回帰分析・t 検定・分散分析

Naoto Yamashita

October 2023

1　はじめに

回帰分析・t 検定・分散分析は，初歩的な統計解析の手法として学習されることが多い．これらの手法は，それぞれが別々に紹介されることもあり，手法間に密接な関係があることはあまり知られていない．本稿では，各種法の理論的な関係を説明することで，3 つの手法の統一的理解を試みる．

2　回帰分析

最も簡単なケースとしての，単回帰分析を取り上げる．単回帰分析は，単一の従属変数と単一の独立変数がペアになったデータに対して，独立変数が従属変数に与える影響を考察したり，従属変数の将来の値を予測するために使われる多変量データ解析法である．単回帰分析は

$$y_i = \alpha + x_i\beta + \epsilon_i \tag{1}$$

というモデルに基づき，切片 α と回帰係数 β を推定する．ここで (y_i, x_i) は第 i 個体 $(i = 1, \cdots, n)$ に対する従属変数と独立変数のペアであり，ϵ_i は予測の残差を表す．α と β は，残差の平方和 $\sum_{i=1}^{n} \epsilon_i^2$ を最小にするように定める（最小二乗法）．α と β の推定量は次式で与えられる．

$$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta},\ \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

また，$p(> 0)$ 個の独立変数を用いる重回帰分析の場合は，切片に対応する 1 列目を含む $n \times (p+1)$ の独立変数行列 $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} 1 & & & & \\ 1 & & & & \\ \vdots & \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ 1 & & & & \\ 1 & & & & \end{bmatrix} \tag{3}$$

を用いる．ここで，$\mathbf{x}_j$ は j 番目の独立変数に関する観測値を並べた n 次元ベクトルである．$\boldsymbol{\theta}$ の第 1 要素が切片であり，第 2 要素以降が p 個の独立変数に対応する偏回帰係数である．また，従属変数に関する観測値を並べた n 次元ベクトルを $\mathbf{y}$ とする．これらを用いて，切片と偏回帰係数ベクトルを含む $(p+1)$ 次元パラメータ行列 $\boldsymbol{\theta}$ の最小二乗推定量は

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{4}$$

で得られる．

3 t 検定

t 検定は，一般には 2 群の平均値差の検定に用いられる手法である．回帰分析におけるデータの記法を用いれば，y_i が検定の対象とする変数（錯視量など）x_i は水準（例えば実験条件）をコーディングしたダミー変数とすれば良い．具体的には，x_i は 0 か 1 の値を取り，実験条件を例にとれば，個体 i が統制条件に割り当てられれば 0，実験条件に割り当てられれば 1 とする．以降の説明では，この実験条件を例として用いる．

t 検定では，$x_i = 0$ の平均 $\bar{y}_0$ と，$x_i = 1$ の平均 $\bar{y}_1$ の間に差を統計的仮説検定の対象とする．検定のためには，検定統計量 t を次式で計算する．

$$t = \frac{\bar{y}_1 - \bar{y}_0}{s^* \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} \tag{5}$$

ここで，s^* はプールした分散と呼ばれ，s_0^2 および s_1^2 は統制条件と実験条件に所属する個体に関する分散を，n_0 および n_1 は各条件の個体数を用いて

$$\sqrt{\frac{n_0 s_0^2 + n_1 s_1^2}{n_0 + n_1 - 2}} \tag{6}$$

と定義される．帰無仮説の元では t は自由度 $n_0 + n_1 - 2$ の t 分布に従うことを利用して，この t 値が有意水準 α の棄却域に属するかどうかで，帰無仮説を棄却するか保持するかを判断する．

4 分散分析

分散分析は，一般的には，水準数が 3 以上の場合の平均値の差の検定として導入される．ここでは例として，実験条件が二つに増え，$x_i = 2$ の場合，実験条件 2 を表すものとする．分散分析では，群内平均 $\bar{y}_0$, $\bar{y}_1$, $\bar{y}_2$ を用いて，検定統計量 F を以下の流れで計算する．まず，全体平均 $\bar{y}$ を用いて，全体平方和

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{7}$$

を計算する．この値は各データが全体平均からどれぐらい離れているかを定量化するものである．さらに，群間平方和

$$SS_B = n_0(\bar{y}_0 - \bar{y})^2 + n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 \tag{8}$$

を求める．これは，群内平均が全体平均とどれぐらい乖離しているかを表す量である．最後に，群内平方和

$$\begin{aligned} SS_W &= \sum_{i \in C_0} (y_i - \bar{y}_0)^2 + \sum_{i \in C_1} (y_i - \bar{y}_1)^2 + \sum_{i \in C_2} (y_i - \bar{y}_2)^2 \\ &= SS_T - SS_B \end{aligned} \tag{9}$$

を計算する．ここで C_0, C_1, C_2 は，各実験水準に所属する個体番号の集合を表す．水準によって y_i の値が異なるということは，群内の平均値が十分に異なることを表す．この時，群内平方和と比べて群間平方和が大きくなる．よって，両平方和をそれぞれの自由度で除した平均平方和を求め（自由度で除する理由は，水準やデータ数が増えると，平均値の乖離とは無関係に平方和が増大するからである），それらの比をとれば，群内の平均値が異なる程度を定量化できる．この考え

方に基づき，検定統計量 F を次式で計算する．

$$F = \frac{\frac{SS_B}{D-1}}{\frac{SS_W}{n-D}} \tag{10}$$

ここで D は水準の数（上の例では $D = 3$）を表す．帰無仮説が真の元では，F は自由度 $(D-1, n-D-1)$ の F 分布に従う．このことを利用して，t 検定と同じロジックにより帰無仮説を棄却するかどうかを判断する．

5 t 検定と回帰分析の関係

x_i を 0 または 1 の 2 値カテゴリ変数とした回帰分析が，t 検定と同等であることを以下に示す．

まず，$x_1, \cdots, x_i$ の平均 $\bar{x}$ は，x_i が 0 と 1 の 2 値であることに注目すると

$$\bar{x} = \frac{n_1}{n} \tag{11}$$

と書ける．これを用いると，(2) の $\hat{\beta}$ の分母は以下のように書ける．

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i \in C_0}\left(0 - \frac{n_1}{n}\right)(y_i - \bar{y}) + \sum_{i \in C_1}\left(1 - \frac{n_1}{n}\right)(y_i - \bar{y}) \\
&= -\frac{n_1}{n}\sum_{i \in C_0} y_i + \frac{n_0 n_1}{n}\bar{y} + \frac{n_0}{n}\sum_{i \in C_1} y_i - \frac{n_0 n_1}{n}\bar{y} \\
&= \frac{n_0 n_1}{n}(\bar{y}_1 - \bar{y}_0) \qquad (12)
\end{aligned}$$

さらに，分子は

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i \in C_0}\left(0 - \frac{n_1}{n}\right)^2 + \sum_{i \in C_1}\left(1 - \frac{n_1}{n}\right)^2 \qquad (13) \\
&= \frac{n_0 n_1^2}{n^2} + \frac{n_0^2 n_1}{n^2} \\
&= \frac{n_0 n_1}{n^2}(n_0 + n_1) \\
&= \frac{n_0 n_1}{n} \qquad (14)
\end{aligned}$$

以上を用いれば，β の推定量は

$$\hat{\beta} = \frac{\frac{n_0 n_1}{n}(\bar{y}_1 - \bar{y}_0)}{\frac{n_0 n_1}{n}} = \bar{y}_1 - \bar{y}_0 \tag{15}$$

と書ける．この $\hat{\beta}$ は実験群の平均 $\bar{y}_1$ から統制群の平均 $\bar{y}_0$ を引いたもの，すなわち群内平均の差を表す．

また，切片である α の推定量は

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \bar{x}\hat{\beta} \\
&= \bar{y} - \frac{n_1}{n}(\bar{y}_1 - \bar{y}_0) \\
&= \frac{1}{n}\left(\sum_{i \in C_0} y_i + \sum_{i \in C_1} y_i\right) - \frac{n_1}{n}(\bar{y}_1 - \bar{y}_0) \\
&= \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_1 - \frac{n_1}{n}(\bar{y}_1 - \bar{y}_0) \\
&= \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_0 \\
&= \bar{y}_0
\end{aligned} \tag{16}$$

となり，統制群の平均に等しい．

ここで，(17) に注目すると，(15) を用いれば

$$t = \frac{\hat{\beta}}{s^* \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} \tag{17}$$

である．つまり，検定統計量 t は β を $\frac{1}{s^* \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$ でスケールしたものである．さらにこのスケーリングのための定数は，その定義上，正である．よって，t 検定で t の大小（すなわち対応する p 値の大小）を論じることは，β，つまり群間平均値の差を論じることと同等である．

以上の議論をまとめると，以下のようになる．

- x_i が 0 と 1 の 2 値の時に回帰分析を実行すると，回帰係数の推定量 $\hat{\beta}$ は群間平均の差に等しくなる．
- t 検定で注目する検定統計量 t は，群間平均の差をスケーリングしたものである．
- $\hat{\beta}$ と t は同じものを評価していることになるため，回帰分析と t 検定は同等である．

6 分散分析と回帰分析の関係

分散分析の節で例として取り上げたように，要因が統制群・実験群 1・実験群 2 の 3 水準を取るケースを考える．この場合，独立変数として各個体がどの条件に割り当てられるかを使うことになるが，そのために，水準をコーディング（このことを，ダミー変数変換特別して，one-hot encoding と呼ぶ）した $n \times 3$ の行列 $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix}
1 & 0 & 0 \\
\vdots & \vdots & \vdots \\
1 & 0 & 0 \\
0 & 1 & 0 \\
\vdots & \vdots & \vdots \\
0 & 1 & 0 \\
0 & 0 & 1 \\
\vdots & \vdots & \vdots \\
0 & 0 & 1
\end{bmatrix} \tag{18}$$

を用いる．第 i 行目の第 j 列に 1 があることは，i 番目の個体が j 番目の条件に割り当てられていることを意味する．ただし，上式では各条件に所属する個体がまとまるように，個体を適宜並び替えている．(4) に上の $\mathbf{X}$ を代入すると

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \begin{bmatrix} \frac{1}{n_0} & & \\ & \frac{1}{n_1} & \\ & & \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{i\in C_0} y_0 \\ \sum_{i\in C_1} y_1 \\ \sum_{i\in C_2} y_2 \end{bmatrix} = \begin{bmatrix} \bar{y}_0 \\ \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} \end{aligned} \tag{19}$$

となり，回帰係数の推定量が群内平均と等しくなる．このことは水準数が 2 であっても一般性を損なうことなく成り立つ．

さて，この推定量を利用すると，予測値 $\mathbf{X}\hat{\boldsymbol{\theta}}$ はどのように計算されるであろうか．

$$\mathbf{X}\hat{\boldsymbol{\theta}} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \bar{y}_0 \\ \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} = \begin{bmatrix} \bar{y}_0 \\ \vdots \\ \bar{y}_1 \\ \vdots \\ \bar{y}_2 \\ \vdots \end{bmatrix} \tag{20}$$

のように，各個体が所属する群における従属変数の平均値（群内平均）が並ぶベクトルとなる．この予測値を用いると，予測の残差ベクトルを $\mathbf{e}$ とすれば，回帰分析モデルに従って

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\theta}} + \mathbf{e} \tag{21}$$

が得られる．この両辺から，従属変数の全体平均が並んだベクトル $\bar{\mathbf{y}} = \bar{y}\mathbf{1}_n$ を引くと

$$\begin{aligned}\mathbf{y} - \bar{\mathbf{y}} &= (\mathbf{X}\hat{\boldsymbol{\theta}} - \bar{\mathbf{y}}) + \mathbf{e} \\ \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ \\ \vdots \\ y_n - \bar{y} \end{bmatrix} &= \begin{bmatrix} \bar{y}_1 - \bar{y} \\ \vdots \\ \bar{y}_2 - \bar{y} \\ \vdots \end{bmatrix} + \begin{bmatrix} y_1 - \bar{y}_1 \\ \vdots \\ y_2 - \bar{y}_2 \\ \vdots \end{bmatrix} \end{aligned} \tag{22}$$

が得られる．$\mathbf{e}$ の各要素は両辺の帳尻が合うように調整した．この式の両辺の平方をとると，右辺のクロス・プロダクトが消失する．このことは，予測値 $\mathbf{X}\hat{\theta}$ と残差 $\mathbf{e}$ が直交することと，残差の和が 0 であるという回帰モデルにおける残差の性質に基づく．従って

$$\begin{aligned}||\mathbf{y} - \bar{\mathbf{y}}||^2 &= ||\mathbf{X}\hat{\boldsymbol{\theta}} - \bar{\mathbf{y}}||^2 + ||\mathbf{e}||^2 \\ \sum_i (y_i - \bar{y})^2 &= \sum_j n_j(\bar{y}_j - \bar{y})^2 + \sum_j \sum_{i\in C_j} (y_i - \bar{y}_j)^2 \end{aligned} \tag{23}$$

が成り立つ．左辺の第 1 項は分散分散における全体平方和 SS_T，右辺の第 1 項は群間平方和 SS_B，第 2 項は群内平方和 SS_W に等しく，これらに基づいて (10) の F 値を計算することができる．

回帰分析の結果の評価には，決定係数が用いられることが多い．決定係数 R は，y_i の予測値を $\hat{y}_i$ とした時

$$R = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \tag{24}$$

によって定義される．上で検討した，one-hot encoding による $\mathbf{X}$ を用いた回帰分析では，(20) に示した通り，$\hat{y}_i$ は i 番目の個体が割り当てられる条件の群内平均である．このことを利用すると，(24) は

$$R = \frac{\sum_j n_j(\bar{y}_j - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{SS_B}{SS_T} \tag{25}$$

と全体平方和と群間平方和の比として表現できる．この量は，全体平方和に占める群間平方和の割合を示しているが，これは本質的に，群内平方和に対する群間平方和の割合を評価する F 値と同じことを表す．

以上の結果をまとめると，以下のようになる．

- 個体の条件への割り当てをダミー変数とした回帰分析を実行すると，回帰係数の推定量として群内平均が得られる．
- この推定量を用いた回帰分析モデルの式を変形すると，分散分析における平方和の分解を導くことができ，それに基づいて分散分析の検定統計量 F を計算できる．
- 分散分析における F 値は，回帰分析における決定係数と本質的に同じものを表している．

7　分散分析と t 検定の関係

前節の分散分析で用いた one-hot encoding は，水準数が 2，つまり t 検定と同じ状況でも使える．よってここでは，3 節で用いたダミー変数変換ではなく，6 節で用いた one-hot encoding によって，個体が割り当てられる群を表現することを考えよう．つまり

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 0 & 1 \\ \vdots & \vdots \end{bmatrix} \tag{26}$$

である．この時，分散分析の時と同様に，得られる回帰係数の推定量は，群内平均が並んだものである．これに基づいて，平方和の分解を行い，平均平方和や F 値を計算すれば，水準数が 2 の場合であっても分散分析を行うことができる．実はこの F 値は，水準数が 2 の時に限って，t 検定で使われる検定統計量 t の 2 乗と等しくなる．よって分散分析と t 検定は同値であるということができる．このことは，以下のように示すことができる．

まず，この状況で t 検定を実行することを考える．t 値は，プールされた分散の平方が，その定義 (6) から

$$s^{*2} = \frac{n_0 s_0^2 + n_1 s_1^2}{n - 2} \tag{27}$$

であることを利用すると，t 値の平方は

$$\begin{aligned} t^2 &= \frac{(\bar{y}_1 - \bar{y}_0)^2}{S^{*2}\left(\dfrac{1}{n_0} + \dfrac{1}{n_1}\right)} \\ &= \frac{(\bar{y}_1 - \bar{y}_0)^2}{\dfrac{n_0 s_0^2 + n_1 s_1^2}{n-2} \times \dfrac{n}{n_0 n_1}} \\ &= \frac{n_0 n_1 (n-2)(\bar{y}_1 - \bar{y}_0)^2}{n(n_0 s_0^2 + n_1 s_1^2)} \end{aligned} \tag{28}$$

一方，上の (26) による回帰分析の結果を用いて分散分析を行うとき，群間の平均平方和 $S\bar{S}_B$ および群内平均平方和 $S\bar{S}_W$ はそれぞれ

$$\begin{aligned} S\bar{S}_B &= ||\mathbf{X}\hat{\boldsymbol{\theta}} - \bar{\mathbf{y}}||^2 \\ &= n_0(\bar{y}_0 - \bar{y})^2 + n_1(\bar{y}_1 - \bar{y})^2 \end{aligned} \tag{29}$$

および

$$\begin{aligned} S\bar{S}_W &= \frac{||\mathbf{e}||^2}{n-2} \\ &= \frac{1}{n-2}\left(\sum_{i\in C_0}(y_i - \bar{y}_0)^2 + \sum_{i\in C_1}(y_i - \bar{y}_1)^2\right) \\ &= \frac{n_0 s_0^2 + n_1 s_1^2}{n-2} \end{aligned} \tag{30}$$

となる．特に，$S\bar{S}_W$ は群内分散 s_0^2 および s_1^2 を条件に所属する個体数で重み付けた和として表現できる．これらを用いると，F 値は

$$F = \frac{S\bar{S}_B}{S\bar{S}_W} = \frac{(n-2)(n_0(\bar{y}_0 - \bar{y})^2 + n_1(\bar{y}_1 - \bar{y})^2)}{n_0 s_0^2 + n_1 s_1^2} \tag{31}$$

と計算できる．ここで，上式の分母について，$\bar{y} = n^{-1}(n_0\bar{y}_0 + n_1\bar{y}_1)$ を利用すれば

$$\begin{aligned} n_0(\bar{y}_0 - \bar{y})^2 + n_1(\bar{y}_1 - \bar{y})^2 &= n_0\bar{y}_0^2 + n_1\bar{y}_1^2 - 2(n_0\bar{y}_0 + n_1\bar{y}_1)\bar{y} + n\bar{y}^2 \\ &= n_0\bar{y}_0^2 + n_1\bar{y}_1^2 - 2\bar{y}^2 \\ &= n_0\bar{y}_0^2 + n_1\bar{y}_1^2 + \frac{n_0^2}{n}\bar{y}_0^2 + \frac{n_1^2}{n}\bar{y}_1^2 - \frac{2n_0n_1}{n}\bar{y}_0\bar{y}_1 \\ &= \left(n_0 - \frac{n_0^2}{n}\right)\bar{y}_0^2 + \left(n_1 - \frac{n_1^2}{n}\right)\bar{y}_1^2 - \frac{2n_0n_1}{n}\bar{y}_0\bar{y}_1 \\ &= \frac{n_0n_1}{n}\bar{y}_0^2 + \frac{n_0n_1}{n}\bar{y}_1^2 - \frac{2n_0n_1}{n}\bar{y}_0\bar{y}_1 \\ &= \frac{n_0n_1}{n}(\bar{y}_1 - \bar{y}_0)^2 \end{aligned} \tag{32}$$

と変形できる．この結果を用いると，(31) は，(28) と比較することにより

$$F = \frac{n-2}{n_0 s_0^2 + n_1 s_1^2} \times \frac{n_0 n_1}{n}(\bar{y}_1 - \bar{y}_0)^2 = t^2 \tag{33}$$

が得られ，F 値が t 値の二乗に等しいことが示された．このことは，2 群の平均値の差を t 値に変換して，その大小で平均値の差の統計的有意性について議論する t 検定の手続きは，F 値による議論，すなわち分散分析の手続きと同値であることを意味する．

8　全体のまとめ

全体をまとめると下の図のようになる．図中の節番号は対応する本文の節番号を表す．

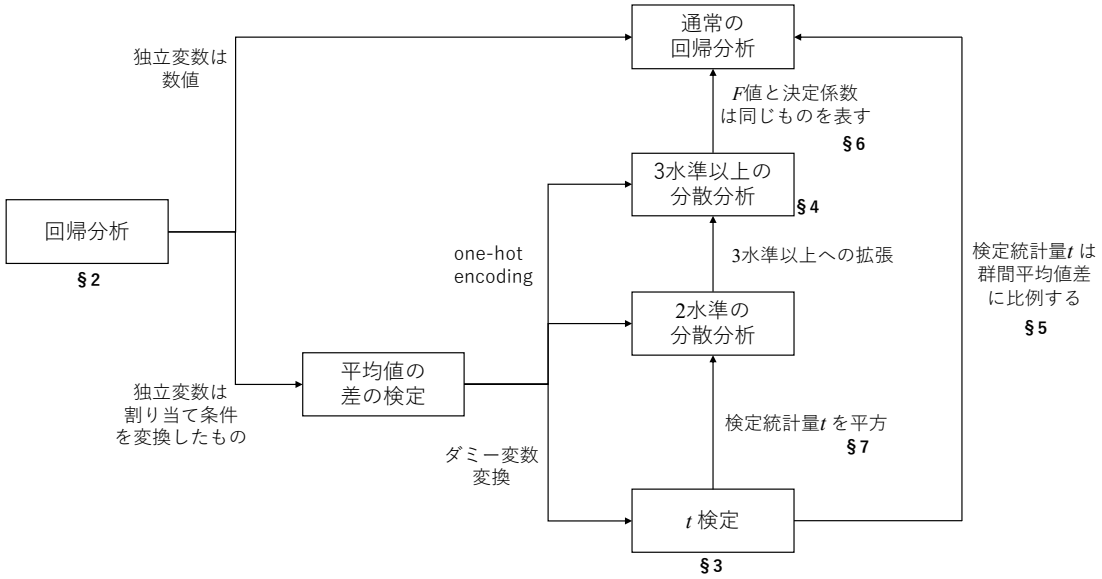

Figure 1: 回帰分析・t 検定・分散分析の関係