

Ⅱ-4. 欠落変数バイアス

2025 年 1 月 19 日

M. Fujii

I. 前提となる状況

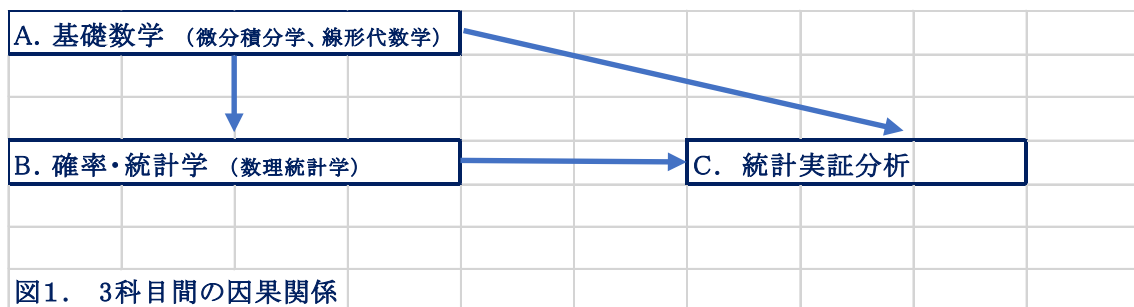
ある理学工業大学の社会工学研究科では、

入学 1 年目の講義で A. 基礎数学（微分積分学、線形代数学）、

入学 2 年目の講義で B. 確率・統計学（数理統計学）、

入学 3 年目の講義で C. 統計実証分析 を行なっています。

いずれの科目も成績評価は、期末試験だけです。これら 3 科目の得点の間に、図 1. のような因果関係を想定しています。



(*) 矢印は直接の因果関係を表します。

X_1 は基礎数学の得点、 X_2 は確率・統計学の得点、 X_3 は統計実証分析の得点です。

図1. の因果関係を線形構造方程式で表したものを、モデル [1] とします。

モデル [1]

$$X_2 = \beta_{12} * X_1 + \varepsilon_2 \quad \text{--- 式 (1-1)}$$

$$X_3 = \beta_{13} * X_1 + \beta_{23} * X_2 + \varepsilon_3 \quad \text{--- 式 (1-2)}$$

X_1 、 X_2 、 X_3 は平均 0、分散 1 に標準化されていることとします。

ε_2 と ε_3 は期待値が 0 で、互いに独立な誤差項とします。

ε_2 は X_1 と無相関です。 ε_3 は X_1 、 X_2 と無相関です。

X_1 、 X_2 、 X_3 の母相関行列は、 $\begin{pmatrix} 1.0 & 0.8 & 0.6 \\ - & 1.0 & 0.7 \\ - & - & 1.0 \end{pmatrix}$ です。

標準化されているので、 $\rho_{ij} = \text{Cov}(X_i, X_j)$ です。

ある年、 X_1 である基礎数学の得点データを入手し忘れました。そこで、 X_2 である確率・統計学の得点、 X_3 である統計実証分析の得点だけを用いて、モデル [2] を最小二乗法で推定しました。 ε'_3 は誤差項です。

モデル [2]

$$X_3 = \gamma_{23} * X_2 + \varepsilon'_3 \quad \text{--- 式 (2)}$$

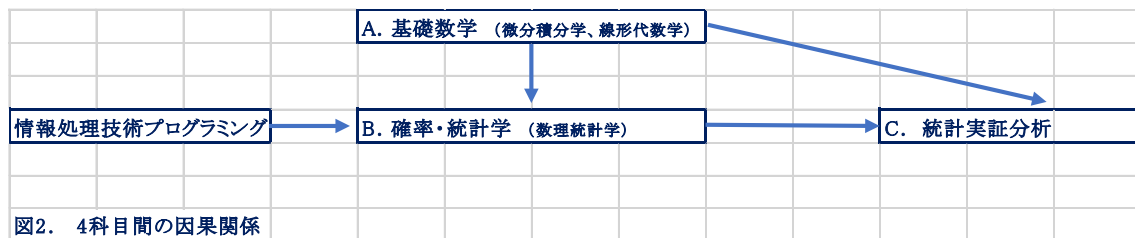
モデル [2] の学年では、情報処理技術プログラミングの講義の履修有無に関するデータ Z が入手可能でした。

$Z = :$ 1 = 情報処理技術プログラミングの講義を履修済

0 = 情報処理技術プログラミングの講義を履修せず

情報処理技術プログラミングの講義の履修有無を含めた因果関係は、

図 2. です。



この場合、基礎数学（微分積分学、線形代数学）は潜在変数となります。また、 Z は X_1 、 ε_3 とも無相関です。

確率・統計学では、情報処理技術プログラミング履修生の平均点が 0.4、未履修生の平均点が -0.8 です。

統計実証分析では、情報処理技術プログラミング履修生の平均点が 0.2、未履修生の平均点が -0.4 です。

II. 考察

<考察 1> β_{12} 、 β_{13} 、 β_{23} の推定値

式 [1-1] の両辺に X_1 を、式 [1-2] の両辺に X_2 を掛けると

$$X_1 X_2 = \beta_{12} * X_1^2 + X_1 * \varepsilon_2$$

$$X_1 X_3 = \beta_{13} * X_1^2 + \beta_{23} * X_1 X_2 + X_1 * \varepsilon_3$$

$$X_2 X_3 = \beta_{13} * X_1 X_2 + \beta_{23} * X_2^2 + X_2 * \varepsilon_3$$

両辺の期待値を取って、 β_{ij} を計算します。

$\rho_{ij} = \text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i) * E(X_j)$ かつ $E(X_i) = 0$ です。

よって、 $\text{Cov}(X_i, X_j) = E(X_i X_j)$ です。

$$\beta_{12} = 0.8 \quad \text{----- 式 (3)}$$

$$\begin{aligned} \because E(X_1 X_2) &= \text{Cov}(X_1, X_2) \\ &= E(\beta_{12} * X_1^2 + X_1 * \varepsilon_2) \\ &= \beta_{12} * E(X_1^2) + E(X_1) * E(\varepsilon_2) \\ &= \beta_{12} * \{V(X_1) + E(X_1)^2\} + E(X_1) * E(\varepsilon_2) \\ &= \beta_{12} * \{1 + 0^2\} + E(X_1) * 0 \\ &= \beta_{12} \end{aligned}$$

$$\beta_{13} + 0.8 \beta_{23} = 0.6 \quad \text{----- 式 (4)}$$

$$\because E(X_1 X_3) = \text{Cov}(X_1, X_3)$$

$$\begin{aligned}
&= E(\beta_{13} * X_1^2 + \beta_{23} * X_1 X_2 + X_1 * \varepsilon_3) \\
&= \beta_{13} * E(X_1^2) + \beta_{23} * E(X_1 X_2) + E(X_1 * \varepsilon_2) \\
&= \beta_{13} * \{V(X_1) + E(X_1)^2\} + \beta_{23} * \{Cov(X_1 X_2) \\
&\quad + E(X_1) * E(X_2)\} + E(X_1) * E(\varepsilon_2) \\
&= \beta_{13} + 0.8 \beta_{23} = 0.6
\end{aligned}$$

$$0.8 \beta_{13} + \beta_{23} = 0.7 \quad \text{----- 式 (5)}$$

$$\begin{aligned}
\because E(X_2 X_3) &= E(\beta_{13} * X_1 X_2 + \beta_{23} * X_2^2 + X_2 * \varepsilon_3) \\
&= \beta_{13} * E(X_1 X_3) + \beta_{23} * E(X_2)^2 + E(X_2 * \varepsilon_3) \\
&= \beta_{13} * \{Cov(X_1 X_3) + E(X_1) * E(X_3)\} \\
&\quad + \beta_{23} * \{V(X_2) + E(X_2)^2\} + E(X_1) * E(\varepsilon_2) \\
&= 0.8 \beta_{13} + \beta_{23} = 0.7
\end{aligned}$$

これらの連立方程式 (1), (2), (3) を解くと

$$\beta_{12} = 0.8, \quad \beta_{13} = \frac{2}{18}, \quad \beta_{23} = \frac{11}{18}$$

<考察 2> 影響を除いたあとの編相関係数

基礎数学の得点の影響を除いたあとの確率・統計学と統計実証分

析の得点の間の編相関係数

$$\begin{aligned}
 \gamma_{23|1} &= \frac{\gamma_{23} - \gamma_{12} \cdot \gamma_{13}}{\sqrt{1-(\gamma_{12})^2} \cdot \sqrt{1-(\gamma_{13})^2}} \text{----- 式 (6)} \\
 &= \frac{0.7 - 0.8 \cdot 0.6}{\sqrt{1-(0.8)^2} \cdot \sqrt{1-(0.6)^2}} \\
 &\doteq 0.4583
 \end{aligned}$$

<考察 3> 欠落変数バイアス

モデル [1] が正しいと仮定できる場合、 γ_{23} の最小二乗推定値は β_{23} の推定値として適切か否かの分析

⇒ このモデルの OLSE は、 $E(X_2 X_3) = 0.8 \beta_{13} + \beta_{23}$

の一致推定量になるので、 $0.8 \beta_{13}$ だけ欠落変数バイアスを含むので、不適切です。

⇒ β_{23} 、 γ_{23} とも「 X_2 が 1 だけ変化するときの X_3 の変化量」なので、意味的に近いと考えるところですが、そうではありません。モデル [1] をベースに考えると、

$$\text{Cov}(X_2, X_3) = \gamma_{23} = \beta_{23} + \beta_{13} \beta_{12} \text{----- 式 (7)}$$

でした。

したがって、 X_1 、 X_2 、 X_3 の標本から最小二乗法で推定値を求め

ると、 γ_{23} の推定値は β_{23} の推定値より $\beta_{13} \beta_{12}$ だけ大きくなり

ます。

X_3 を表す式での X_2 の係数を統計データから推定したい場合、本来モデルに含めないといけない変数 X_1 が欠けたモデル [2] で推定した値で モデル [1] の値を代用してしまうと、 $\beta_{13} \beta_{12}$ だけ過大評価となります。 欠落変数バイアスです。

<考察 4> β_{12} の一致推定値

モデル [1] は次のように修正されます。

モデル [3] とし、構造方程式は以下ようになります。

モデル [3]

$$X_2 = \beta_{12} * X_1 + \alpha Z + \varepsilon_2' \quad \text{--- 式 (8-1)}$$

$$X_3 = \beta_{13} * X_1 + \beta_{23} * X_2 + \varepsilon_3 \quad \text{--- 式 (8-2)}$$

ここで、 ε_2' は期待値 0 で、 X_1 や Z と無相関です。

ε_3 は期待値 0 で、 X_1 や X_2 や Z と無相関です。

<2値変数の場合の定理>

B が 2 値変数の場合、次が成り立つ。

$$\text{Cov}(A, B) = 0 \quad \Leftrightarrow \quad E(A \mid B) = E(A)$$

Z は 0, 1 の 2 値なので、 Z と X_1 、 ε_3 が無相関であることから、

$$E(X_1 \mid Z) = E(X_1) = 0, \quad E(\varepsilon_3 \mid Z) = E(\varepsilon_3) = 0$$

与えられた平均点は、 $Z = 1$ や $Z = 0$ で条件付けした期待値なので

$$E(X_2 \mid Z = 1) = 0.4, \quad E(X_2 \mid Z = 0) = -0.8$$

$$E(X_3 \mid Z = 1) = 0.2, \quad E(X_3 \mid Z = 0) = -0.4$$

両辺の $Z = 1$ 、 $Z = 0$ のもとでの条件期待値を取ると、

$$\begin{aligned} E(X_3 \mid Z = 1) = 0.2 &= \beta_{13} * E(X_1 \mid Z = 1) \\ &\quad + \beta_{23} * E(X_2 \mid Z = 1) \\ &\quad + E(\varepsilon_3 \mid Z = 1) \end{aligned}$$

$$\begin{aligned} E(X_3 \mid Z = 0) = -0.4 &= \beta_{13} * E(X_1 \mid Z = 0) \\ &\quad + \beta_{23} * E(X_2 \mid Z = 0) \\ &\quad + E(\varepsilon_3 \mid Z = 0) \end{aligned}$$

$$\therefore \beta_{23} = 0.5$$

以上