パネル単位根検定における問題と多重検定の適用

松木　　隆*

Some Issues of Panel Unit Root Tests and Application of Multiple Testing

Takashi MATSUKI

要旨：本稿では、実証研究において一般的に用いられるようになってきたパネル単位根検定について、それが抱える解釈上の問題点を指摘する。そして、その問題の修正について、多重検定を用いた方法を考察する。

Abstract :

This paper points out that panel unit root tests have some substantial issues in them, which have become commonly used in empirical analysis. Then, it considers the application of multiple testing to resolve these issues.

キーワード：パネル単位根検定、多重性問題、Familywise Error Rate、多重検定

1. はじめに

時系列データを用いた実証分析において、データに単位根が存在するかどうかを確認する単位根検定は、計量経済学の一般的な分析手法の 1 つとして広く用いられている。そして、それをパネルデータへ拡張したパネル単位根検定も、今や多くの実証分析ソフトにおいて実装されており、標準的な手法になりつつある。例えば、実質為替レートの定常性を検定することによる購買力平価説の検証を行う研究や、失業率の履歴現象（hysteresis）を単位根の有無に帰着させた研究、1 人当たり実質所得の 2 国間差や平均偏差の安定性を確認することによるコンバージェンス（収斂）研究など、さまざまな研究課題においてパネル単位根検定が適用されている。

実証分析でよく用いられる手法としては、Levin, Lin, and Chu（2002）検定（LLC 検定）、Im, Pesaran, and Shin（2003）検定（IPS 検定）や Fisher-type 検定（Maddala and Wu（1999）、Choi（2001））が挙げられる。この利用傾向の 1 つの理由は、実証分析ソフトにこれらの手法が実装されているかどうかに依存する部分もあるであろう。例えば計量分析に良く用いられる Eviews（ver 9.5）では、LLC 検定、IPS 検定、Fisher-type 検定に加えて、Breitung（2000）と Hadri（2000）も実行できるようになっている[1)]。

しかしながら、これらのパネル単位根検定は、実証結果から何らかの結論を導く段階において本質的な問題を抱えている。この問題は、Pesaran（2014）においても指摘がなされている。それは、パネル単位根検定実行の際に、パネルデータを構成する 1 つ 1 つの系列（時間軸方向にデータを捉えた場合の系列）の単位根の有無を同時帰無仮説（joint null hypothesis）として検定するため、仮に帰無仮説が有意に棄却されたとしても、「個別にはどの系列が定常なのかを特定できない」という技術的な欠陥である。そして、このことは、特に

*大阪学院大学経済学部教授

1）STATA（ver 14）では、上記の検定に加えて、Harris and Tzavalis（1999）検定も行える。

実証分析の結果解釈の際に大きな問題となってくる。例えば、前述の購買力平価説の検証において、仮にパネル単位根検定によって単位根帰無仮説が有意に棄却されたとしても、分析対象国中のどの国の為替レートが定常であるのか（つまり購買力平価説が成立しているのか）を判断できない、ということになる。

この事実を受けて、実証分析では個々の仮説の正誤を判定する手がかりとして、パネルデータを構成する個々の時系列の個別検定（Augmented Dickey-Fuller（ADF）検定など）の結果を参考にする場合が見られる。しかしながら、この方法は別の問題（多重性問題）を引き起こす可能性がある。第3節において詳述するが、複数の仮説がありそれを1つ1つ検定する場合、すべての検定作業を1つの包括的検定と見なすと、少なくとも1つの仮説を誤って棄却する確率は、個々の仮説を検定する際に設定する有意水準を大きく上回る場合がある。これが多重性問題である。

上記2つの問題を同時に解決する有効な方法の1つは、多重検定（multiple testing）の適用である。最近になって、Hanck（2009）、Moon and Perron（2012）、Matsuki and Sugimoto（2013）、Matsuki（2016）において、それが行われている[2)]。多重検定を利用することで、実証分析者は複数の仮説に関する検定全体の過剰棄却問題を回避しつつ、個々の仮説の正誤を判定することができるため、仮説棄却に伴う結果解釈上の曖昧さから解放されることが期待できる。

本稿では、パネル単位根検定を用いた実証分析において、分析者が直面する結果解釈上の問題を指摘し、次に多重検定のアイデアとその適用手順を紹介する。さらに、実際のデータに対して期待した通りに検定が働いているのかどうか、検定のパフォーマンスを検証し、また実証例も示す。

2. パネル単位根検定の仮説設定と問題点

本節では、パネルデータ（$\{y_{i,t}\}$, $i = 1,\ldots,N$, $t = 1,\ldots,T$）を用いた単位根検定として、Levin, Lin, and Chu（2002）検定に代表される pooled test と Im, Pesaran, and Shin（2003）検定に代表される averaged test および Fisher-type 検定の p-value combination test を紹介し、それらが持つ問題点を指摘したい[3)]。Pooled test はデータをプールして回帰を行うことにより得られる t 統計量を利用し、averaged test、p-value combination test は各クロスセクション単位（i）について個別に回帰を行い、得られた N 個の t 統計量またはその p 値を集約する。Averaged test は t 統計量のクロスセクション平均を基準化した統計量を使用し、p-value combination test は N 個の t 統計量の p 値を集約して一つのパネル単位根検定統計量を構築する。

2.1 Pooled test

パネルデータの生成過程（Data Generating Process（DGP））を以下の（1）式と仮定する。また、回帰式には以下の（2）式を用いる。

$$\Delta y_{i,t} = \phi_i y_{i,t-1} + \varepsilon_{i,t}\ \varepsilon_{i,t} \sim i.i.d.(0,1)\quad for\quad all\quad i\ and\ t\ i = 1,\ldots,N\quad 、t = 1,\ldots,T \qquad (1)$$

$$\Delta \mathbf{y} = \phi_{pool}\,\mathbf{y}_{-1} + \mathbf{e}\ （\text{pooled regression}） \qquad (2)$$

ここで、$\varepsilon_{i,t} \sim i.i.d.(0,1)$ *for all i and t* とは、$\varepsilon_{i,t}$ が独立同一分布に従い、その分布の期待値が0、分散が1であり、その性質はクロスセクション方向（i）にも、時系列方向（t）にも保たれる、の意味である。また、（2）式の **y**、**e** は以下のベクトルを表す。

$$\mathbf{y} = (y_{1,1}\ \cdots\ y_{N,1}\ \cdots\ y_{1,T}\ \cdots\ y_{N,T})' = (\mathbf{y}_1'\ \cdots\ \mathbf{y}_T')'、$$

2）別のアプローチとして、Chortareas and Kapetanios（2009）、Smeekes（2010）らは、検定を sequential に行うことによる方法を提案している。

3）本稿で想定するデータ生成過程及び検定構築手順は、LLC 検定、IPS 検定のそれよりも単純化したものである。そのため、厳密に言えば、ここで対象とする検定は LLC 検定、IPS 検定そのものとは異なる。しかし、本質的な検定の設定は保たれており、それらが抱える問題点を考察するには問題がない。

$\mathbf{e} = (e_{1,1} \cdots e_{N,1} \cdots e_{1,T} \cdots e_{N,T})' = (\mathbf{e}_1' \cdots \mathbf{e}_T')'$。

ここでの仮説は全てのクロスセクション単位に共通であり、以下のように設定される。

$$H_0\text{（帰無仮説）}：\phi_i = \phi_{pool} = 0 \ \ \textit{for all } i^{4)}$$
$$H_1\text{（対立仮説）}：\phi_i = \phi_{pool} < 0 \ \textit{for all } i \qquad (3)$$

また、このときの（2）式の係数推定量と検定統計量は、それぞれ以下のようになる。

$$\hat{\phi}_{pool} = \frac{\mathbf{y}_{-1}'\Delta\mathbf{y}}{\mathbf{y}_{-1}'\mathbf{y}_{-1}} = \frac{\sum_{t=2}^{T}\mathbf{y}_{t-1}'\Delta\mathbf{y}_t}{\sum_{t=2}^{T}\mathbf{y}_{t-1}'\mathbf{y}_{t-1}} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i,t-1}\Delta y_{i,t}}{\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i,t-1}^2} \quad (4)$$

$$t_{pool} = \frac{\mathbf{y}_{-1}'\Delta\mathbf{y}}{s\sqrt{\mathbf{y}_{-1}'\mathbf{y}_{-1}}} \frac{\sum_{t=2}^{T}\mathbf{y}_{t-1}'\Delta\mathbf{y}_t}{s\sqrt{\sum_{t=2}^{T}\mathbf{y}_{t-1}'\mathbf{y}_{t-1}}} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i,t-1}\Delta y_{i,t}}{s\sqrt{\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i,t-1}^2}} \quad (5)$$

ここで、
$s^2 = (N(T-1))^{-1}(\Delta\mathbf{y} - \hat{\phi}_{pool}\,\mathbf{y}_{-1})'(\Delta\mathbf{y} - \hat{\phi}_{pool}\,\mathbf{y}_{-1})$ である。t_{pool} は時系列数（T）が大きくなる時（$T \to \infty$）、以下の分布に分布収束する。

$$t_{pool} \xrightarrow{d} \frac{\sum_{i=1}^{N}\int_0^1 W_i(r)dW_i}{\sqrt{\sum_{i=1}^{N}\int_0^1 W_i^2(r)dr}}$$

ここで、$W(r)$ は標準ブラウン運動であり、$\xrightarrow{d}$ は分布収束を表す[5)]。また、上式の分子を $\sqrt{N}$ で割ったものは、クロスセクション数（N）が大きくなる時（$N \to \infty$）、中心極限定理により正規分布に収束する。

$$\frac{\sum_{i=1}^{N}\int_0^1 W_i(r)dW_i}{\sqrt{N}} = \frac{\frac{1}{2}\sum_{i=1}^{N}(\chi_i^2(1)-1)}{\sqrt{N}} \xrightarrow{d} N(0,1/2)。$$

また、$E(\int_0^1 W^2(r)dr) = 1/2$ であるから[6)]、クロスセクション数が大きくなると、大数の法則より

$$\frac{\sum_{i=1}^{N}\int_0^1 W_i^2(r)dr}{N} \xrightarrow{p} \frac{1}{2}$$

も示される。ここで、$\xrightarrow{p}$ は確率収束を表す。これより、$N \to \infty$ の時、

$$t_{pool} \xrightarrow{d} N(0,1)$$

となる。したがって、t_{pool} 統計量は、時系列数、クロスセクション数がともに大きくなる時、標準正規分布を用いて仮説検定を行うことができる。

（3）式の仮説設定において、それの意味するところは、帰無仮説が「全ての系列（$i = 1,\ldots,N$）について単位根がある」であり、対立仮説が「全ての系列について単位根がない（定常である）」である。これより、本検定において帰無仮説が棄却される場合、考察対象の N 系列は全て定常系列であると結論付けるが、これは実データの性質を考えると現実的な解釈ではないだろう。例えば、実質為替レートのパネルデータを用いた実証分析において、N 個の系列とは N ヵ国の実質為替レートデータのことであり、これが全て定常であるとは、N ヵ国全てで実質為替レートが定常であり、したがって、（weaker version の）購買力平価説が成立することを意味する。しかしながら、現実には、定常なデータと非定常なデータ（単位根を持つデータ）の混合によりパネルデータが構成されていることが多く、つまり、購買力平価説が成立する国とそうではない国が分析対象国に混在するのが一般的であろう。しかし、ここでの仮説設定は非常に強い制約を最初から課しているため、仮に帰無仮説が棄却されてもそれが現実を適切に反映した結論なのかは疑問が残る。

4）回帰式が（2）式である場合、帰無仮説と対立仮説の両方において、係数に homogeneity 制約を課すことになる。つまり、帰無仮説の下では、全てのクロスセクション単位 i において $\phi_i = 0$ を要求し、対立仮説の下では、全ての i について $\phi_i = \phi < 0$ を要求する。

5）標準ブラウン運動とは、$0 \le r \le 1$ の範囲に定義される連続確率過程の 1 つであり、以下の性質を満たすものである。①原点から始まる（$W(0) = 0$）。②時点の重ならない増分は互いに独立である。③$0 \le s < r \le 1$ に対して、増分 $W(r) - W(s)$ は期待値 0、分散 $r - s$ の正規分布に従う。

6）この値は解析的に計算可能。Levin et al.（2002）によっても与えられている。

2.2 Averaged test

DGP を（1）式とし、回帰式を以下の式とする。

$$\Delta y_{i,t} = \phi_i y_{i,t-1} + e_{i,t} \quad t = 1, \ldots, T \tag{6}$$

ここでの帰無仮説及び対立仮説は

$$H_0 : \phi_i = 0 \quad \textit{for all } i$$
$$H_1 : \phi_i < 0 \quad \textit{for some } i \text{（}\phi_i \text{ のいくつかは } \phi_i = 0 \text{ でもよい）} \tag{7}$$

である。検定統計量の構築は、まず各 i 系列について（6）式の回帰から以下の t 統計量を得る。

$$t_i = \frac{\sum_{t=2}^{T} y_{i,t-1}\Delta y_{i,t}}{s_i \sqrt{\sum_{t=2}^{T} y_{i,t-1}^2}}$$

ここで、$s_i^2 = (T-1)^{-1}\sum_{t=2}^{T}(y_{i,t} - \hat{\rho}_i y_{i,t-1})^2$。この t_i は、時系列数が大きくなる時、以下の Dickey-Fuller t 分布に分布収束する（Phillips（1987））。

$$t_i \xrightarrow{d} \frac{\int_0^1 W(r)dW(r)}{\sqrt{\int_0^1 W^2(r)dr}} = \eta$$

また、t_i は独立であるから、$\bar{t} = \frac{1}{N}\sum_{i=1}^{N} t_i$ を考えると、帰無仮説の下で中心極限定理が使えて、$N \to \infty$ の時、

$$Z_{bar} = \frac{\sqrt{N}\{\bar{t} - E(\eta)\}}{\sqrt{Var(\eta)}} \xrightarrow{d} N(0,1)$$

を得る（$E(\eta)$、$Var(\eta)$ は Nabeya（1999）により与えられている）[7]。したがって、上記の統計量も時系列数とクロスセクション数が大きい時、標準正規分布を用いて検定が行える。

（7）式の仮説設定において、帰無仮説は「全ての系列について単位根がある」であり、対立仮説は「いくつかの系列について単位根がない」である。本検定における仮説設定、特に対立仮説の設定は、先の（3）式よりも制約が緩いが、この場合には別の問題がある。つまり、帰無仮説が棄却された時、N 個の考察対象系列のうちのいくつかは定常であるわけだが、本検定ではそれが具体的にどの系列かは特定できない。さらに、N 個の系列中にいくつ定常系列が含まれるかも分からない。これでは、先の購買力平価説の実証分析の例において、考察対象国中のいくつかには購買力平価説が成立しているが、それがどの国であり、合計何ヵ国において成立しているかは不明であるから、極めて曖昧な結果である。したがって、この結果から有意義な結論を導くことはできないであろう。

2.3 P-value combination test

本検定の DGP、回帰式、仮説は averaged test と同じである。最初に Maddala and Wu（1999）の手法を紹介する。本手法のアイデアは Fisher（1932）に基づいている。Fisher（1932）は N 個の連続かつ独立な検定統計量の p 値（p_i、$i = 1, \ldots, N$）が独立に一様分布 $U(0,1)$ に従うことを利用し、以下の統計量を構築した。

$$p_\lambda = -2\sum_{i=1}^{N} \log p_i$$

上記の統計量は自由度 $2N$ のカイ 2 乗分布に従う。Maddala and Wu（1999）は上式の p 値について、ADF_t 統計量の p 値を用いて p_λ をパネル単位根検定とした。彼らの統計量は、$T \to \infty$ のとき（ここで N は固定）、漸近的に自由度 $2N$ のカイ 2 乗分布に従う。

一方、Choi（2001）は標準正規分布の分布関数を利用して以下の統計量を提案した。

$$Z = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \Phi^{-1}(p_i)$$

ここで、$\Phi(\cdot)$ は標準正規分布の分布関数である。時系列数、クロスセクション数が大きい時、Z は標準正規分布に従う。

$$Z \xrightarrow{d} N(0,1)$$

7）有限の T のとき、$E(\eta)$、$Var(\eta)$ はシミュレーションにより求める。

本検定においても、仮説の設定は先の averaged test と同じであるから、同じ解釈上の問題点が残る。

ここで紹介した 3 つの検定手法は、クロスセクション間の独立性を前提としているが、実際のパネルデータではクロスセクション間に何らかの従属性（相関）を持つ場合が一般的である。実際の分析では、パネルデータにおけるクロスセクション間の従属性を考慮して構築された検定手法、例えば Bai and Ng（2004）、Moon and Perron（2004）、Breitung and Das（2005）、Pesaran（2007）などを用いる必要があろう。しかしながら、これらの検定もクロスセクション間に独立性を仮定した検定の拡張であるから、パネル単位根検定の仮説設定及びその解釈について、同様の問題を抱えている。

3. 多重性問題

2 節で述べたように、パネル単位根検定における解釈上の問題は実証結果に満足な意味を与えない。そこで、N 個の系列があり、それぞれの系列についてサンプル数（時系列数）が一変量検定において十分な検出力を得られるものであれば、一変量検定を個々の仮説（H_i）に繰り返し適用して検定結果を得ることも考えられる。実証分析においては、しばしば見られる方法である。

しかしながら、この方法においては別の問題（多重性問題）が起こりうる。多重性問題（multiplicity problem）とは、複数の仮説を検定するために、一変量時系列を対象とした検定を繰り返し適用する場合、これらの適用を包括的に 1 つの検定として考えると、検定全体として少なくとも 1 つの仮説を間違って棄却する確率（familywise error rate（FWER））が事前に設定した個々の検定の有意水準を上回ってしまう現象である。

多重性問題の例を 1 つ紹介しよう。いま 10 の独立な検定があるとし、その各々を 5% の有意水準の下で仮説を検定するとする。T_i を検定統計量、c_i を判定値（$i = 1, \ldots, 10$）とすると、$\Pr(T_i \geq c_i) = 0.05$ である。ここで、もし $T_i \geq c_i$ ならば、対応する帰無仮説は棄却される。もし全ての仮説が正しいならば、全ての検定において少なくとも 1 つの仮説が間違って棄却される確率は

$$
\begin{aligned}
&\Pr\{(T_1 \geq c_1) \cup (T_2 \geq c_2) \cup \cdots \cup (T_{10} \geq c_{10})\} \\
&= 1 - \Pr\{(T_1 < c_1) \cap (T_2 < c_2) \cap \cdots \cap (T_{10} < c_{10})\} \\
&\qquad\qquad \text{（排反事象より）} \\
&= 1 - \Pr(T_1 < c_1)\Pr(T_2 < c_2) \cdots \Pr(T_{10} < c_{10}) \\
&\qquad\qquad \text{（検定の独立性より）} \\
&= 1 - 0.95^{10} = 0.4013。
\end{aligned}
$$

ここで、最後の確率 0.4013 が familywise error rate である。この値は、検定前に設定した 5% の有意水準よりもずっと大きい。

これはシミュレーションを使っても示すことができる。データが以下の（8）式のランダム・ウォークに従っているとし、回帰式を（9）式として、その係数 ϕ_i の t 統計量（Dickey-Fuller t 統計量）を求めて単位根仮説を検定する。そして、この手順を N 回繰り返す。

$$
y_{i,t} = y_{i,t-1} + \varepsilon_{i,t} \qquad \varepsilon_{i,t} \sim i.i.d.N(0,1) \quad for\ all\ i\ and\ t \quad i = 1, \ldots, N、t = 1, \ldots, T \tag{8}
$$

$$
\Delta y_{i,t} = \phi_i y_{i,t-1} + \varepsilon_{i,t} \tag{9}
$$

このとき、N 回の仮説検定を 1 つの包括的な検定とみなし、少なくとも 1 回は間違って帰無仮説を棄却する確率（FWER）を計算すると表 1 のようになる。表 1 より、最も過剰棄却が少ない $N = 5$ のどのケースでも FWER は 0.2 を超えており、クロスセクション数 N が増加するに従い

表 1　DFt 検定の FWER

T	N	FWER
50	5	0.238
	10	0.396
	20	0.635
100	5	0.222
	10	0.402
	20	0.633
200	5	0.218
	10	0.405
	20	0.637

（出所）筆者計算。
（注）各（T, N）に対して、2000 回の繰り返し計算を行っている。

FWER が上昇することがわかる。また、時系列数 T に関係なく FWER が上昇することも明らかである。特に、T に関係なく、$N = 20$ のときの FWER は 0.63 であり、非常に大きい。この場合、N 個の仮説について、約 63% の確率で少なくとも 1 つは間違って棄却してしまう。

4. 問題の解決法：多重検定の適用

4.1 多重検定のアイデア

2、3 節において指摘した 2 つの問題について、これらを同時に解決する 1 つの方法は、多重検定（multiple testing）の適用である。多重検定は、複数の仮説がある場合に、FWER をある望ましい水準（例えば 5% など）にコントロールしつつ検定を行うための方法である。仮説を同時に検定する、または逐次検定することができる。

多重検定の最も基本的なものに Bonferroni 不等式を利用した方法がある。Bonferroni 不等式とは、例えば、N 個の仮説 $H_{1,0}, H_{2,0}, \ldots, H_{N,0}$ があるとき、A_i を正しい帰無仮説が誤って棄却される事象であるとすると、以下のように表される。

$$Pr(\bigcup_{i=1}^{N} A_i) \leq \sum_{i=1}^{N} Pr(A_i) \qquad (10)$$

上式の左辺は「正しい帰無仮説のうち少なくとも 1 つの仮説が誤って棄却される確率」であり、右辺は「個々の正しい帰無仮説が誤って棄却される確率の合計」である。個々の仮説を有意水準 α で検定を行っているとすると、$Pr(A_i) \leq \alpha$ であるから、(10) 式の右辺の確率を有意水準以下にコントロールしようとすれば $\sum_{i=1}^{N} Pr(A_i) \leq N \cdot \frac{\alpha}{N} = \alpha$ とすればよい。つまり、個々の仮説について、有意水準を α/N としてそれぞれ検定を行えば、検定全体として有意水準 α で検定していることとなる。これが Bonferroni 不等式を用いた多重検定である[8]。パネルデータに対して多重検定を適用する際の特徴としては、検定統計量を得る際に推計する推計式についての柔軟性の高さが挙げられる。具体的には、モデルにおける定数項やタイム・トレンドの有無、構造変化の有無やその時期について、各系列 i について異なっていてもよい。また、パネルデータを用いた場合でも検定統計量の計算は比較的容易な場合が多い[9]。

一方、多重検定は α 以下に検定の第 1 種の過誤の確率がコントロールされ、一般的な正規検定や t 検定のように有意水準に等しくはならない（このような性質を持つ検定を保守的な検定という）。そのため、検定の検出力が低くなる傾向にある。また、関連する性質として、仮説数（N）が大きい場合には、仮説を棄却するには非常に厳しい有意水準を個々の検定または検定全体が満たさないといけなく、検出力が低下する問題がある。例えば、Bonferroni 法では、$\alpha = 0.05$、$N = 10$ のとき、$\frac{\alpha}{N} = \frac{0.05}{10} = 0.005$ であるから、個々の仮説を 0.5% の有意水準で検定しなければならない。これが、$N = 20$ であれば 0.25% の有意水準、$N = 50$ であれば 0.1% の有意水準での検定となり、仮説を棄却し難くなることが分かるであろう。

4.2 適用手順

ここで多重検定の適用手順を 1 つ示す。パネルデータ $(\{y_{i,t}\},\ i = 1, \ldots, N,\ t = 1, \ldots, T)$ について、(8) 式をデータ生成過程として、(9) 式を各 i について回帰し、ϕ_i の t 統計量（またはその p 値）を求める。各帰無仮説 $H_{i,0}$ を多重検定の FWER を α に設定して検定する。検定の概念図は図 1 のようになる。

上から順に説明をすると、N 組あるデータ $\{y_{i,1}, \ldots, y_{i,T}\}$ について、それぞれ検定統計量 T_i（またはその p 値 p_i）を計算する[10]。次に、対応する仮説 $H_{i,0}$ について、与えられた有意水準または FWER の下で、それぞれ検定を行う。その際、one step 法は複数の仮説を同時に検定し結果

8）その他にも Holm（1979）、Simes（1986）、Benjamini and Hochberg（1995）など、多くの多重検定がある。多重検定のサーベイについては、Hochberg and Tamhane（1987）、Tamhane（1996）などが参考になる。
9）ただし、p 値を bootstrap 法を用いて計算する場合は計算量が多くなる。
10）ここで考える多重検定は、クロスセクション間における従属性を考慮する。

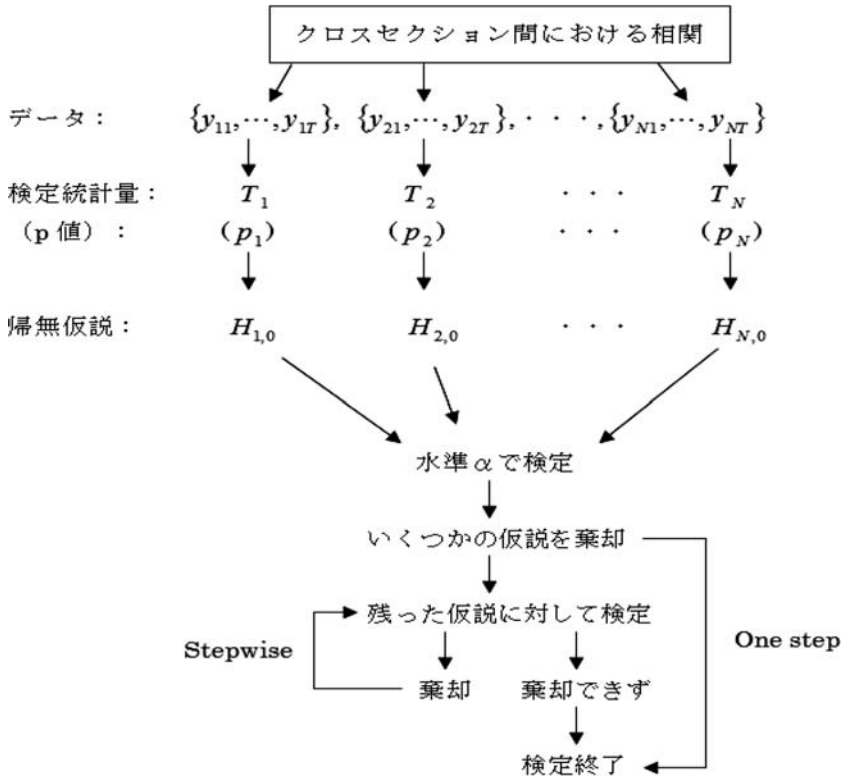

図 1　多重検定概念図

（出所）筆者作成。

を得る方法であり、仮説検定の便宜的な実行順に関係なく同じ条件（有意水準）の下で全ての仮説を検定する。Stepwise 法は仮説を逐次検定していく方法であり、最も棄却しやすい仮説（または最も棄却しにくい仮説）から順に検定を行う。その際に、検定の有意水準を少しずつ緩和していくことができるため、一般的に one step 法よりも検出力が高い。

5. 検定パフォーマンスと実証例

ここでは、Romano and Wolf（1995）が提案した多重検定を利用して、パネル単位根検定とする。Romano and Wolf 法は、同時信頼区間を用いた逐次検定であり、最初のステップでいくつかの仮説を棄却した場合には、次のステップでそれらを除外して新たに同時信頼区間を再計算し、残った仮説の検定を行う。そして、仮説が棄却できなくなるまでこの手順を繰り返す。また、同時信頼区間の計算の際には bootstrap 法が用いられる。

5.1　検定パフォーマンス

検定の FWER と検出力を見ることで、多重検定の検定パフォーマンスを確認する[11]。ここでは、帰無仮説の下でデータがランダム・ウォーク過程で生成され、対立仮説の下で定常過程である場合（表 2 a）と、1 回の構造変化を持つ定常過程である場合（表 2 b）について、それぞれ個別検定と多重検定の検定特性を考察する。表中の値は、各（T,N）に対して 2000 回の繰り返し計算により得られている。

表 2 a より、個別検定の FWER は、時系列数やクロスセクション数が大きくなるにつれて値が大きくなり、過剰棄却がより深刻になることが分かる。一方、多重検定においては、どのケースで

11）検出力として、棄却された仮説数を総仮説数で割った average power を用いる。

も FWER は 0.05～0.08 程度であり、ほぼ適切にコントロールされている。検出力を見ると、全てのケースで多重検定の検出力が個別検定のそれを上回っている。特に、時系列数が少ない場合（$T = 50$）、その差は大きい。

表 2 b において、一部例外はあるものの、総じて多重検定の方が個別検定よりも FWER を低くコントロールしている。特に、クロスセクション数が大きい場合（$N = 10$）には、個別検定の過剰棄却は大きくなる。検出力では、構造変化の大きさ（δ）が大きくなるにつれて、両検定の検出力は高まる傾向がある。ただし、全てのケースにおいて、多重検定の検出力が個別検定のそれより大きい。特に、時系列数が小さく、構造変化の大きさが小さい時、その差はより顕著になる（例えば、$T = 50$ または 100、$\delta = 1$ の場合など）。

ここでは 2 つの結果を観察したが、いずれの場合も多重検定は個別検定の繰り返し適用と比較して、FWER を適切にコントロールしつつ、より高い検出力を達成していることが確認できる。

5.2 実証例

実証例として、OECD の 13 ヵ国について、失業率の持続性（persistence）の確認を行う。データは 1996 年第 1 四半期から 2013 年第 3 四半期であり、多くのデータについて 2008 年に発生した

表 2 a　個別検定と多重検定の FWER と検出力

T	N	FWER		検出力	
		個別検定	多重検定	個別検定	多重検定
50	5	0.154	0.078	0.302	0.750
	10	0.278	0.053	0.304	0.572
100	5	0.172	0.089	0.724	0.987
	10	0.290	0.074	0.713	0.979
200	5	0.180	0.088	0.941	1.000
	10	0.346	0.082	0.940	1.000

（注）帰無仮説、対立仮説の下でのモデルは、$y_{i,t} = y_{i,t-1} + u_{i,t}(under\ H_0)$、$y_{i,t} = 0.85y_{i,t-1} + u_{i,t}(under\ H_1)$。ここで、$u_{i,t} = \phi(L)\varepsilon_{i,t}$ $(\phi(L) = 1 + \phi L + \phi^2 L^2 + \cdots + \phi^{t-1}L^{t-1})$ であり、$\varepsilon_t \sim N(0, \Omega)$ $(\varepsilon_t = (\varepsilon_{i,t}, \cdots, \varepsilon_{N,t})')$、$\Omega = \begin{pmatrix} 1 & 0.3 & \cdots & 0.3 \\ 0.3 & 1 & & \vdots \\ \vdots & & \ddots & 0.3 \\ 0.3 & \cdots & 0.3 & 1 \end{pmatrix}$。回帰式は $\Delta y_{i,t} = \hat{\phi}_i y_{i,t-1} + \sum_{l=1}^{\overline{l}_i} \hat{\alpha}_{i,l}\Delta y_{i,t-1} + error$ である。ラグ次数 $\overline{l}_i$ は modified AIC で決定。

表 2 b　個別検定と多重検定の FWER と検出力（構造変化がある場合）

T	N	FWER		検出力					
				$\delta = 1$		$\delta = 3$		$\delta = 5$	
		個別検定	多重検定	個別検定	多重検定	個別検定	多重検定	個別検定	多重検定
50	5	0.133	0.106	0.095	0.487	0.241	0.707	0.528	0.902
	10	0.267	0.065	0.104	0.308	0.244	0.605	0.529	0.859
100	5	0.133	0.150	0.247	0.939	0.421	0.972	0.679	0.989
	10	0.246	0.087	0.244	0.909	0.417	0.964	0.678	0.987
200	5	0.145	0.163	0.683	1.000	0.789	1.000	0.900	1.000
	10	0.259	0.072	0.680	0.999	0.784	1.000	0.898	1.000

（注）帰無仮説の下でのモデル及び $u_{i,t}$ は表 2 a と同じ。対立仮説の下でのモデルは、$y_{i,t} = 0.85y_{i,t-1} + \delta DU_t(\tau_i) + u_{i,t}$。ここで、$DU_t(\tau_i)$ は $DU_t(\tau_i) = 1$ for $t > \tau_i T$ or zero otherwise である（$\tau_i = TB_i/T = 0.5$ とする）。回帰式は $\Delta y_{i,t} = \hat{\phi} y_{i,t-1} + \hat{\delta}_i DU_t(\hat{\tau}_i) + \sum_{l=1}^{\overline{l}_i} \hat{\alpha}_{i,l}\Delta y_{i,t-1} + error$ である。τ_i は、$0.35 < \tau_i < 0.65$ の範囲において $\hat{\phi}_i$ の t 統計量を最小にする時点で決定される。ラグ次数 $\overline{l}_i$ は modified AIC で決定。

表 3　OECD 13 ヵ国の失業率の定常性

Country	Zivot and Andrews（1992）統計量	個別検定[a]	多重検定[b]
Australia	−3.152		
Austria	−2.903		
Canada	−2.622		
Chile	−2.616		
Denmark	−4.197		＋＋
Germany	−1.556		
Greece	−2.053		
Ireland	−6.483	***	＋＋
Japan	−2.209		
Korea	−6.382	***	＋＋
Mexico	−2.614		
Zew Zealand	−3.308		
US	−3.473		

a　***，**，*は、それぞれ 1%，5%，10% の有意水準の下で有意であることを表す。臨界値は Zivot and Andrews（1992）による。

b　＋＋，＋は、それぞれ 5%，10% の familywise error rate の下で有意であることを表す。

リーマン・ショックの影響があると思われるため、1 回の構造変化を持つ単位根検定（Zivot and Andrews（1992）、定数項を含む場合）を用いて、その繰り返し適用と Zivot and Andrews 統計量を用いた多重検定を行う。

検定結果を表 3 に示す。結果より、個別検定においては Ireland と Korea の 2 ヵ国の棄却であるが、多重検定において、さらに Denmark も有意に棄却できている。先の表 2 b の検定パフォーマンスより、構造変化がある場合でも多重検定は FWER を個別検定よりも適切にコントロールし、かつ検出力が高いとの示唆を得たが、この結果はそれを裏付けるものであろう。

6. おわりに

本稿では、実証分析におけるパネル単位根検定の利用について、分析者が直面する結果解釈上の根本的な問題を取り上げた。そして、その問題の解決方法の 1 つとして、多重検定の適用を考察した。特に、検定における FWER のコントロールと検出力について、個別検定の繰り返し適用との比較を行い、多重検定のパフォーマンスの良さが明らかとなった。また、実証例においてもそれが裏付けられた。

しかしながら、多重検定の非定常パネルデータを用いた研究への応用は、まだ端緒に就いたばかりであり、研究の蓄積は現在のところ極めて少ない。そのため、今後多くの改善や発展が行われる必要がある。

特に、よく知られた多重検定の欠点として、検定する仮説数が多くなる時、帰無仮説を棄却し難くなるという性質がある。現在、その改善のための試みがいくつかなされている。その 1 つが、FWER を緩和する方向での改良である。それは一般化 FWER（k-FWER）（Dudoit et al.（2004）、Romano et al.（2008））を用いた多重検定であり、k-FWER とは帰無仮説を少なくとも誤って k 回棄却する確率である。2-FWER や 3-FWER を考えることで、仮説数の増加に伴う検出力の低下を補うことができる。しかしながら、実証分析の現場においては、仮説数に応じた k の最適値がいくらであるのかが不明であり、恣意的に k を選ぶことは好ましくないため、この値の選択基準が決まらないと実際の利用は難しい。このテーマは今後なされるべき研究課題の 1 つであろう。

参考文献

Bai, J. and Ng, S.（2004）“A panic attack on unit roots and cointegration,” *Econometrica,* vol.72, pp.1127-1177.

Benjamini, Y. and Hochberg, J. C.（1995）“Controlling the false discovery rate : A practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society, Series* B, vol.57, pp.289-300.

Breitung, J.（2000）“The local power of some unit root tests for panel data,” *Advances in Econometrics,* vol.15, pp 161-178.

Breitung, J. and Das, S.（2005）“Panel unit root tests under cross sectional dependence,” *Statistica,* vol.59, pp.414-433.

Chortareas, G. and Kapetanios, G.（2009）“Getting PPP right : identifying mean-reverting real exchange rates in panels,” *Journal of Banking and Finance,* vol.33, pp.390-404.

Choi, I.（2001）“Unit Root Tests for Panel Data,” *Journal of International Money and Finance,* vol.20, pp.249-272.

Dudoit, S., van der Laan, M. J. and Pollard, K. S.（2004）“Multiple testing. Part I. Single-step procedures for

control of general type I error rates,” *Statistical Applications in Genetics and Molecular Biology,* vol.3.

Fisher, R. A.（1932）Statistical methods for research workers, 4th edn. Oliver&Boyd.

Hadri, K.（2000）“Testing for Stationarity in Heterogeneous Panel Data,” *Econometrics Journal,* vol.3, pp.148-161.

Harris, R. D. F. and Tzavalis, E.（1999）“Inference for unit roots in dynamic panels where the time dimension is fixed,” *Journal of Econometrics,* vol.91, pp.201-226.

Hanck, C.（2009）“For which countries did PPP hold? A multiple testing approach,” *Empirical Economics,* vol.37, pp.93-103.

Hochberg, Y. and Tamhane, A. C.（1987）*Multiple Comparison Procedures.* New York : John Wiley & Sons, 1987.

Holm, S.（1979）“A simple sequentially rejective multiple test procedure,” *Scandinavian Journal of Statistics,* vol.6, pp.65-70.

Im, K.-S., Pesaran, H. and Shin, Y.（2003）“Testing for unit roots in heterogeneous panels,” *Journal of Econometrics,* vol.115, pp.53-74.

Levin, A., Lin, C.-F. and Chu, C.（2002）“Unit root test in panel data : asymptotic and finite sample results,” *Journal of Econometrics,* vol.108, pp.1-24.

Maddala, G. S., and Wu, S.（1999）“A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test,” *Oxford Bulletin of Economics and Statistics,* Special Issue, pp.631-652.

Matsuki, T. and Sugimoto, K.（2013）“Stationarity of Asian real exchange rates : An empirical application of multiple testing to nonstationary panels with a structural break,” *Economic Modelling,* vol.34, pp.52-58.

Matsuki, T.（2016）“Linear and nonlinear comovement in Southeast Asian local currency bond markets : a stepwise multiple testing approach,” *Empirical Economics,* vol.51, pp.591-619.

Moon, H. R. and Perron, B.（2004）“Testing for a unit root in panels with dynamic factors,” *Journal of Econometrics,* vol.122, pp.81-126.

Moon, H. R. and Perron, B.（2012）“Beyond panel unit root tests : Using multiple testing to determine the nonstationarity properties of individual series in a panel,” *Journal of Econometrics,* vol.169, pp.29-33.

Nabeya, S.（1999）“Asymptotic Moments of Some Unit Root Test Statistics in the Null Case,” *Econometric Theory,* vol.15, pp.139-149.

Pesaran, M. H.（2007）“A simple panel unit root test in the presence of cross section dependence,” *Journal of Applied Econometrics,* vol.22, pp.265-312.

Pesaran, M. H.（2014）“On the interpretation of panel unit root tests,” *Economics Letters,* vol.116, pp.545-546.

Phillips, P. C. B.（1987）“Time Series Regression with a Unit Root,” *Econometrica,* vol.55, no.2, pp.277-230.

Romano, J. P. and Wolf, M.（2005）“Stepwise multiple testing as formalized data snooping,” *Econometrica,* vol.73, pp.1237-1282.

Romano, J. P., Shaikh, A. M. and Wolf, M.（2008）“Formalized data snooping based on generalized error rates,” *Econometric Theory,* vol.24, pp.404-447.

Simes, R. J.（1986）“An improved Bonferroni procedure for multiple tests of significance,” *Biometrika,* vol.73, pp.751-754.

Smeekes, S.,（2010）Bootstrap sequential tests to determine the stationary units in a panel. METEOR Research Memorandum 11-003.

Tamhane, A. C.（1996）“Multiple Comparisons,” *Handbook of Statistics,* vol.13, pp.587-630.

Zivot, E. and Andrews, D.W.K.（1992）“Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis,” *Journal of Business & Economic Statistics,* vol.10, pp.251-270.