldsymbol{\beta}}}{\sqrt{s^2/n}} \sim t(n-1)$

- $\sqrt{\widehat{\text{Var}}(\mathbf{c}_1' \hat{\boldsymbol{\beta}})} = \sqrt{s^2 \mathbf{c}_1' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{c}_1} = \sqrt{s^2/n}$

結果の一般化

- 一番簡単な場合は説明できた： $\mathbf{c}'_1 = (1)$,
 $p = 1, \mathbf{X} = \mathbf{1} = (1, 1, \dots, 1)'$
- 今度はこれの一般化を考えたい
- 行列による表記ができないだろうか？
 - $\mathbf{c}'_1 \hat{\boldsymbol{\beta}}$ の方は正規分布だったので簡単そう
 - s^2 は行列表記できたが, この分布はどうなるのだろうか？

$\mathbf{c}'_1 \hat{\boldsymbol{\beta}}$ の分布

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

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

二次形式の分布

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

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, $r = \text{rank}(\mathbf{A})$ 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)

- $A = AA$ をみたす正方行列のことを冪等行列とよぶ

- 冪等行列 A については

$$\text{rank}(A) = \text{tr}(A)$$

がなりたつ

- 冪等行列の例

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

二次形式の分布

- ちょっと易しめのケースから検討
- $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ のとき
 - $\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)' \mathbf{A} \left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right) \sim \chi^2(r)$
- ちなみに
 - $\frac{\mathbf{Y}' \mathbf{A} \mathbf{Y}}{\sigma^2} \sim \chi^2\left(r, \frac{1}{2\sigma^2} \boldsymbol{\mu}' \mathbf{A} \boldsymbol{\mu}\right)$: 非心カイ二乗分布

二次形式による偏差平方和の分布

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

二次形式による偏差平方和の分布

- $\mathbf{j} = (1 \quad \dots \quad 1)'$, $\mathbf{J} = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}$

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

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
- $\text{rank}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) = \text{tr}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) = n - 1$
- $\left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right)' \left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) \left(\frac{\mathbf{Y}-\boldsymbol{\mu}}{\sigma}\right) \sim \chi^2(n - 1)$

練習問題

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

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

- 冪等行列 $\mathbf{A}$ を用い s^2 は二次形式で書けた
 - $s^2 = \frac{1}{n-p} \mathbf{Y}' \mathbf{A} \mathbf{Y} = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{A} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$
 - $\mathbf{A} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
 - $\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A}) = n - p$
- したがって
 - $(n - p) \frac{s^2}{\sigma^2} = \mathbf{Z}' \mathbf{A} \mathbf{Z} \sim \chi^2(n - p)$
 - $\mathbf{Z} = \frac{\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}}{\sigma}$

t 統計量の導出：練習問題

- $Z = \frac{\mathbf{c}_1' \hat{\boldsymbol{\beta}} - \mathbf{c}_1' \boldsymbol{\beta}}{\sqrt{\sigma^2 \mathbf{c}_1' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{c}_1}} \sim N(0,1)$ [スライド17]

- $V = (n - p) \frac{s^2}{\sigma^2} \sim \chi^2(n - p)$

- $$\frac{Z}{\sqrt{V/(n-p)}} = \frac{\mathbf{c}_1' \hat{\boldsymbol{\beta}}}{\sqrt{\sigma^2 \mathbf{c}_1' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{c}_1}} \frac{1}{\sqrt{(n-p) \frac{s^2}{\sigma^2} / (n-p)}} =$$
$$\frac{\mathbf{c}_1' \hat{\boldsymbol{\beta}}}{\sqrt{s^2 \mathbf{c}_1' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{c}_1}} = \frac{\mathbf{c}_1' \hat{\boldsymbol{\beta}}}{\sqrt{\widehat{\text{Var}}(\mathbf{c}_1' \hat{\boldsymbol{\beta}})}} \sim t(n - p)$$

- $\sqrt{\widehat{\text{Var}}(\mathbf{c}_1' \hat{\boldsymbol{\beta}})} = \sqrt{s^2 \mathbf{c}_1' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{c}_1}$

具体例：練習問題

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

冪等行列の性質

- 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：
$$A\mathbf{x} = \lambda\mathbf{x}$$
- $A\mathbf{x} = AA\mathbf{x} = A\lambda\mathbf{x} = \lambda^2\mathbf{x}$
- $\lambda\mathbf{x} = \lambda^2\mathbf{x} \Leftrightarrow \lambda(1 - \lambda)\mathbf{x} \Leftrightarrow \lambda = 0, 1$

冪等行列の性質

- $\Lambda = \text{diag}(\lambda_1, \dots)$: 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 : $\mathbf{A}\mathbf{X} = \mathbf{X}\Lambda \Leftrightarrow \mathbf{X}'\mathbf{A}\mathbf{X} = \Lambda$
 - $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{X}'\mathbf{A}\mathbf{X}) = \text{rank}(\Lambda)$
正則行列をかけてもランクは不変
 - $\text{rank}(\Lambda) = \text{tr}(\Lambda) = \text{tr}(\Lambda\mathbf{X}'\mathbf{X}) = \text{tr}(\mathbf{A})$
冪等行列の固有値は0または1のみ