A Note on Ommited Variable Bias

Hiroshi Hamada
Tohoku University

2016/02/07

1 問題の所在

統計モデルに本来含まれるべきはずの変数が含まれないとき，OLS 推定や最尤推定による母数の推定値にバイアスが生じることはよく知られている[1].

また, この欠落変数バイアスは特定条件下では，明示的（代数的）に説明変数の関数として表すことができる (Theil 1957; Griliches 1957).

1.1 問題のある論法?

社会学者 (と一部の経済学者) が好む論法として，過小定式化モデルに順次変数を追加した分析を比較して,「追加された説明変数 Z によって，初期段階で投入された変数 X の効果が弱まった（ゆえに Z は Y の変動を説明する上で重要な変数だ)」というものがある.

```
      M1          M2          M3
-----------------------------------------
X_1  5.12***    4.85**      3.12
X_2  3.48***    2.58*       1.78
X_3             1.67***     0.86
Z                           1.48***
-----------------------------------------
*$TeX$の表組みが面倒なので verbatim を使った
```

必要条件という意味では誤りではないが，冗長ではないだろうか？ そもそも X との共分散がない Z を追加したところで意味はないし，最初から過小定式化しなければよいだけの話ではないだろうか[2].

[1] Griliches (1957) によれば，欠落変数バイアスは遅くとも Theil の謄写版論文によって定式化されている (Theil 1956)．Theil の論文は後に *Review of the International Statistical Institute*, 25: 41-51. として 1957 年に刊行された. なお Theil は 1953 年に two-stage least squares を考案している.

[2] 追加した変数に星がつくからといって，因果的な効果があるわけではないので，そもそもよく分からないロジックである．必要条件と十分条件を取り違えている？

むしろ「欠落変数によって生じたバイアスが Z の追加によって補正された（ただし Z の係数自体にまだバイアスが残っている可能性がある）」と言うべきであろう．

以上の違和感から，バイアスについての数学的な理解を深めようと思い立った．以下に簡単な計算から分かった事を示す[3]．

1.2　過小定式化が単回帰の場合

ベースラインとして, Theil が示した《「過小定式化」による推定量のバイアス》の具体例を紹介しておく．

いま真のモデルが y の m 個の説明変数に対する回帰式

$$\begin{aligned} y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + u_i \\ i &= 1, 2, \ldots, n \end{aligned}$$

であると仮定する[4]．このとき本来必要な $x_2, x_3, ..., x_m$ が欠落した過小定式化

$$\begin{aligned} y_i &= \tilde{\beta}_1 x_{i1} + v_i \\ i &= 1, 2, \ldots, n \end{aligned}$$

を用いてパラメータを推定した仮定する．過小定式化の最小自乗推定量を b_1 とおく．生じたバイアス量 $E(b_1) - \beta_1$ は,

$$\begin{aligned} E(b_1) - \beta_1 &= \beta_2 \frac{\sum_{i=1}^n x_{i1}x_{i2}}{\sum_{i=1}^n x_{i1}^2} + \cdots + \beta_m \frac{\sum_{i=1}^n x_{i1}x_{im}}{\sum_{i=1}^n x_{i1}^2} \\ &= \sum_{j=2}^{m} \beta_j \frac{\sum_{i=1}^n x_{i1}x_{ij}}{\sum_{i=1}^n x_{i1}^2} \end{aligned}$$

である．

証明として次のような計算を考える．

$$\begin{aligned} E(b_1) &= E(\frac{\sum_{i=1}^n x_{i1}y_{i1}}{\sum_{i=1}^n x_{i1}^2}) \\ &= E(\frac{\sum_{i=1}^n x_{i1}(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + u_i)}{\sum_{i=1}^n x_{i1}^2}) \quad \text{本来の正しい } y_{i1} \text{に置き換える} \\ &= \beta_1 E(\frac{\sum_{i=1}^n x_{i1}^2}{\sum_{i=1}^n x_{i1}^2}) + \beta_2 E(\frac{\sum_{i=1}^n x_{i1}x_{i2}}{\sum_{i=1}^n x_{i1}^2}) + \cdots + \beta_m E(\frac{\sum_{i=1}^n x_{i1}x_{im}}{\sum_{i=1}^n x_{i1}^2}) + E(\frac{\sum_{i=1}^n x_{i1}u_i}{\sum_{i=1}^n x_{i1}^2}) \\ &= \beta_1 + \beta_2 \frac{\sum_{i=1}^n x_{i1}x_{i2}}{\sum_{i=1}^n x_{i1}^2} + \cdots + \beta_m \frac{\sum_{i=1}^n x_{i1}x_{im}}{\sum_{i=1}^n x_{i1}^2} + \frac{\sum_{i=1}^n x_{i1}}{\sum_{i=1}^n x_{i1}^2} E(\sum_{i=1}^n u_i) \\ &= \beta_1 + \beta_2 \frac{\sum_{i=1}^n x_{i1}x_{i2}}{\sum_{i=1}^n x_{i1}^2} + \cdots + \beta_m \frac{\sum_{i=1}^n x_{i1}x_{im}}{\sum_{i=1}^n x_{i1}^2} + 0 \end{aligned}$$

このことから分かるように，バイアスの正負は x_1 と $x_j (j = 2, 3, ..., m)$ の相関と，$\beta_j (j = 1, 2, ..., m)$ の符号に依存することがわかる．

[3] 操作変数法を含め，バイアスを取り除く方法については別に論じる．

[4] ここでは定数項のないモデルを考えているが，あってもなくても結論は大きくは変わらない．定数項を含む一般型については次節で述べる．

2　欠落変数バイアスの数学的構造

本来必要な説明変数 $x_j(j=1,2,3,...,m)$ のうち，ごっそり $x_j(j=2,3,...,m)$ が欠如しているという例は稀で，実際の研究の文脈では，必要な変数のうち少数が欠落してるパタンが多いと予想される．

過小定式化したモデルが単回帰ではなく重回帰になっている場合は，線形代数の記法を用いればバイアスを明示的に示すことができる[5]．

例 1. いま説明変数が 5 個ある $(k=5)$ と仮定して，そのうちの 2 個が欠落した場合の推定を考える．まず全ての説明変数を含むモデル（以下単にフルモデルと呼ぶ）を

$$\boldsymbol{y} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{e}$$

とおく．ここで $\boldsymbol{X_1}, \boldsymbol{\beta_1}, \boldsymbol{X_2}, \boldsymbol{\beta_2}$ は具体的には,

$$\boldsymbol{X_1} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix}, \quad \boldsymbol{\beta_1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad \boldsymbol{X_2} = \begin{pmatrix} x_{14} & x_{15} \\ x_{24} & x_{25} \\ \vdots & \vdots \\ x_{n4} & x_{n5} \end{pmatrix}, \quad \boldsymbol{\beta_2} = \begin{pmatrix} \beta_4 \\ \beta_5 \end{pmatrix}$$

である．

ε_i は個体 i の誤差であり，回帰モデルの**仮定**は以下の通りである (古典的回帰モデル).

- 説明変数 x_i は**確率変数ではなく**，固定された値をとる．
- 誤差項 e_i(error term) は確率変数で期待値は 0 である．$E(\varepsilon_i) = 0$.
- 異なる固体の誤差項は無相関である．$C(\varepsilon_i, \varepsilon_j) = 0\,(i \neq j)$.
- 誤差項の分散は一定である．$\forall i \in N, V(\varepsilon_i) = \sigma^2$.

過小定式化モデルを

$$\boldsymbol{y} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{e_1}$$

とおく．その最小自乗推定量を $\boldsymbol{b_1}$ とおくと,

$$\boldsymbol{b_1} = (\boldsymbol{X_1'}\boldsymbol{X_1})^{-1}\boldsymbol{X_1'}\boldsymbol{y}$$

を得る[6]．$\boldsymbol{X_1'}$ は $\boldsymbol{X_1}$ の転置行列である．

さて真に正しいモデルがフルモデルであるならば，過小定式化モデルにおける最小自乗推定量は

$$\begin{aligned} \boldsymbol{b_1} &= (\boldsymbol{X_1'}\boldsymbol{X_1})^{-1}\boldsymbol{X_1'}\boldsymbol{y} \\ &= (\boldsymbol{X_1'}\boldsymbol{X_1})^{-1}\boldsymbol{X_1'}(\boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{e}) \quad \boldsymbol{y} \text{ をフルモデルで置換} \\ &= \boldsymbol{\beta_1} + (\boldsymbol{X_1'}\boldsymbol{X_1})^{-1}(\boldsymbol{X_1'}\boldsymbol{X_2})\boldsymbol{\beta_2} + (\boldsymbol{X_1'}\boldsymbol{X_1})^{-1}\boldsymbol{X_1'}\boldsymbol{e} \end{aligned}$$

[5]以下の計算は Griffiths, Hill, and Judge(1993) による考察に浜田が若干の修正を加えたものである
[6]この推定値を得るのに，実は誤差の分布はなんであってもかまわない.

である．その期待値をとると

$$E(\boldsymbol{b_1}) = \boldsymbol{\beta_1} + (\boldsymbol{X_1'X_1})^{-1}(\boldsymbol{X_1'X_2})\boldsymbol{\beta_2}$$

s であり，真のパラメータからのズレが欠落変数バイアスである．

$$E(\boldsymbol{b_1}) = \boldsymbol{\beta_1} + \underbrace{(\boldsymbol{X_1'X_1})^{-1}(\boldsymbol{X_1'X_2})\boldsymbol{\beta_2}}_{\text{欠落変数バイアス}}$$

以上の例から，欠落変数バイアスは次のように一般的に定式化できる．

命題 1. k 個の説明変数うち，q 個 ($(k-p)$ 個) が欠落している場合には，欠落変数バイアスは

$$(\boldsymbol{X_1'X_1})^{-1}(\boldsymbol{X_1'X_2})\boldsymbol{\beta_2}$$

で与えられる．ただし X_1 は切片と過小定式化に含まれる説明変数のみからなる行列，X_2 は欠落した説明変数からなる行列，β_2 は欠落した真のパラメータからなるベクトルである．

証明. k 個の説明変数を持つモデルを

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

とおく．上式が n 個のユニット数だけ並んでいる．つまり n 本の式を同時に仮定している．これを行列とベクトルで一度に表現すると

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I_n).$$

ここで I_n は n 次の単位行列である．ベクトルと行列の定義は次の通りである．

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

この式を x_1 から x_p までの説明変数と，x_{p+1} から x_k までの説明変数に分解する．すなわち

$$\boldsymbol{y} = \boldsymbol{X_1\beta_1} + \boldsymbol{X_2\beta_2} + \boldsymbol{e}$$

とおく．ここで $\boldsymbol{X_1}, \boldsymbol{\beta_1}, \boldsymbol{X_2}, \boldsymbol{\beta_2}$ は具体的には,

$$\boldsymbol{X_1} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta_1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\boldsymbol{X_2} = \begin{pmatrix} x_{1p+1} & x_{1p+2} & \dots & x_{1k} \\ x_{2p+1} & x_{2p+2} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{np+1} & x_{np+2} & \dots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta_2} = \begin{pmatrix} \beta_{p+1} \\ \beta_{p+2} \\ \vdots \\ \beta_k \end{pmatrix}$$

とおく. 例 1 の計算より,

$$E(\boldsymbol{b_1}) = \boldsymbol{\beta_1} + \underbrace{(\boldsymbol{X_1'X_1})^{-1}(\boldsymbol{X_1'X_2})\boldsymbol{\beta_2}}_{\text{欠落変数バイアス}}$$

である．明示的に書けば

$$(\boldsymbol{X_1'X_1})^{-1}(\boldsymbol{X_1'X_2})\boldsymbol{\beta_2} = (X_1'X_1)^{-1}\begin{pmatrix} \sum_{j=p+1}^{k}\beta_j\left(\sum_{i=1}^{n}x_{ij}\right) \\ \sum_{j=p+1}^{k}\beta_j\left(\sum_{i=1}^{n}C(x_1,x_j)\right) \\ \vdots \\ \sum_{j=p+1}^{k}\beta_j\left(\sum_{i=1}^{n}C(x_p,x_j)\right) \end{pmatrix}$$

なお，この $(\boldsymbol{X_1'X_1})^{-1}(\boldsymbol{X_1'X_2})\boldsymbol{\beta_2}$ の計算結果は $(k-p)$ 行$\times$ 1 列のベクトルになる．したがって，過小定式化モデルに含まれる第 j 説明変数の推定値に対するバイアスは，$(k-p)$ 行$\times$ 1 列ベクトルの j 行目を見ればよい.

以上の計算によって，任意の欠落変数による任意の説明変数の推定値のバイアスが明示的に特定できる.

□

過小定式化モデルに含まれる説明変数 $x_1,\ldots,x_p$ と欠落した説明変数 $x_{p+1},\ldots,x_k$ の共分散だけでなく，欠落した真のパラメータ $\beta_{p+1},\ldots,\beta_k$ の大きさも，実は重要だという事が分かる.

3　Artificial Data に基づくバイアスの推定

さて前節の命題によって，欠落変数バイアスの内容は数学的に特定できた．しかし抽象過ぎて具体的イメージがわかないという読者もいるだろう．そこで次にバイアスがどの程度生じるのかを，Artificial Data を使って検討する.

手続き（モンテカルロ法シミュレーション）は以下の通りである.

- m 次元正規分布に従う確率変数の実現値ベクトル (m 次元) を n 個生成する
- m 個の変数を真のパラメータによって線型結合し，誤差項 (平均 0，分散 σ^2 の正規分布に従う確率変数) を加えて，応答変数を作る.
- データ行列（n 行$\times$ (m 列）を使って応答変数 y を $m-1$ 個の説明変数に回帰する（わざと過小定式化によって推定する）
- 得られた推定量が真のパラメータと，どのくらいズレているのかを確認する

以下に R のソースコードを例示する.

```
library(MASS) #多次元正規分布用
#真の説明変数は 2 個．説明変数は標準化している
omit2<-function(n,sd,r,b1,b2)
{#n:サンプル数,sd:誤差項標準偏差,r:説明変数間の相関,b1,b2:真のパラメータ
  mu  <- c(0,0)  #説明変数の平均ベクトルの定義
  Sigma <- matrix(c(1, r, r, 1), 2, 2)#分散共分散行列の定義
```

```
  exv<-mvrnorm(n, mu, Sigma)#説明変数用に 2 次元正規乱数を n 個生成
  er<-rnorm(n, mean = 0, sd)#攪乱項の生成
  y<- b1*exv[,1]+ b2*exv[,2]+er #真の関数形によるデータ生成，ベクトル演算
  data1<-data.frame(x1=exv[,1],x2=exv[,2],y)#data.frame を使って人工データを格納
  out1<-lm(y~x1,data=data1)#x2 をわざと欠落させた OLS 推定
  out2<-lm(y~x2,data=data1)#x1 をわざと欠落させた OLS 推定
  out3<-lm(y~x1+x2,data=data1)#フルモデルによる推定
  print(summary(out1));print(summary(out2));print(summary(out3))#強制出力
  print(cor(exv[,1],exv[,2]))#説明変数間の相関係数の確認
}

omit2(1000,5,0.3,10,10)
```

4　モンテカルロ・シミュレーションによるバイアス推定

さて既存のデータ分析に対して，欠落変数によるバイアスが予想される場合に，欠落変数行列 $\boldsymbol{X_2}$ と欠落した真のパラメータベクトル $\boldsymbol{\beta_2}$ を仮想的に挿入することで, おおよそのバイアス量が分かる[7].

計算に必要なパラメータは

$$(\boldsymbol{X_1'X_1})^{-1}(\boldsymbol{X_1'X_2})\boldsymbol{\beta_2}$$

のうち，$(\boldsymbol{X_1'X_1})^{-1}$ はデータから計算可能な分散共分散行列だから，仮想的に計算すべき量は

$$(\boldsymbol{X_1'X_2})\boldsymbol{\beta_2}$$

の部分だけである.

多くの説明変数と相関すると考えられる個人特性 (遺伝的特性など) は，おおよその相関を予想できるから，そこから共分散を逆算できるし，真のパラメータも変数のレンジから平均的な数値はある程度推測できるだろう.

5　今後の課題

- 先行研究に含まれていたバイアスのヴァーチャルな検証
- 操作変数法によるバイアス除去の効果の検証
- 不要な変数の効果が欠落変数バイアスによって有意になるかどうかの実験
- 交互作用の欠落についての数値実験
- 同時性バイアスを持った人工データによる数値実験

[7]操作変数法やプロキシ代入と何が違うのか？　というと，この方法はシミュレーションなので，実際にはデータを取得しなくてもよい，という点である（つまり実は VRA(浜田・石田 2005) の応用なのである）．適切な操作変数を見つけるのは難しいが，上記の方法ならば，先行研究を参照してバイアスを推測することができる.

References

Griffiths, William, E., R. Carter,Hill, and George, G. Judge, 1993, *Learning and Practising Econometrics*, John Wiley & Sons.

Griliches, Zvi, 1957, "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics*, 39(1):8-20.

羽森茂之, 2009,『ベーシック計量経済学』中央経済社.

永田靖・棟近雅彦, 2001,『多変量解析法入門 (ライブラリ新数学大系)』サイエンス社.

Stock, James, H., and Mark, M. Watson, 2012, *Introduction to Econometrics, Global Edition*, Pearson Education.

Theil, Henri, 1978, *Introduction to Econometrics*, Prentice-Hall.

Wooldridge, Jeffrey, M., 2010,*Econometric Analysis of Cross Section and Panel Data, Second Edition*, The MIT Press.