重回帰 (1)
Multiple Regression Models

別所俊一郎

2006 年 5 月 19 日

単回帰と重回帰

- 被説明変数に影響を与える他の要因を明示的に考慮しない OLS 推定量はミスリーディング
  - 不偏性・一致性を持たない
  - “omitted variable bias”
- 計測可能な他の要因を説明変数として回帰分析に加える
  - 単回帰の拡張
  - 推定や検定の手続きは単回帰に似ている

単回帰のOLS推定量の「偏り」

単回帰では、被説明変数に影響を与える説明変数以外の要因はすべて誤差項に集約

- 児童数とテストの点数の例では、学校の特性／生徒の特性／地域要因などなど
- ここでは、「英語を学習中の生徒の比率」に着目

ところが、いくつかの要因を無視してしまうとOLS推定量は偏りをもつかもしれない

- $E[\hat{\beta}_1] \neq \beta_1$ かもしれない
- $\hat{\beta}_1$ は真の β_1 よりも大きく（小さく）出やすい可能性
- 政策の効果は思ったようなものにならなくなってしまう

単回帰の OLS 推定量の「偏り」（例）

たとえば、英語を学習中の生徒の比率を無視すると、

- 英語を母国語としない生徒は、母国語とする生徒よりも平均的に点数が低い
- クラス児童数が多い地区には英語を母国語としない生徒が多い傾向がある（相関係数は正）
- OLS 推定量は、クラス児童数の効果だけではなく、英語を母国語としない生徒の比率の関係を間違って検出する（「拾ってしまう」）可能性
- $\hat{\beta}_1$ は真の β_1 よりも大きく出やすい可能性
- クラスの児童数を減らしても思うような効果は出ないかもしれない

"Omitted variable bias"

説明変数が、被説明変数に影響を与え、かつ分析から外れた変数と相関を持つとき、OLS 推定量は不偏性を持たない

以下の 2 つの条件をともに満たすとき、omitted variable bias が発生

1. omitted variable が説明変数と相関
2. omitted variable が被説明変数に直接に影響

どちらかの条件を満たさなければ、omitted variable bias は発生しない

“Omitted variable bias”（例）

例１： 英語を母国語としない生徒の比率

1. クラスの人数とは相関（因果関係ではない）
2. テストの点数とも直接相関

例２： テストの日の曜日

1. クラスの人数とは無相関
2. テストの点数とは相関する可能性

例３： 一人当たりの駐車場の面積

1. クラスの人数とは相関（因果関係ではない）
2. テストの点数とは無相関

Omitted variable bias と OLS の仮定

OLS の 3 つの仮定のうち、$E[u_i|X_i] = 0$ が成り立っていない

- omitted variable が被説明変数と直接に相関している $\Longleftrightarrow$ omitted variable が u_i に含まれている
- u_i と X_i が相関している $\Longrightarrow E[u_i|X_i] \neq 0$

したがって、OLS 推定量は不偏性も一致性も持たない

Omitted variable biasの大きさ

いま、$\mathrm{corr}(X_i, u_i) = \rho_{xu}$ とし、OLS 推定の残りの仮定は満たされているとすると、OLS 推定量は

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{xu}\frac{\sigma_u}{\sigma_X}$$

となり、第 2 項が偏り（bias）の大きさを示す

- Omitted variable bias は小標本でも大標本でも存在（不偏性も一致性も持たない）
- Bias の大きさは誤差項と説明変数の相関係数 ρ_{xu} の絶対値に依存
- Bias の方向は誤差項と説明変数の相関係数 ρ_{xu} の符号に依存

単回帰による対処方法：サンプル分割

重回帰モデル

単回帰モデルを拡張し、説明変数の数を増やす

- 他の説明変数 $X_{2i}, X_{3i}, ...$ の値を一定に保ったときの X_{1i} の Y_i への効果を推定

Population regression line

- さしあたって説明変数が 2 個（X_{1i}, X_{2i}）のケース
- 線形の重回帰モデルでは、

$$\underbrace{E[Y_i|X_{1i} = x_1, X_{2i} = x_2]}_{\text{条件付き期待値}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

β_0：切片（intercept）、定数項（constant）

β_1, β_2：X_{1i}, X_{2i} の傾き、係数（coefficient）

X_{1i}, X_{2i}：独立変数（independent variable）、制御変数（control variable）

Population multiple regression model

説明変数に入っていない決定要因を誤差項を用いて表すと、各観測値 i に対して

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, 2, ..., n$$

つねに 1 をとる変数 X_{0i}（$X_{0i} = 1 \forall i$）を考えると

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, 2, ..., n$$

より一般的には、k 個の説明変数を持つモデルを

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

ここで、誤差項が分散均一ならば、以下が成り立つ

$$\mathrm{var}(u_i | X_{1i}, X_{2i}, ..., X_{ki}) = \sigma_u \text{（一定で説明変数に依存しない）}$$

推定したいのは $\beta_0, \beta_1, ..., \beta_k, (\sigma_u)$ の係数たち。

重回帰モデルの OLS 推定量

単回帰モデルと同様、予測誤差の 2 乗和を最小化する推定量を考える
$(\beta_0, \beta_1, ..., \beta_k)$ の推定量を一般に $(b_0, b_1, ..., b_k)$ とする
このとき当てはめ値は

$$b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

予測誤差は

$$Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki})$$

最小化すべき 2 乗和は

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - \cdots - b_k X_{ki})^2$$

これを最小化する $(b_0, b_1, ..., b_k)$ を OLS 推定量といい、
$(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ で表す

重回帰モデルの OLS の基礎用語

単回帰モデルと同じ

OLS 回帰直線 2 次元では表現できないが、

$$\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

当てはめ値

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$$

残差

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki}$$

OLS 推定量 $n \times k$ の説明変数行列 $\mathbf{X}$ と $n \times 1$ の被説明変数ベクトル $\mathbf{y}$ を用いて

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

たいがいの計量ソフトウェアには組み込まれている