第 7 回　重回帰分析（6.1–6.3）

村澤 康友

2024 年 5 月 21 日

今日のポイント

1. 結果を処置ダミーに回帰すれば平均処置効果が求まる．結果と処置の両方に影響する共変量が存在する場合は重回帰分析で共変量調整を行う．
2. 自由度修正済み決定係数は $\bar{R}^2 := 1 - [\mathrm{RSS}/(n-k)]/[\mathrm{TSS}/(n-1)]$．2 つの説明変数の相関が極めて高く，それらの回帰係数の OLS 推定値が不安定になる問題を多重共線性という．
3. 説明変数の欠落によって生じる OLS 推定量の偏りを欠落変数バイアスという．
4. ある説明変数の偏回帰係数の OLS 推定値は，その説明変数を残りの説明変数に回帰した OLS 残差に被説明変数を単回帰しても求まる．
5. 誤差項が無相関で分散が均一な線形回帰モデルを古典的線形回帰モデルという．被説明変数の線形関数で表される推定量を線形推定量という．不偏な線形推定量を線形不偏推定量という．分散が最小となる線形不偏推定量を最良線形不偏推定量（BLUE）という．古典的線形回帰モデルの回帰係数の OLS 推定量は BLUE（ガウス＝マルコフ定理）．

目次

1　重回帰分析

1.1　ダミー変数（pp. 53, 166）

ある条件に該当するか否かの 2 値変数はベルヌーイ確率変数で表せる．すなわち

$$D := \begin{cases} 1 & \text{該当} \\ 0 & \text{非該当} \end{cases}$$

処置の有無を D，結果を Y とする．

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

定義 2. 処置群と対照群に対する効果の差を**処置**

（介入）**効果**という．

定義 3. 処置効果の平均を**平均処置効果**（*Average Treatment Effect, ATE*）という．

注 1. 処置群と対照群の母平均の差に等しい．すなわち Y に対する D の ATE は

$$\mathrm{ATE} = \mathrm{E}(Y|D=1) - \mathrm{E}(Y|D=0)$$

実験データなら母平均の差の推測のみ（2 標本問題）．

注 2. $\mu_0 := \mathrm{E}(Y|D=0)$，$\mu_1 := \mathrm{E}(Y|D=1)$ とすると

$$\begin{aligned}\mathrm{E}(Y|D) &= D\mu_1 + (1-D)\mu_0 \\ &= \mu_0 + (\mu_1 - \mu_0)D \\ &= \mu_0 + \mathrm{ATE} \cdot D\end{aligned}$$

これは単回帰モデル．したがって 2 標本問題は単回帰分析で実行できる．また k 標本問題は重回帰分析で実行できる（＝分散分析）．

1.2 共変量調整（p. 131）

実験データと異なり，観察データでは D を直接コントロールできない．Y と D の両方に影響する変数 X が存在する場合，Y の (D, X) 上への重回帰モデルを考える．

$$\mathrm{E}(Y|D,X) = \alpha + \mathrm{ATE} \cdot D + \beta X$$

定義 4. 関心の対象外の説明変数を**共変量**という．

定義 5. 分析の際に共変量の影響を調整することを**共変量調整**という．

1.3 MM（＝ OLS）推定量（p. 133）

次の重回帰モデルを考える．

$$\mathrm{E}(Y|X_1,\dots,X_k) = \alpha + \beta_1 X_1 + \dots + \beta_k X_k$$

回帰の誤差項は $U := Y - \mathrm{E}(Y|X_1,\dots,X_k)$.

定理 1.

$$\mathrm{E}(U|X_1,\dots,X_k) = 0$$

証明. 復習テスト． □

定理 2.

$$\mathrm{E}(U) = \mathrm{E}(X_1 U) = \dots = \mathrm{E}(X_k U) = 0$$

証明. 復習テスト． □

注 3. $U = Y - \alpha - \beta_1 X_1 - \dots - \beta_k X_k$ を代入すると

$$\begin{aligned}\mathrm{E}(Y - \alpha - \beta_1 X_1 - \dots - \beta_k X_k) &= 0 \\ \mathrm{E}(X_1(Y - \alpha - \beta_1 X_1 - \dots - \beta_k X_k)) &= 0 \\ &\vdots \\ \mathrm{E}(X_k(Y - \alpha - \beta_1 X_1 - \dots - \beta_k X_k)) &= 0\end{aligned}$$

この連立方程式が解けるなら，$(\alpha, \beta_1, \dots, \beta_k)$ は MM 法で推定できる（OLS と同値）．

1.4 自由度修正済み決定係数（p. 135）

決定係数は

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

ただし

$$\begin{aligned}\mathrm{TSS} &:= \sum_{i=1}^n (y_i - \bar{y})^2 \\ \mathrm{RSS} &:= \sum_{i=1}^n e_i^2\end{aligned}$$

推定する係数の数（＝定数項を含む説明変数の数）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 6. 自由度修正済み決定係数は

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

注 4. 無作為標本なら

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

$\mathrm{var}(u_i|x_i) = \mathrm{var}(u_i)$ なら

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

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

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

1.5 多重共線性（p. 138）

次の重回帰モデルを考える．

$$\mathrm{E}(Y|X,Z)=\alpha+\beta X+\gamma Z$$

ここで $X=Z$ とすると，任意の w について

$$\mathrm{E}(Y|X,Z)=\alpha+w(\beta+\gamma)X+(1-w)(\beta+\gamma)Z$$

すなわち X,Z の係数は一意に定まらない．$Z=a+bX$ でも同様.

より一般的に，次の重回帰モデルを考える．

$$\mathrm{E}(Y|X_1,\dots,X_k)=\alpha+\beta_1 X_1+\dots+\beta_k X_k$$

ここで $X_1=a+b_2X_2+\dots+b_kX_k$ の場合も係数は一意に定まらない.

定義 7. 実質的に同じ説明変数が 2 つあり，それらの回帰係数が定まらない問題を**完全な多重共線性**という．

定義 8. 2 つの説明変数の相関が極めて高く，それらの回帰係数の OLS 推定値が不安定になる問題を**（準）多重共線性**という．

2 欠落変数バイアス（p. 139）

次の重回帰モデルを考える．

$$\mathrm{E}(Y|X,Z)=\alpha+\beta X+\gamma Z$$

ここで $\mathrm{E}(Z|X)=a+bX$ とし，Z を説明変数に含めないと，繰り返し期待値の法則より

$$\begin{aligned}
\mathrm{E}(Y|X) &= \mathrm{E}(\mathrm{E}(Y|X,Z)|X) \\
&= \mathrm{E}(\alpha+\beta X+\gamma Z|X) \\
&= \alpha+\beta X+\gamma\,\mathrm{E}(Z|X) \\
&= \alpha+\beta X+\gamma(a+bX) \\
&= \alpha+\gamma a+(\beta+\gamma b)X
\end{aligned}$$

すなわち X の回帰係数は β でなく $\beta+\gamma b$ となる．

定義 9. 説明変数の欠落によって生じる OLS 推定量の偏りを**欠落変数バイアス**という．

3 偏回帰

3.1 重回帰モデル

$(1+k)$ 変量データを $\{(y_i,x_{i,1},\dots,x_{i,k})\}_{i=1}^n$ とする．y_i の $(x_{i,1},\dots,x_{i,k})$ 上への重回帰モデルは

$$\mathrm{E}(y_i|x_{i,1},\dots,x_{i,k})=\beta_1 x_{i,1}+\dots+\beta_k x_{i,k}$$

β_1 の推定を考える（$\beta_2,\dots,\beta_k$ に関心はない）．

3.2 MM（= OLS）推定量

繰り返し期待値の法則より

$$\begin{aligned}
\mathrm{E}(x_{i,1}(y_i-\beta_1x_{i,1}-\dots-\beta_kx_{i,k})) &= 0 \\
&\vdots \\
\mathrm{E}(x_{i,k}(y_i-\beta_1x_{i,1}-\dots-\beta_kx_{i,k})) &= 0
\end{aligned}$$

$(\beta_1,\dots,\beta_k)$ の MM（= OLS）推定量を $(b_1,\dots,b_k)$ とすると

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n x_{i,1}(y_i-b_1x_{i,1}-\dots-b_kx_{i,k}) &= 0 \\
&\vdots \\
\frac{1}{n}\sum_{i=1}^n x_{i,k}(y_i-b_1x_{i,1}-\dots-b_kx_{i,k}) &= 0
\end{aligned}$$

3.3 OLS 残差

y_i の回帰予測は

$$\hat{y}_i := b_1x_{i,1}+\dots+b_kx_{i,k}$$

OLS 残差は

$$\begin{aligned}
e_i &:= y_i-\hat{y}_i \\
&= y_i-b_1x_{i,1}-\dots-b_kx_{i,k}
\end{aligned}$$

定理 3.

$$\sum_{i=1}^n x_{i,1}e_i=\dots=\sum_{i=1}^n x_{i,k}e_i=0$$

証明. 復習テスト． □

系 1.

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

証明. 変形すると

$$\begin{aligned}\sum_{i=1}^{n} \hat{y}_i e_i &= \sum_{i=1}^{n} (b_1 x_{i,1} + \cdots + b_k x_{i,k}) e_i \\ &= b_1 \sum_{i=1}^{n} x_{i,1} e_i + \cdots + b_k \sum_{i=1}^{n} x_{i,k} e_i\end{aligned}$$

前定理より各項は 0. □

3.4 偏回帰（p. 158）

$x_{i,1}$ の $(x_{i,2}, \ldots, x_{i,k})$ 上への重回帰モデルを考える．すなわち

$$\mathrm{E}(x_{i,1} | x_{i,2}, \ldots, x_{i,k}) = \gamma_2 x_{i,2} + \cdots + \gamma_k x_{i,k}$$

繰り返し期待値の法則より

$$\begin{aligned}\mathrm{E}(x_{i,2}(x_{i,1} - \gamma_2 x_{i,2} - \cdots - \gamma_k x_{i,k})) &= 0 \\ &\vdots \\ \mathrm{E}(x_{i,k}(x_{i,1} - \gamma_2 x_{i,2} - \cdots - \gamma_k x_{i,k})) &= 0\end{aligned}$$

$(\gamma_2, \ldots, \gamma_k)$ の MM (= OLS) 推定量を $(c_2, \ldots, c_k)$ とすると

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^{n} x_{i,2}(x_{i,1} - c_2 x_{i,2} - \cdots - c_k x_{i,k}) &= 0 \\ &\vdots \\ \frac{1}{n} \sum_{i=1}^{n} x_{i,k}(x_{i,1} - c_2 x_{i,2} - \cdots - c_k x_{i,k}) &= 0\end{aligned}$$

$x_{i,1}$ の回帰予測は

$$\hat{x}_{i,1} := c_2 x_{i,2} + \cdots + c_k x_{i,k}$$

OLS 残差は

$$\begin{aligned}x_{i,1}^* &:= x_{i,1} - \hat{x}_{i,1} \\ &= x_{i,1} - c_2 x_{i,2} - \cdots - c_k x_{i,k}\end{aligned}$$

OLS 残差の性質より

$$\sum_{i=1}^{n} x_{i,2} x_{i,1}^* = \cdots = \sum_{i=1}^{n} x_{i,k} x_{i,1}^* = 0$$

かつ

$$\sum_{i=1}^{n} \hat{x}_{i,1} x_{i,1}^* = 0$$

補題 1.

$$\sum_{i=1}^{n} \hat{x}_{i,1} e_i = 0$$

証明. 変形すると

$$\begin{aligned}\sum_{i=1}^{n} \hat{x}_{i,1} e_i &= \sum_{i=1}^{n} (c_2 x_{i,2} + \cdots + c_k x_{i,k}) e_i \\ &= c_2 \sum_{i=1}^{n} x_{i,2} e_i + \cdots + c_k \sum_{i=1}^{n} x_{i,k} e_i\end{aligned}$$

前定理より各項は 0. □

定理 4 (偏回帰)**.**

$$b_1 = \frac{\sum_{i=1}^{n} x_{i,1}^* y_i}{\sum_{i=1}^{n} {x_{i,1}^*}^2}$$

証明. 補題より

$$\begin{aligned}\sum_{i=1}^{n} x_{i,1} e_i &= \sum_{i=1}^{n} \left(\hat{x}_{i,1} + x_{i,1}^*\right) e_i \\ &= \sum_{i=1}^{n} \hat{x}_{i,1} e_i + \sum_{i=1}^{n} x_{i,1}^* e_i \\ &= \sum_{i=1}^{n} x_{i,1}^* e_i \\ &= \sum_{i=1}^{n} x_{i,1}^* (y_i - b_1 x_{i,1} - \cdots - b_k x_{i,k}) \\ &= \sum_{i=1}^{n} x_{i,1}^* y_i - b_1 \sum_{i=1}^{n} x_{i,1}^* x_{i,1} \\ &\quad - b_2 \sum_{i=1}^{n} x_{i,1}^* x_{i,2} - \cdots - b_k \sum_{i=1}^{n} x_{i,1}^* x_{i,k} \\ &= \sum_{i=1}^{n} x_{i,1}^* y_i - b_1 \sum_{i=1}^{n} x_{i,1}^* x_{i,1} \\ &= \sum_{i=1}^{n} x_{i,1}^* y_i - b_1 \sum_{i=1}^{n} x_{i,1}^* \left(\hat{x}_{i,1} + x_{i,1}^*\right) \\ &= \sum_{i=1}^{n} x_{i,1}^* y_i - b_1 \sum_{i=1}^{n} x_{i,1}^* \hat{x}_{i,1} - b_1 \sum_{i=1}^{n} {x_{i,1}^*}^2 \\ &= \sum_{i=1}^{n} x_{i,1}^* y_i - b_1 \sum_{i=1}^{n} {x_{i,1}^*}^2\end{aligned}$$

左辺= 0 より b_1 について解けば結果が得られる. □

注 5. 定理より β_1 の OLS 推定量 b_1 は以下の手順でも求まる.

1. $x_{i,1}$ を $(x_{i,2},\dots,x_{i,k})$ 上へ回帰し，OLS 残差 $x^*_{i,1}$ を求める．
2. y_i を $x^*_{i,1}$ 上へ回帰．

したがって b_1 は，$(x_{i,2},\dots,x_{i,k})$ と相関する部分を取り除いた上での y_i と $x_{i,1}$ の関係を表す．

4 OLS 推定量の性質

4.1 古典的線形回帰モデル（p. 146）

$(1+k)$ 変量データを $((y_1,\boldsymbol{x}_1),\dots,(y_n,\boldsymbol{x}_n))$ とする．ただし $\boldsymbol{x}_i := (x_{i,1},\dots,x_{i,k})'$．$x_{i,1} := 1$ を定数項とすると，y_i の $\boldsymbol{x}_i$ 上への重回帰モデルは

$$\begin{aligned}\mathrm{E}(y_i|\boldsymbol{x}_i) &= \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} \\ &= \boldsymbol{\beta}'\boldsymbol{x}_i \\ &= \boldsymbol{x}_i'\boldsymbol{\beta}\end{aligned}$$

または

$$\begin{aligned}y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \\ \mathrm{E}(u_i|\boldsymbol{x}_i) &= 0\end{aligned}$$

すなわち重回帰モデルをベクトルで表記すれば，定数項なしの単回帰モデルと同様に扱える．

定義 10. $(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$ を所与として $u_1,\dots,u_n$ が無相関で分散が均一な線形回帰モデルを**古典的線形回帰モデル**という．

注 6. すなわち

$$\begin{aligned}y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \\ \mathrm{E}(u_i|\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) &= 0 \\ \mathrm{var}(u_i|\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) &= \sigma^2 \\ \mathrm{cov}(u_i,u_j|\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) &= 0 \quad \text{for } i \neq j\end{aligned}$$

4.2 MM（= OLS）推定量

繰り返し期待値の法則より

$$\mathrm{E}(\boldsymbol{x}_i u_i) = \boldsymbol{0}$$

$u_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$ を代入すると

$$\mathrm{E}(\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})) = \boldsymbol{0}$$

$\boldsymbol{\beta}$ の MM（= OLS）推定量を $\boldsymbol{b}$ とすると

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{b}) = \boldsymbol{0}$$

すなわち

$$\sum_{i=1}^n \boldsymbol{x}_i y_i = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{b}$$

逆行列を用いて連立方程式を解くと

$$\boldsymbol{b} = \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\sum_{i=1}^n \boldsymbol{x}_i y_i$$

定理 5.

$$\mathrm{E}(\boldsymbol{b}|\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) = \boldsymbol{\beta}$$

証明. 省略（定数項のない単回帰モデルと同じ）． □

系 2.

$$\mathrm{E}(\boldsymbol{b}) = \boldsymbol{\beta}$$

証明. 省略（繰り返し期待値の法則）． □

定理 6. 古典的線形回帰モデルなら

$$\mathrm{var}(\boldsymbol{b}|\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) = \sigma^2\left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}$$

証明. 省略（定数項のない単回帰モデルと同じ）． □

4.3 ガウス＝マルコフ定理（p. 146）

定義 11. 被説明変数の線形関数で表される推定量を**線形推定量**という．

注 7. $\boldsymbol{b}$ は $y_1,\dots,y_n$ の線形関数だから線形推定量．

定義 12. 不偏な線形推定量を**線形不偏推定量**という．

注 8. $\mathrm{E}(\boldsymbol{b}) = \boldsymbol{\beta}$ より $\boldsymbol{b}$ は線形不偏推定量．

定義 13. 分散が最小となる線形不偏推定量を**最良線形不偏推定量**（*Best Linear Unbiased Estimator, BLUE*）という．

定理 7 (ガウス＝マルコフ定理)**.** 古典的線形回帰モデルの回帰係数の *OLS* 推定量は *BLUE*.

証明. 省略（行列を使うと簡単）． □

5　今日のキーワード

ダミー変数，処置（介入）効果，平均処置効果（ATE），共変量，共変量調整，自由度修正済み決定係数，完全な多重共線性，（準）多重共線性，欠落変数バイアス，偏回帰，古典的線形回帰モデル，線形推定量，線形不偏推定量，最良線形不偏推定量（BLUE），ガウス＝マルコフ定理

6　次回までの準備

復習　教科書第 6 章 1–3 節，復習テスト 7

予習　教科書第 6 章 4–5 節