

第 26 回 分散分析と決定係数 (13.4)

村澤 康友

2024 年 1 月 22 日

今日のポイント

- 1 元配置分散分析は 2 標本問題の k 標本問題への拡張. 各群の母平均に対する要因効果の有無の検定は, 各群のダミー変数を説明変数とし, すべての回帰係数が等しいかどうかを F 検定すればよい. 1 元配置分散分析表は F 検定の理解に役立つ.
- 2 元配置分散分析は 2 つの要因効果を分析する. 交互作用を捉えるには各群のダミー変数の交差項を説明変数に加える.
- 総変動 (TSS) は回帰変動 (ESS) と残差変動 (RSS) に分解できる ($TSS = ESS + RSS$). 決定係数は $R^2 := ESS/TSS = 1 - RSS/TSS$. 自由度修正済み決定係数は $\bar{R}^2 := 1 - [RSS/(n-k)]/[TSS/(n-1)]$.
- y_i と \hat{y}_i の相関係数を y_i と x_i の重相関係数という. 重相関係数の 2 乗 = 決定係数.

- 2.3 自由度修正済み決定係数 5
- 2.4 重相関係数 (pp. 63, 272) 5

- 3 今日のキーワード 6
- 4 次回までの準備 6

1 分散分析 (ANOVA)

1.1 1 元配置分散分析

2 標本問題を k 標本問題に拡張し, k 個の正規母集団 $N(\mu_1, \sigma^2), \dots, N(\mu_k, \sigma^2)$ の母平均を比較したい. この問題を回帰分析で考える. 各母集団 (群) から独立に抽出した大きさ n_1, \dots, n_k の無作為標本を $(y_{1,1}, \dots, y_{1,n_1}), \dots, (y_{k,1}, \dots, y_{k,n_k})$ とする. $n := n_1 + \dots + n_k$ とする.

定義 1. μ_1, \dots, μ_k の総平均は

$$\mu := \frac{\mu_1 + \dots + \mu_k}{k}$$

定義 2. μ_1, \dots, μ_k が異なる原因を **因子 (要因)** という.

例 1. 薬の投与, 教育.

定義 3. $h = 1, \dots, k$ を因子の水準という.

例 2. 処置の有無や程度, 教育水準 (最終学歴).

定義 4. $\alpha_h := \mu_h - \mu$ を水準 h の効果という.

例 3. 処置効果, 学歴収益率.

定義 5. 1 元配置分散分析モデルは $h = 1, \dots, k, i = 1, \dots, n_h$ について

$$y_{h,i} = \mu_h + u_{h,i}$$
$$u_{h,i} \sim N(0, \sigma^2)$$

目次

| | | |
|-----|------------------------------|---|
| 1 | 分散分析 (ANOVA) | 1 |
| 1.1 | 1 元配置分散分析 | 1 |
| 1.2 | ダミー変数 | 2 |
| 1.3 | 群間変動と群内変動 | 2 |
| 1.4 | 要因効果の F 検定 | 2 |
| 1.5 | 2 元配置分散分析 | 3 |
| 2 | 決定係数と重相関係数 | 4 |
| 2.1 | 回帰残差 (p. 262) | 4 |
| 2.2 | 決定係数 (pp. 60, 272) | 4 |

または

$$y_{h,i} = \mu + \alpha_h + u_{h,i}$$

$$u_{h,i} \sim N(0, \sigma^2)$$

ただし $\alpha_1 + \dots + \alpha_k = 0$.

1.2 ダミー変数

$h = 1, \dots, k, i = 1, \dots, n_h$ について $x_{h,i} := h$ として群を表すと

$$y_{h,i} = \sum_{j=1}^k \mu_j [x_{h,i} = j] + u_{h,i}$$

$$= \mu_1 + \sum_{j=2}^k (\mu_j - \mu_1) [x_{h,i} = j] + u_{h,i}$$

ただし $[\cdot]$ は中の命題が真なら 1, 偽なら 0 を返す指示関数.

定義 6. ある条件に該当するなら 1, 該当しないなら 0 とした変数を **ダミー変数** という.

例 4. 女性ダミー (女性なら 1, 男性なら 0), 大卒ダミー (大卒なら 1, それ以外なら 0).

注 1. 1 元配置分散分析モデルは k 個の群ダミー変数 (または定数項と $k-1$ 個の群ダミー変数) を説明変数とした重回帰モデルで表せる.

1.3 群間変動と群内変動

各群の標本平均は $h = 1, \dots, k$ について

$$\bar{y}_h := \frac{1}{n_h} \sum_{i=1}^{n_h} y_{h,i}$$

全群の標本平均は

$$\bar{y} := \frac{1}{n} \sum_{h=1}^k \sum_{i=1}^{n_h} y_{h,i}$$

定義 7. 全 (総) 変動は

$$S := \sum_{h=1}^k \sum_{i=1}^{n_h} (y_{h,i} - \bar{y})^2$$

定義 8. 群間変動は

$$S_b := \sum_{h=1}^k \sum_{i=1}^{n_h} (\bar{y}_h - \bar{y})^2$$

注 2. $(\bar{y}_h - \bar{y})^2$ は i に依存しないので

$$S_b := \sum_{h=1}^k n_h (\bar{y}_h - \bar{y})^2$$

定義 9. 群内変動は

$$S_w := \sum_{h=1}^k \sum_{i=1}^{n_h} (y_{h,i} - \bar{y}_h)^2$$

定理 1.

$$S = S_b + S_w$$

証明.

$$S = \sum_{h=1}^k \sum_{i=1}^{n_h} (y_{h,i} - \bar{y}_h + \bar{y}_h - \bar{y})^2$$

$$= \sum_{h=1}^k \sum_{i=1}^{n_h} (y_{h,i} - \bar{y}_h)^2$$

$$+ 2 \sum_{h=1}^k \sum_{i=1}^{n_h} (y_{h,i} - \bar{y}_h) (\bar{y}_h - \bar{y})$$

$$+ \sum_{h=1}^k \sum_{i=1}^{n_h} (\bar{y}_h - \bar{y})^2$$

$$= S_w + 2 \sum_{h=1}^k (\bar{y}_h - \bar{y}) \sum_{i=1}^{n_h} (y_{h,i} - \bar{y}_h) + S_b$$

ここで

$$\sum_{i=1}^{n_h} (y_{h,i} - \bar{y}_h) = \sum_{i=1}^{n_h} y_{h,i} - n_h \bar{y}_h$$

$$= \sum_{i=1}^{n_h} y_{h,i} - \sum_{i=1}^{n_h} y_{h,i}$$

$$= 0$$

したがって第 2 項は 0. □

1.4 要因効果の F 検定

次の検定問題を考える.

$$H_0 : \mu_1 = \dots = \mu_k$$

$$\text{vs } H_1 : \mu_h \neq \mu \text{ for some } h = 1, \dots, k$$

$k = 2$ なら 2 標本問題の母平均の差の両側検定.

補題 1. H_0 の下で

$$\sum_{h=1}^k \frac{n_h (\bar{y}_h - \mu)^2}{\sigma^2} \sim \chi^2(k)$$

証明. H_0 の下で $h = 1, \dots, k$ について

$$\bar{y}_h \sim N\left(\mu, \frac{\sigma^2}{n_h}\right)$$

すなわち

$$\frac{\bar{y}_h - \mu}{\sqrt{\sigma^2/n_h}} \sim N(0, 1)$$

または

$$\frac{(\bar{y}_h - \mu)^2}{\sigma^2/n_h} \sim \chi^2(1)$$

各群からの標本は独立なので

$$\sum_{h=1}^k \frac{(\bar{y}_h - \mu)^2}{\sigma^2/n_h} \sim \chi^2(k)$$

注 3. 各群の標本平均の標本分布から導出.

定理 2. H_0 の下で

$$\frac{S_b}{\sigma^2} \sim \chi^2(k-1)$$

証明. 補題の μ を \bar{y} に置き換えると

$$\sum_{h=1}^k \frac{n_h(\bar{y}_h - \bar{y})^2}{\sigma^2} \sim \chi^2(k-1)$$

(詳細は略). 左辺は S_b/σ^2 .

系 1.

$$E\left(\frac{S_b}{k-1}\right) = \sigma^2$$

証明. 定理より $E(S_b/\sigma^2) = k-1$.

定理 3.

$$\frac{S_w}{\sigma^2} \sim \chi^2(n-k)$$

証明. $h = 1, \dots, k$ について

$$\frac{\sum_{i=1}^{n_h} (y_{h,i} - \bar{y}_h)^2}{\sigma^2} \sim \chi^2(n_h - 1)$$

各群からの標本は独立なので

$$\sum_{h=1}^k \frac{\sum_{i=1}^{n_h} (y_{h,i} - \bar{y}_h)^2}{\sigma^2} \sim \chi^2\left(\sum_{h=1}^k (n_h - 1)\right)$$

注 4. 各群の標本分散の標本分布から導出.

系 2.

$$E\left(\frac{S_w}{n-k}\right) = \sigma^2$$

証明. 定理より $E(S_w/\sigma^2) = n-k$. □

定理 4. S_b と S_w は独立.

証明. 「統計学入門」の範囲を超えるので省略. □

定理 5. H_0 の下で

$$\frac{S_b/(k-1)}{S_w/(n-k)} \sim F(k-1, n-k)$$

□ 証明. 前 3 定理より明らか. □

注 5. 1 元配置分散分析の考え方は, 1 元配置分散分析表に整理できる (表 1).

注 6. 定数項と $k-1$ 個の群ダミー変数を説明変数とした重回帰モデルの回帰係数の F 検定とも理解できる.

1.5 2 元配置分散分析

2 つの因子 A, B を考える (例えば性別と最終学歴). 両者の水準の効果は独立とは限らない. A の水準を $j = 1, \dots, J$, B の水準を $k = 1, \dots, K$ とする.

定義 10. 2 元配置分散分析モデルは $j = 1, \dots, J$, $k = 1, \dots, K$, $i = 1, \dots, n_{j,k}$ について

$$y_{j,k,i} = \mu + \alpha_j + \beta_k + \gamma_{j,k} + u_{j,k,i}$$

$$u_{j,k,i} \sim N(0, \sigma^2)$$

ただし

$$\alpha_1 + \dots + \alpha_J = 0$$

$$\beta_1 + \dots + \beta_K = 0$$

$$\gamma_{j,1} + \dots + \gamma_{j,K} = 0, \quad j = 1, \dots, J$$

$$\gamma_{1,k} + \dots + \gamma_{J,k} = 0, \quad k = 1, \dots, K$$

注 7. α_j, β_k を主効果, $\gamma_{j,k}$ を交互作用という.

□ 注 8. $j = 1, \dots, J$, $k = 1, \dots, K$, $i = 1, \dots, n_{j,k}$

表 1 1元配置分散分析表

| | 変動 | 自由度 | 分散 | F 値 |
|----|-------|---------|---------------|-------------------------------|
| 群間 | S_b | $k - 1$ | $S_b/(k - 1)$ | $[S_b/(k - 1)]/[S_w/(n - k)]$ |
| 群内 | S_w | $n - k$ | $S_w/(n - k)$ | |
| 計 | S | $n - 1$ | $S/(n - 1)$ | |

について $x_{j,k,i} := j$, $z_{j,k,i} = k$ とすると

$$\begin{aligned}
 y_{j,k,i} &= \mu + \sum_{j'=1}^J \alpha_j [x_{j,k,i} = j'] \\
 &+ \sum_{k'=1}^K \beta_j [z_{j,k,i} = k'] \\
 &+ \sum_{j'=1}^J \sum_{k'=1}^K \gamma_{j,k} [x_{j,k,i} = j'] [z_{j,k,i} = k'] \\
 &+ u_{j,k,i}
 \end{aligned}$$

1階の条件より

$$\begin{aligned}
 \sum_{i=1}^n (y_i - a^* - b^* x_i) &= 0 \\
 \sum_{i=1}^n x_i (y_i - a^* - b^* x_i) &= 0
 \end{aligned}$$

□

すなわち $J + K$ 個の群ダミー変数と JK 個の交差項を説明変数とした重回帰モデルとなる。

2 決定係数と重相関係数

2.1 回帰残差 (p. 262)

2変量データを $((y_1, x_1), \dots, (y_n, x_n))$ とする。 y_i の x_i 上への単回帰モデルは

$$E(y_i | x_i) = \alpha + \beta x_i$$

(α, β) の OLS 推定量 (値) を (a^*, b^*) , 回帰予測を $\hat{y}_i := a^* + b^* x_i$, 回帰残差を $e_i := y_i - \hat{y}_i$ とする。

補題 2.

$$\begin{aligned}
 \sum_{i=1}^n e_i &= 0 \\
 \sum_{i=1}^n x_i e_i &= 0
 \end{aligned}$$

証明. OLS 問題は

$$\begin{aligned}
 \min_{a,b} \sum_{i=1}^n (y_i - a - b x_i)^2 \\
 \text{and } a, b \in \mathbb{R}
 \end{aligned}$$

2.2 決定係数 (pp. 60, 272)

定義 11. (y_1, \dots, y_n) の全 (総) 変動 (Total Sum of Squares, TSS) は

$$TSS := \sum_{i=1}^n (y_i - \bar{y})^2$$

定義 12. (y_1, \dots, y_n) の回帰変動 (Explained Sum of Squares, ESS) は

$$ESS := \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

注 9. 分散分析の群間変動.

定義 13. (y_1, \dots, y_n) の残差変動 (Residual Sum of Squares, RSS) は

$$RSS := \sum_{i=1}^n e_i^2$$

注 10. 分散分析の群内変動.

定理 6.

$$TSS = ESS + RSS$$

証明. 総変動は

$$\begin{aligned} \text{TSS} &:= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \sum_{i=1}^n [(\hat{y}_i - \bar{y}) + e_i]^2 \\ &= \sum_{i=1}^n [(\hat{y}_i - \bar{y})^2 + 2(\hat{y}_i - \bar{y})e_i + e_i^2] \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (\hat{y}_i - \bar{y})e_i + \sum_{i=1}^n e_i^2 \end{aligned}$$

補題より

$$\begin{aligned} \sum_{i=1}^n (\hat{y}_i - \bar{y})e_i &= \sum_{i=1}^n [(a^* + b^*x_i) - (a^* + b^*\bar{x})]e_i \\ &= b^* \sum_{i=1}^n (x_i - \bar{x})e_i \\ &= b^* \sum_{i=1}^n x_i e_i - b^* \bar{x} \sum_{i=1}^n e_i \\ &= 0 \end{aligned}$$

□

注 11. 重回帰の場合も同様.

定義 14. 回帰の決定係数は

$$R^2 := \frac{\text{ESS}}{\text{TSS}}$$

2.3 自由度修正済み決定係数

前定理より

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

説明変数の数 (定数項を含む) を k とすると, RSS は k の減少関数. また一般に $k \geq n$ なら RSS は 0. したがって R^2 は説明変数の選択に役立たない.

定義 15. 自由度修正済み決定係数は

$$\bar{R}^2 := 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)}$$

注 12. 無作為標本なら

$$E \left(\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \right) = \text{var}(y_i)$$

古典的線形回帰モデルなら

$$E \left(\frac{1}{n-k} \sum_{i=1}^n e_i^2 \right) = \text{var}(u_i)$$

したがって \bar{R}^2 は $1 - \text{var}(u_i) / \text{var}(y_i)$ の推定量 (値) となっている. ただし

$$\begin{aligned} E(\bar{R}^2) &= 1 - E \left(\frac{[1/(n-k)] \sum_{i=1}^n e_i^2}{[1/(n-1)] \sum_{i=1}^n (y_i - \bar{y})^2} \right) \\ &\neq 1 - \frac{E([1/(n-k)] \sum_{i=1}^n e_i^2)}{E([1/(n-1)] \sum_{i=1}^n (y_i - \bar{y})^2)} \\ &= 1 - \frac{\text{var}(u_i)}{\text{var}(y_i)} \end{aligned}$$

2.4 重相関係数 (pp. 63, 272)

定義 16. y_i と \hat{y}_i の相関係数を, y_i と x_i の重相関係数という.

注 13. 重回帰で y_i と $(x_{i,1}, \dots, x_{i,k})$ の関係の強さを測る. 単回帰なら重相関係数 = 相関係数の絶対値.

定理 7. 決定係数 $R^2 =$ 重相関係数 R の 2 乗.

証明. $(\hat{y}_1, \dots, \hat{y}_n)$ の平均は

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \hat{y}_i &= \frac{1}{n} \sum_{i=1}^n (a^* + b^*x_i) \\ &= a^* + b^*\bar{x} \\ &= \bar{y} \end{aligned}$$

$((y_1, \hat{y}_1), \dots, (y_n, \hat{y}_n))$ の共分散は

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^n [(\hat{y}_i - \bar{y}) + e_i](\hat{y}_i - \bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \frac{1}{n} \sum_{i=1}^n e_i(\hat{y}_i - \bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \end{aligned}$$

$((y_1, \hat{y}_1), \dots, (y_n, \hat{y}_n))$ の相関係数は

$$\begin{aligned} & \frac{(1/n) \sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{(1/n) \sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{(1/n) \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}} \\ &= \frac{(1/n) \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sqrt{(1/n) \sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{(1/n) \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}} \\ &= \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \sqrt{\frac{\text{ESS}}{\text{TSS}}} \end{aligned}$$

□

3 今日のキーワード

総平均, 因子 (要因), (因子の) 水準, (水準の) 効果, 1 元配置分散分析モデル, ダミー変数, 全 (総) 変動, 群間変動, 群内変動, 1 元配置分散分析表, 2 元配置分散分析モデル, 主効果, 交互作用, 回帰変動, 残差変動, 決定係数, 自由度修正済み決定係数, 重相関係数

4 次回までの準備

復習 教科書第 13 章 4 節, 復習テスト 26

試験 (1) 教科書を読む (2) 用語の定義を覚える (3) 復習テストを自力で解く (4) 過去問に挑戦