第２講　クロス･セクション分析の基礎

時間を一時点に固定して止め、その時点で区切って各地点、場所、グループで起こっているデータを記録したものを**クロス･セクション（横断面）データ (cross section data)** という。これに対し、時間に従って取ったデータを**タイムシリーズ（時系列）データ (time series data)** という。時系列データには時間の要素が入るが、クロス・セクション・データには場所空間の要素が入る。一般的には、時系列データの方が、クロス・セクション・データより記録するコストは低いとされている。

官庁統計の多くは、クロス・セクション・データであるが、それを一定の間隔で行っており、その意味では時系列的要素もある。この両方の要素を意図的に取り入れたのが**パネル・データ (panel data)** といわれるものである。すなわち、同一主体（個人、家計、企業等）を複数年に亘って調査し続けることで、ある政策の効果や外生的なショックに対して経済主体がどのように反応するかをより厳密に観察できる。このようなパネル・データの収集は、わが国では最近始まったばかりであり、十分に利用されているとは言えないが、企業データを中心に現在のクロス・セクション・データを時系列方向に接続（マッチング）してパネル・データを作成するという試みも始まっている。

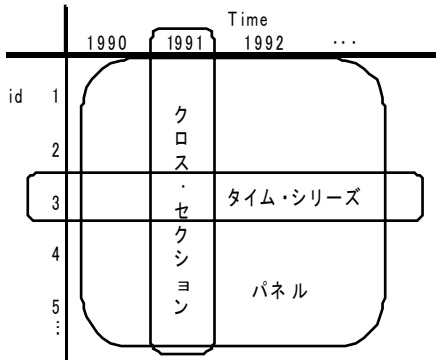
パネル・データの性格とその利用法、統計的手法については昨年の比較統計システム論で論じたので、今年は主として単一年度のクロス・セクション・データについて論じたい。もちろんクロス・セクション・データとパネル・データは密接に関係があり、必要に応じてパネル・データの話もする。パネル・データについて関心のある方は、私の個人ホームページの中に講義録を載せるのでそれを参照していただきたい。

パネル・データのように同一主体を複数年に亘って追跡したものではないが、同じ生年でカテゴリー化したデータに**コーホート・データ (cohort data)** がある。先に述べたようにパネル・データの蓄積はまだ少ないので、既存のクロス・セクション・データからコーホート・データを作り、時系列や単体のクロス・セクション・データでは分からなかった**コーホート効果**を調べることができる。

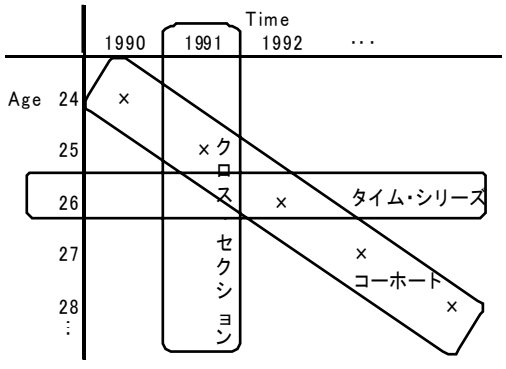
複数年のクロス･セクション・データを調査年の区別なく、全てプールして用いるデータを**プールド・データ (pooled data)** という。この方法では観察値が増え、推定量が安定化することもある。

	個票データ micro data	集計データ aggregate data	擬似データ
クロス・セクション	調査世帯数	一時点の調査を集計したもの 地域、年齢、職業	
タイム・シリーズ	調査時点数	調査時点数	
パネル	同一世帯×調査時変数		(例)都道府県別消費(47)の年次推移(10年)=470
コーホート	同一世帯×調査時変数		集計データの年齢別データを複数年に亘りコーホートに整理する

パネル・データのイメージ



コーホート・データのイメージ



実証研究を行う場合に注意すべき点

1. 何を調べたいのか、実証的に面白い問題なのか。
2. どのような経済理論を用いるのか。
3. どのような統計データが利用可能か、それは経済理論と整合的か。多くの場合、理論は経済主体の異時点間での最適化問題から導かれたものを用い、実証ではパネル・データがないので、集計した時系列データを用いることが多いが、これは厳密には誤りである。
4. 逆に、クロス・セクション・データを用いる場合に、適切な経済理論とは何だろうか。一般的には各経済主体（X_{id}）がそれぞれ何らかの最適化を行って、観察される値は**均衡値**であると考える。そして各主体間での違いは経済属性の違いによって説明されるという考え方である。とりわけ、それが**切片**の違いとして表れる場合には**ダミー変数**が用いられる。
5. ４では最適化そのものについての特定化はしなかった。異時点間の最適化問題を経済主体が解いていると考えても、クロス・セクション・データを用いることは可能である。
6. ４のモデルはマクロ経済学で用いられる代表的個人、代表的企業のモデルであってはならない。ミクロ統計を用いるということはマクロの平均的行動を分析するというより、経済主体間の差異や分配の問題を分析することに主眼があることに注意すべきである。

データに関する問題

1. **無回答**

   ミクロ統計データはマクロデータと違い、個別の調査票の回答に基づいて作成されたものであり、当然ながら無回答の調査項目も含まれる。その無回答項目が分析にとって中心的関心事であれば、この問題は深刻である。一般には無回答の主体をサンプルから外す。

2. **外れ値**

   研究にとって中心的関心事である変数が全サンプルから見て異常な値をとる場合がある。入力ミスでないとすれば、何らかの経済学的な意味を持つケースには、外れ値をとる主体が存在することを十分に意識して分析する必要がある。外れ値を含んだまま平均や分散を算出すると大きなバイアスを生じる。外れ値をとるような極端な経済行動をとる主体に関心があれば、そのデータは残すべきだが、全サンプルの分布等に関心がある場合は $\mu \pm 4\sigma$ の外に出るサンプルを除外するなど外れ値の調整が必要になる。

3. **入力ミス：重複、桁違い**

   近年ではミクロ統計の入り口は機会であるようになったが、過去のデータでは手入力の結果生じたと思われる入力ミスが見出された (重複や桁違い等)。しかし機械に入力する前の個票レベルでの入力ミスについてはいまだに残されており、各データ、変数の分布等をていねいに調べて入力ミスと思われるデータについては改正するなり、除外する必要がある。

4. **データ・マッチング**

   ミクロ統計データを独立して個別に使う場合には問題はないが、複数年のミクロ統計データを合併して、パネル・データやコーホート・データを作

る場合には、データのマッチングをする必要がある。各年の調査が同じことを繰り返していれば問題は少ないが、調査対象となる主体が変わり、調査項目自体が変更されることも多々あるので、マッチングは予想以上に困難な作業となる。