第 6 章 多変数の線形回帰分析

James H. Stock 著・Mark W. Watson 著・宮尾 龍蔵 訳

『入門計量経済学』（共立出版、2016）
https://www.kyoritsu-pub.co.jp/book/b10003746.html

2025-10-15

1. 除外された変数のバイアス

はじめに

- 分析から除外されている要因があると、最小二乗法 (Ordinary Least Squares, OLS) 推定量にバイアスが発生する。
- この問題を**除外された変数のバイアス (omitted variable bias)** と呼び、対処法として多変数回帰分析がある。
- 多変数回帰の基本的な考え方は、除外された変数を説明変数として回帰モデルに追加し、ある説明変数の効果を他の変数を一定の下で推計することである。

除外された変数のバイアスの定義

OLS 推定量に除外された変数のバイアスが生じるのは、以下の 2 つの条件が同時に成り立つ場合である。

1. **除外された変数が、すでに含まれている説明変数と相関がある。**
2. **除外された変数が、被説明変数の決定要因である。**

具体例（テスト成績と生徒・教師比率）

- **例 1：英語学習者の割合**
  - 生徒・教師比率と英語学習者の割合は相関があり（条件 1）、英語学習者の割合はテスト成績の決定要因であると考えられる（条件 2）ため、この変数を分析に含めないと**バイアスが発生する可能性がある。**
- **例 2：テストの実施時刻**
  - テスト時刻は成績に影響するかもしれないが（条件 2）、クラス規模とは相関しないと考えられる（条件 1 を満たさない）ため、**バイアスは発生しない。**
- **例 3：生徒 1 人当たりの駐車場スペースの数**
  - 駐車スペースの数はクラス規模と相関するかもしれないが（条件 1）、テスト成績に直接影響するとは考えにくいため（条件 2 を満たさない）、**バイアスは発生しない。**

バイアスの公式

標本数が大きくなるとき、OLS 推定量 $\hat{\beta}_1$ は以下のように収束する。

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \left(\frac{\sigma_u}{\sigma_X} \right)$$

この式は、以下の点を示唆している。

1. 除外された変数のバイアスは、標本数の大小にかかわらず問題である。
2. バイアスの大きさは、説明変数と誤差項の相関 ρ_{Xu} に依存する。
3. バイアスの方向は、ρ_{Xu} の符号に依存する。

2. 多変数回帰モデル

多変数回帰モデル (multiple regression model)

- 1 変数回帰モデルを拡張し、複数の説明変数を追加したモデルである。
- ある変数 (X_{1i}) の被説明変数 (Y_i) への影響を、他の説明変数 $(X_{2i}, X_{3i}$ など$)$ を**一定とした下で (holding constant)** 推定できる。

母集団の回帰線

2 つの説明変数を持つ多変数回帰モデルの母集団の回帰線 (population regression line) は次式で表される。

$$E(Y_i|X_{1i}=x_1, X_{2i}=x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- β_1 は、X_{2i} を一定にした下で、X_{1i} が 1 単位変化したときの Y_i への**部分的な効果 (partial effect)** を表す。
- β_0 は切片 (intercept) であり、すべての説明変数がゼロのときの Y_i の期待値である。

母集団の多変数回帰モデル

誤差項 u_i を含んだ母集団の多変数回帰モデル (population multiple regression model) は以下のように表される。

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots, n$$

- ここで、k は説明変数の数である。
- **均一分散 (homoscedastic)** と**不均一分散 (heteroscedastic)** の定義は 1 変数モデルと同様である。

3. 多変数回帰モデルにおける OLS 推定量

OLS 推定量

多変数回帰モデルの係数 $\beta_0, \beta_1, \dots, \beta_k$ は、予測の誤りの二乗和を最小化することで推定される。

$$\min_{b_0, b_1, \dots, b_k} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2$$

- この最小化によって得られる推定量を $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ とする。
- OLS 回帰線 (OLS regression line): $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$。
- 予測値 (predicted value): $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$。
- OLS 残差 (OLS residual): $\hat{u}_i = Y_i - \hat{Y}_i$。

応用例：テスト成績

- 1 変数モデル: $\widehat{TestScore} = 698.9 - 2.28 \times STR$。
- 多変数モデル（英語学習者の割合 $PctEL$ を追加）：

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL$$

- 英語学習者の割合をコントロールすることで、生徒・教師比率 (STR) の効果の推定値は約半分になった。これは 1 変数モデルに除外された変数のバイアスが存在したことを示唆する。

4. 多変数回帰の当てはまりの指標

1. 回帰の標準誤差 (Standard Error of the Regression, SER)

誤差項 u_i の標準偏差の推定量である。

$$SER = s_{\hat{u}} = \sqrt{\frac{SSR}{n-k-1}}$$

- ここで $SSR = \sum_{i=1}^{n} \hat{u}_i^2$（残差の二乗和)、$k$ は説明変数の数である。
- 分母が $n-k-1$ となるのは、推定される係数の数（$k+1$ 個）に応じた**自由度の修正 (degrees of freedom adjustment)** である。

2. R^2 (決定係数)

Y_i の標本分散のうち、説明変数によって説明される割合を示す。

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$（説明された二乗和）、$TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$（全体の二乗和）。
- **説明変数を追加すると、R^2 は決して減少しない**（通常は増加する）。

3. 修正済み R^2 (Adjusted R^2)

R^2 が変数の追加によって増加する傾向を修正した指標である。

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

- 説明変数を追加しても、必ずしも増加するわけではない。
- $\bar{R}^2$ は R^2 よりも常に小さい。
- **負の値になることもある。**

5. 多変数回帰モデルにおける最小二乗法の仮定

仮定１：条件付き期待値がゼロ

説明変数 $X_{1i}, \dots, X_{ki}$ が与えられた下で、誤差項 u_i の条件付き期待値はゼロである。

$$E(u_i | X_{1i}, X_{2i}, \dots, X_{ki}) = 0$$

- この仮定は、OLS 推定量が不偏であることを示すための鍵となる。

仮定 2：i.i.d.

$(X_{1i}, \dots, X_{ki}, Y_i)$ は、独立かつ同一の分布 (independently and identically distributed, i.i.d.) に従う。

仮定 3：大きな異常値はほとんど起こりえない

説明変数と被説明変数は、ゼロでない有限の 4 次のモーメントを持つ。

仮定 4：完全な多重共線性はない

完全な多重共線性 (perfect multicollinearity) とは、ある説明変数が他の説明変数の完全な線形関数で表される状況である。

- 完全な多重共線性があると、OLS 推定量を計算することはできない。

6. 多変数回帰モデルにおける OLS 推定量の分布

大標本における性質

- 基本概念 6.4 の 4 つの最小二乗法の仮定の下で、多変数回帰モデルの OLS 推定量 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ は、対応する母集団パラメータ $\beta_0, \beta_1, \ldots, \beta_k$ の**不偏かつ一致推定量**である。
- 標本数が十分大きい場合、OLS 推定量の結合標本分布は、**多変数正規分布**で近似される。
- OLS 推定量間には一般に相関があり、その相関は説明変数間の相関に依存する。

7. 多重共線性

完全な多重共線性 (Perfect Multicollinearity)

ある説明変数が他の説明変数の完全な線形結合で表される場合に発生する。

- **ダミー変数のわな (dummy variable trap)**:
  - G 個のカテゴリがあり、それぞれに対応する G 個のダミー変数をすべて説明変数に含め、かつ定数項（切片）も回帰式に含めると発生する。
  - 回避策：ダミー変数のうち 1 つを除くか、定数項を除く。
- **解決策**: 完全な多重共線性は通常、回帰モデルの特定化の誤りが原因であるため、説明変数を修正して問題を除去する必要がある。

不完全な多重共線性 (Imperfect Multicollinearity)

2 つ以上の説明変数が高い相関を示す状況を指す。

- **OLS 推定量を計算することは可能**であり、理論上の問題はない。
- しかし、高い相関を持つ変数に関する係数の**分散が大きくなり、推定が不正確になる**可能性がある。

8. 結論

- 説明変数が 1 つの回帰分析では、**除外された変数のバイアス**が常に懸念される。
- **多変数回帰分析**は、除外された変数をモデルに含めることで、このバイアスを軽減できる。
- 多変数回帰における係数は、他の説明変数を一定とした下での**部分的な効果**を表す。
- 多変数回帰の統計理論は 1 変数回帰の理論を基礎としており、**完全な多重共線性の排除**という 4 つ目の仮定が追加される。