重回帰 (2)

別所俊一郎

2006 年 5 月 24 日

重回帰の仮定

1. $E[u_i|X_{1i}, X_{2i}, \cdots, X_{ki}] = 0$
   $\{X_{1i}, X_{2i}, \cdots, X_{ki}\}$ を所与としたときの u_i の条件付き分布の期待値がゼロ
2. $(X_{1i}, X_{2i}, \cdots, X_{ki}, Y_i) \sim$ i.i.d.
3. $0 < E[X_{1i}], E[X_{2i}], \cdots, E[X_{ki}], u_i < \infty$
   説明変数と誤差項の 4 次モーメントは有限
4. 完全な多重共線性（perfect multicollinearity）がない

これらの仮定が満たされると，OLS 推定量は望ましい性質を持つ

仮定 1〜3

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

説明変数を所与としたときの誤差項の条件付き分布の平均がゼロ．Y_i は平均的には population regression line の上にある

$(X_{1i}, X_{2i}, \cdots, X_{ki}, Y_i) \sim$ **i.i.d.** ：

単純な無作為抽出によって得られるサンプルであれば満たされる

$0 < E(X_{1i}), E(X_{2i}), \cdots, E(X_{ki}), u_i < \infty$ ：

説明変数と誤差項は有限の 4 次モーメントを持ち，極端な外れ値はない．中心極限定理を応用するためのやや技術的な仮定.

これらの仮定は基本的には単回帰の仮定の単純な拡張

仮定 4：完全な多重共線性がない

- ある説明変数が他の説明変数の完全な線形関数にはなっていない
- 完全な多重共線性があるとき，OLS は定義できない
  - 「ゼロで割る」ことになるから
- 計量アプリケーションでは，多重共線性がある場合には
  1. 多重共線を起こしている変数のいずれかを自動的に落とす
  2. 「計算できない」というエラーメッセージを返す
  3. Crash する
- 多重共線が起こるのは以下のケースが多い
  - 本質的に同義の変数が入っている
  - とくにダミー変数の場合，**サンプルのなかで**，定数項と区別できない

多重共線性の例

テストの点数に対するクラスあたり児童数（STR_i）の効果を見るため，英語を母国語としない児童のパーセンテージ（$PctEL_i$）の変数を加え，さらにもうひとつの変数を追加しようとしている状況を考えてみよう

英語を母国語としない人の比率（$FracEL_i$）： この変数は $PctEL_i$ と本質的に同義で,

$$PctEL_i = 100 \times FracEL_i$$

OLS 推定量が定義できないのは，非論理的な値を求めようとするから

- 「英語を母国語としない人の比率を一定にしたまま，英語を母国語としない人のパーセンテージをあげると，テストの点数はどうなりますか？」

多重共線性の例

「小さくない」クラス（NVS_i）： ダミー変数として，$STR_i < 12$ であれば 0，そうでなければ 1 の値を取るような変数を考える．このとき，データの中に $STR_i < 12$ なる観測値は存在せず，

$$NVS_i = 1 \quad \text{for all } i$$

定数項も $X_{0i} = 1, \forall i$ であるから，これらが多重共線を起こす
母集団に $NVS_i = 0$ なる地区があったとしても，サンプルのなかにそのような地区がなければ，多重共線性が発生してしまうために分析の対象とはなりえない

英語を母国語とする児童のパーセンテージ（$PctES_i$）：

$$PctES_i = 100 \times X_{0i} - PctEL_i$$

であるから，多重共線を発生する．多重共線は，2 つの説明変数間の現象ではなく，説明変数すべての組み合わせによって定義される．この場合，$PctES_i$，$PctEL_i$，定数項のいずれかが回帰式から外されれば，推定は可能．

多重共線性への対処

- 回帰式の特定化のミスであれば，簡単な場合も（どちらかの変数を落としても分析に支障がない）
- サンプルの問題であれば簡単ではない（ダミー変数の場合）
- 計量ソフトウェアの警告等で気づくことも多いが，そのつど考えることが必要

「不完全な」多重共線性

- 説明変数間に高い相関があること
- OLS 推定において，理論的には，とくに問題はない
- むしろ，潜在的に相関のある説明変数があるときに，それぞれの単独での効果を抽出するための分析手法が OLS.
- ただし，不完全な多重共線があるとき，係数の標準誤差が大きくなる傾向（後述）.

OLS 推定量の確率分布

サンプルが異なれば，計算される OLS 推定値も異なる
→OLS 推定量は確率変数

- 単回帰の場合
  - 適切な 3 つの仮定のもとで，OLS 推定量 $(\hat{\beta}_0, \hat{\beta}_1)$ は不偏推定量・一致推定量
  - $n \to \infty$ で漸近的に 2 変数正規分布に従う
- 重回帰でも同様
  - 前述の 4 つの仮定のもとで，OLS 推定量 $(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$ は $(\beta_0, \beta_1, \cdots, \beta_k)$ の不偏推定量・一致推定量
  - $n \to \infty$ で漸近的に多変量正規分布に従う
  - 証明は中心極限定理の応用（Ch. 16）

OLS 推定量の標準誤差

単回帰の場合

- 大数の法則が成り立つので，期待値を sample counterpart に代えれば

$$\hat{\sigma}^2_{\hat{\beta}} / \sigma^2_{\hat{\beta}} \xrightarrow{p} 1$$

重回帰の場合

- 基本的な考え方は単回帰のケースと同じ．大数の法則を利用すれば $SE(\hat{\beta}_j)$ の一致推定量をえられる
- OLS 推定量 $(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$ は $n \to \infty$ で漸近的に多変量正規分布に従い，その相関係数が存在（共分散行列）
- 説明変数間に相関があるため，係数の推定量も相関を持つ

係数の１つについての仮説検定と信頼区間

他の条件を一定にしたときのある説明変数が被説明変数に与える効果についての仮説検定

- （例）英語を母国語としない児童のパーセンテージを一定としたときのクラス児童数の変化に対する標準テストの点数の変化
- クラスの児童数の係数 β_1 の大きさについての仮説検定
- より一般的には，両側検定のためには

$$H_0 : \beta_j = \beta_{j,0} \quad \text{v.s. } H_1 : \beta_j \neq \beta_{j,0}$$

単回帰の仮説検定と基本的な手続きは同じ

係数の 1 つについての仮説検定と信頼区間

仮説検定の手続き

⇒OLS 推定量は H_0 が真であるときに漸近的に既知の正規分布に従う

1. 標準誤差 $\mathrm{SE}(\hat{\beta}_j)$ を求める
2. t 値を求める

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{\mathrm{SE}(\hat{\beta}_j)}$$

3. p 値を求める（両側検定）→有意水準 p で棄却される

$$p = 2\Phi(-|t|)$$

信頼区間の形成

$$(\hat{\beta}_j - 1.96\mathrm{SE}(\hat{\beta}_j), \hat{\beta}_j + 1.96\mathrm{SE}(\hat{\beta}_j))$$

（例）児童数の点数への効果

LS // Dependent Variable is TESTSCR, Date: 05/22/06 Time: 21:29
Sample: 1 420, Included observations: 420
White Heteroskedasticity-Consistent Standard Errors and Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	649.5779	15.45834	42.02119	0.0000
STR	-0.286399	0.482073	-0.594100	0.5528
EXPN	3.867902	1.580722	2.446920	0.0148
EL PCT	-0.656023	0.031784	-20.63975	0.0000

R-squared	0.436592	Mean dependent var	654.1565
Adjusted R-squared	0.432529	S.D. dependent var	19.05335
S.E. of regression	14.35301	Akaike info criterion	5.337398
Sum squared resid	85699.71	Schwarz criterion	5.375876
Log likelihood	-1712.808	F-statistic	107.4547
Durbin-Watson stat	0.742238	Prob(F-statistic)	0.000000

不完全な多重共線性の評価

- 説明変数間に高い相関があることをいい，理論的にはとくに問題はない
- ただし，不完全な多重共線があるとき，係数の標準誤差が大きくなる傾向．説明変数が 2 個のとき，標準誤差は

$$\sigma^2_{\hat{\beta}} = \frac{1}{n}\left[\frac{1}{1-\rho^2_{x_1x_2}}\frac{\sigma^2_u}{\sigma^2_{x_1}}\right]$$

となるので，X_1 と X_2 の相関係数 $\rho_{x_1x_2}$ が大きくなると標準誤差は大きくなる
片方だけの効果を取り出しにくくなる（係数が不安定になる）

係数制約検定

Joint null hypothesis

- 例：テストの点数には教育支出もクラスの人数も影響を与えない

  $H_0 : \beta_1 = 0$ かつ $\beta_2 = 0$ v.s. $H_1 : \beta_1 \neq 0$ または $\beta_2 \neq 0$

- 2 つの制約を同時にかけた仮説
- より一般的には

  $H_0 : \beta_j = \beta_{j,0}$ かつ $\beta_m = \beta_{m,0}$ など q 個の制約

  v.s. H_1 : 少なくとも 1 つの制約が成り立たない

- H_0 を構成する q 個の制約条件のうち，少なくとも一つが成り立たなければ H_0 は偽

1 つずつ検定してはいけないのか?

$H_0 : \beta_1 = 0$ かつ $\beta_2 = 0$ を検定するとき，t 検定を 2 回やっては？

- (t_1, t_2) は 2 変量正規分布に従い，各周辺分布が標準正規分布
- (t_1, t_2) が互いに相関を持たないとき，H_0 が真であるのに H_0 を棄却してしまう確率は，有意水準 5%の両側検定を 2 回行うと 5%より大きい
- H_0 が棄却されないのは，$|t_1| < 1.96$ かつ $|t_2| < 1.96$ のとき

$$\begin{aligned} \Pr(|t_1| < 1.96, |t_2| < 1.96) &= \Pr(|t_1| < 1.96) \times \Pr(|t_2| < 1.96) \\ &= 0.95^2 = 0.9025 \end{aligned}$$

- H_0 を棄却する確率は 9.75%で，棄却しやすい
- (t_1, t_2) が互いに相関していればもっと複雑だが，H_0 のもとでの棄却確率と有意水準が異なってしまう
- Bonferroni の方法

F 統計量（F statistics）

- Joint null hypothesis の検定方法
- H_0 が $q = 2$ 個の制約（$\beta_1 = 0$ かつ $\beta_2 = 0$）であるとき

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1 t_2} t_1 t_2}{1 - \hat{\rho}_{t_1 t_2}}\right) \sim F_{2,\infty}$$

(t_1, t_2) が互いに相関を持たないとき，t 値の 2 乗の平均値

$$F = \frac{1}{2}\left(t_1^2 + t_2^2\right) \sim F_{2,\infty}$$

- より一般に，q 個の線形制約（$\mathbf{R}\beta = \mathbf{r}$）に対して

$$F = (\mathbf{R}\hat{\beta})'[\mathbf{R}\hat{\mathbf{\Sigma}}_{\hat{\beta}}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}) \sim F_{q,\infty}$$

- p 値の算出には F 分布を用いるが，大標本の χ^2 近似（Chi-squared approximation）を用いてもよい

$$\chi_q^2 = qF_{q,\infty}$$

よくつかう F 統計量

"Overall" regression F-statistics

- 説明変数はどれひとつとして被説明変数の変動を説明していない，という仮説の検定

$$H_0 : \beta_1 = 0 \text{ かつ } \beta_2 = 0 \text{ かつ } \cdots \beta_k = 0$$

$$H_1 : \text{少なくとも 1 つの } j \text{ に対して } \beta_j \neq 0$$

- このとき，F 統計量は $F_{k,\infty}$ に従う

$q = 1$ のとき

- 帰無仮説は 1 つの係数についての仮説になる

$$F = t^2$$

Heteroskedasticity-robust な F 統計量

- 共分散行列が Heteroskedasticity-robust
- 計量ソフトではしばしばオプション指定が必要