パネル講演

パネルデータの長所とその分析方法†

――常識の誤りについて

山口 一男

（シカゴ大学社会学部教授）

1．はじめに

パネルデータ分析法は学問的には学際領域ですが、実際には社会科学の中でも専門によって特徴を持ちながら発展してきました。例えば計量経済学ではfixed-effects modelsや random effects modelsや selection models（Hsiao 1986; Baltagi 1995）などのモデルを中心に特に説明変数の状態への選択バイアスの問題の取り扱いや因果分析の点で大きく貢献してきました。一方計量心理学では因子分析から発展したconfirmatory factor-analytical modelsや linear growth modelsや人と時間を二つのレベルとするmultilevel modelsやlatent Markov modelsなど、変数の不完全な計測を前提とする潜在変数を用いたモデルのパネルデータ分析への応用の発展に貢献してきました。計量社会学ではGoodman等の伝統により対数線形モデルや潜在クラスモデルなどのカテゴリカルデータのモデルの発展が特徴で、著者は統計学者ですが、例えばHagenaars（1990）などにその成果を見ることができます。また生命統計学・計量経済学・計量社会学での共通領域の方法であるイベントヒストリー分析もパネルデータに基づくものが多く、従ってこれもパネルデータ分析手法の一つといえるでしょう。

このような様々な分析方法を包括的に議論するのは時間的制約上不可能なので、今日はむしろパネルデータについて通常信じられている「常識」の誤りを幾つか指摘することでパネルデータ分析の「深さ」に触れてみたいと思います。

まずパネルデータに関する幾つかの「神話」もしくは「誤った常識」、あるいは誤りとはいえないまでも「広範に存在する不十分な理解」について以下が考えられます。

神話１．パネルデータはマクロな時系列的変化を分析するのに優れている。

神話２．パネルデータは例えば転職などのイベントXが収入などの従属変数Yの変化にどう影響するかを見るのに適しているが、これはパネル調査をしなくても一回調査で転職者について前職の収入を調査すれば同様の変化の情報が得られる。

神話３．XのYへの影響について、例えば離婚（X）が健康（Y）に与える影響について、離婚が平均的に見て健康のレベルを減少させるとき、それが離婚が健康な人が健康を損なう傾向を増大させることからくるのか、それとも離婚が健康状態の良くない人が健康を回復する傾向を減少させることからくるのかは、回帰分析では区別できない。

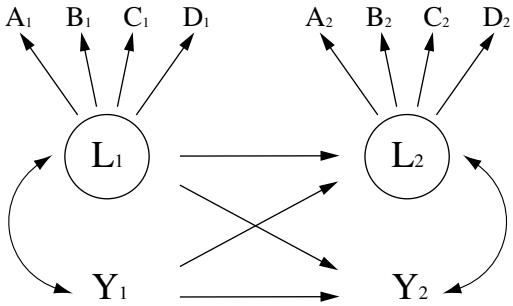
神話４．態度や意識Yの安定性が例えば教育レベルなどのXに依存するか否かはY_tの予測においてY_{t-1}とXの交互作用効果を見ればよい。

神話５．時間とともに変化するイベント（例えば転職や結婚など）Xの（収入や健康など）Yへの影響を見るとき、選択バイアスとは、Xを経験したグループと経験しなかったグループについて、Xの経験以前に２つのグループ間に存在していた個人差により生じるYのグループ差のことをいう。

神話６－１．パネルデータによる回帰分析でY_tの予測を、Y_{t-1}を制御して（説明変数に加えて）行えば、Yの変化の予測の分析をすることになる。

神話６－２．パネルデータでY（例えば収入や健康）の変動について、２時点間で起こったイベ

図表-1　二変数間の相互影響のモデル



ントX（転職、離婚など）の影響を計るのに、イベントを経験することになる者と経験しない者との間にすでに時点t-1で存在していた個人差の影響を排除してXの影響を見るには、Y_tを予測する回帰分析でY_{t-1}を制御して（Y_{t-1}をY_tの説明変数に加えて）Xの影響を見ればよい。

神話７．$Y_t - Y_{t-1}$を従属変数とする回帰分析は観察されない個人の異質性を制御できる点で長所があるが、Y_tのY_{t-1}からの独立を仮定する上に「平均値への回帰（regression to the mean）」の問題もあるので利用すべきでない。

２．パネル調査の長所について

まずパネル調査の長所に関する神話の１ですが、マクロな時系列分析なら独立な標本の繰り返し調査（repeated cross-sectional survey、以下RCS調査）のほうがパネル調査より優れています。なぜなら（1）パネルデータでは年代とともに年齢分布が移動してしまうし、（2）標本脱落（sample attrition）も起こる、という理由で各時点で同等な母集団を代表することがRCS調査に比べてより困難だからです。実際パネルデータでは標本脱落を補塡したり、若いコーホートの標本を随時に追加していかなければ母集団の同等性や代表性に問題が起き、その点コストも手間も余計にかかります。年齢効果と時代効果の区別もパネルデータでなくともRCS調査データで効率的にできます。ただし従属変数Yの年齢による変化の出生世代間の比較については、同じ人々に起こった変化を分析できるのでパネルデータの方がより精度が高く優れています。

パネル調査の真の利点は、マクロの時系列変化そのものでなく、変化をミクロな個人のレベルで捉えることができる点にあります。このことは、（1）一回調査の大きな問題の一つである従属変数Yに対する個人に起こる変化X（結婚、離婚、就業、転職、失業など）の影響を個人間の違いから説明するという誤りを犯さないですむ手段を与え（具体例として樋口（2001）における「転職コスト」の分析を参照）、（2）パネルデータが個人レベルで起こるXのYへの影響に対し、その因果関係の有無の判定とその程度の測定に格段の進展をもたらすことを示し（Winship and Morgan（1999）を参照）、（3）マクロな変化についてどの程度個人に起こった変化から説明できるかを分析することを可能にする点にあります。

またその他の利点として共変動する２変数YとZで、どちらがどちらに影響を与えているかは、少なくとも２時点の観察値があれば、相互的影響の同時推定モデルでcross-variable, time-lagged effectsを調べることで分析ができます。例えば、カテゴリカルデータ分析の例ですが、既婚女性の従業上の地位Y（フルタイム、パートタイム、無職の別）と幾つかの指標（A, B, C, D）で潜在クラスが特徴づけられる性別役割態度Lの間で、どちらがどちらに主に影響するかは図表-１のようなモデルに対し例えば*lem*という統計プログラムを用いてパネルデータからL_1のY_2に対する影響とY_1のL_2に対する影響を同時に推定することで判定できます。図表-１では省きましたが外生変数ももちろん制御できます。この例のように態度や意識や心理の変数を含む共変動の分析は一回調査で過去の状態を回顧して調査することは信頼できないので、パネルデータが必要となります。

また他の利点として態度や意識などの変数もその持続性や安定性に個人差があるのですが、パネル調査ではその情報が得られます。例えばColeman（1964）は、いまだパネルデータがアメリカでも少なかった時期に世論調査の結果70%が

現政権を支持しているとわかった時、それは全体の70%の人が確率１で支持し30%の人が確率１で支持しないことを意味するのか、それとも各人が確率70%で支持し確率30%で支持しないことを意味するのか、あるいはその混合か、の区別が重要で、その区別はパネルデータによってできると考えました。実際には支持確率にも態度の安定度にも個人間に異質性があり、Colemanの２分法的区別よりは複雑ですが、基本的には彼の考えた通り、より態度の安定した人々に支持されているほうが、より安定していない人々に支持されるより政権基盤がより強固であるといえ、パネル調査では態度や意識の安定度についての個人差の情報が得られることが変化の予測に対して極めて有利であるといえます。

神話の２について、一回調査で転職者について前職の収入などを調査すればパネルデータと同様の変化の適切な情報が得られるかどうかという点ですが、これは得られません。パネル調査における各時点での状態は人と時点の組み合わせを単位として標本抽出しています。そのため各時点では継続年数の長い職ほど「現職」になりやすく、「現職」は（これは他の現在の状態でも全く同じですが）継続年数に比例して抽出されます。一回調査でも現職についてはこの点は同じです。しかし、一回調査で転職経験者から前職についての情報を得ると「前職」は継続年数に比例して抽出されません。継続時間が短くても長くても、長さに関係なく「前職」となるからです。従って「現職」と「前職」の比較（例えば収入の変化）は、職の標本抽出の仕方が「現職」と「前職」で違うので、例えば長く就業する職ほど平均的に年間収入などYは高くなるなどの傾向があるとYの現職と前職の差の推定にバイアスをもたらします。従って正しい比較方法は共に時点を単位として、例えば現職と一年前の職について、過去一年での転職経験者と継続就業者を比べる方法です。この比較のためのデータは一回調査でも回顧によって得られますが、正確さの点でパネル調査のほうが明らかに優れています。

3. XとY_{t-1}との交互作用効果の利用について

神話の３と４はともにXとY_{t-1}との交互作用効果の利用に関係しています。神話の３はXのYの変化への影響について動きの方向を区別して回帰分析で推定できるか否かという点ですが、これは線形のYの場合はわかりにくいけれどもYを順序のついたカテゴリーで表し、ロジスティック回帰（二分法の時）か累積ロジット（cumulative logit）回帰分析で行えば容易に区別ができます。いまYが値０と１をとる場合は$P_t=Prob(Y_t=1)$とし、

$$\log(P_t/(1-P_t)) = a+bY_{t-1}+\sum_{i=1}^{I} c_i X_{it}+\sum_{i=1}^{I} d_i Y_{t-1} X_{it}+ \sum_{j=1}^{J} e_j Z_{jt}$$

とすると、$Y_{t-1}=0$の場合はYの状態が時点tで０から1へ変化する確率の対数オッズへのX_{it}の効果はc_iとなりZ_{jt}の効果はe_jとなります。また、$Y_{t-1}=1$の場合Yの状態が時点tで１から０へ変化する確率の対数オッズへのX_{it}の効果は$-(c_i+d_i)$となりZ_{jt}の効果は$-e_j$となります。したがってY_{t-1}との交互作用がない変数の両方向への影響は向きが逆で程度が同じ（例えば離婚は健康を損なわせるのと同程度に健康回復を妨げる）になりますが、交互作用を入れることによって非対称的効果を測定できます。

より一般的な場合も同様でいまYが1からnまでの順序のついたカテゴリー値で$P_{k,t}=Prob(Y_t=k)$とし、$k>1$に対し$Q_{k,t}=\sum_{j=k}^{n} P_{j,t}$, $1-Q_{k,t}=\sum_{j=1}^{k-1} P_{j,t}$と定義します。また$Y^*$を

$$Y^*_{k,t-1}\begin{cases}=1 \text{ if } Y_{t-1}\geq k\\ =0 \text{ if } Y_{t-1}\leq k-1\end{cases}$$

で定義されるダミー変数とします。いま

$$\log(Q_{k,t}/(1-Q_{k,t})) = a_k+bY^*_{k,t-1}+\sum_{i=1}^{I} c_i X_{it}+\sum_{i=1}^{I} d_i Y^*_{k,t-1} X_{it}+\sum_{i=1}^{J} e_j Z_{jt}$$

というモデルを考えると、$Y^*_{k,t-1}=0$の場合Yの状態が時点tでk未満のレベルからk以上のレベルへ変化する確率の対数オッズへのX_{it}の効果はc_iとなりZ_{jt}の効果はe_jとなります。また、$Y^*_{k,t-1}=1$の場合Yの状態が時点tでk以上のレベルからk未満のレベ

ルへ変化する確率の対数オッズへのX_{it}の効果は$-(c_i+d_i)$となりZ_{jt}の効果は$-e_j$となります。したがって一般の場合も$Y^*_{k,t-1}$との交互作用をモデルに入れることによって同様に非対称的効果を測定できます。

一般にXがYの上方方向の動きを促進（減少）させるのか、下方方向の動きを減少（促進）させるのかは、単にXがYの変化に正または負に影響するという以上に、一歩変化のメカニズムに踏み込んでいるだけ理論的に重要です。パネルデータの利用によりこういった分析が可能になります。

神話の4は、XとY_{t-1}との交互作用効果のもう一つの側面について、例えば態度や意識Yの安定性がXに依存するか否かはY_tの予測においてY_{t-1}とXの交互作用効果を見ればよいという理解でよいかという点ですが、態度や意識の安定性はY_{t-1}のY_tへの影響の程度の異質性だけでは適切に計れない可能性が大と考えます。詳しくは述べませんが個人の態度や意識の安定度を潜在変数Zで表し、Yの回帰モデルとZの回帰モデルの同時モデルを応用するといった分析が考えられます（例えば山口（2002）を参照）。

4. 因果分析に対するパネルデータの貢献について

神話の5、6、7は統計的因果分析にパネルデータが何をなし得るかという問題に関連しています。まず神話の5は、選択バイアスについて、Xを経験したグループと経験しなかったグループについて、Xの経験以前に2つのグループ間に存在していた違いによりYについて生じる差、という理解でよいかという点ですが、この理解は誤りとはいえないが不十分な点があります。

実際の社会経済変数の標本調査データは制御された心理実験とは異なるのですが、概念として同様の言葉を用いることとし、X（例えば転職や離婚）を経験したグループをトリートメントグループ（略称Tグループ）、Xを経験しなかったグループをコントロールグループ（略称Cグループ）と呼ぶことにします。いま、

$\overline{Y}_{T|X=1}$をTグループがXを経験したときのYの平均値、

$\overline{Y}_{T|X=0}$をTグループがXを経験しなかったなら、実現したであろうYの平均値、

$\overline{Y}_{C|X=1}$をCグループがXを経験したなら、実現したであろうYの平均値、

$\overline{Y}_{C|X=0}$をCグループがXを経験しなかったときのYの平均値

とします。実際に観察されるのは$\overline{Y}_{T|X=1}$と$\overline{Y}_{C|X=0}$の2値で、あとの二つの$\overline{Y}_{T|X=0}$と$\overline{Y}_{C|X=1}$は**事実と反する**（**counterfactualな**）仮定での、観察されない値です。また$\delta_T \equiv \overline{Y}_{T|X=1}-\overline{Y}_{T|X=0}$、$\delta_C \equiv \overline{Y}_{C|X=1}-\overline{Y}_{C|X=0}$と定義します。$\delta_T$と$\delta_C$はそれぞれTグループとCグループにおけるXの因果的な効果と考えられます。またTグループの割合をπとし、$\overline{\delta} \equiv \pi\delta_T+(1-\pi)\delta_C$とします。$\overline{\delta}$は全体における$X$の因果的効果の平均値です。すると観察される2値の差、$\overline{Y}_{T|X=1}-\overline{Y}_{C|X=0}$と$X$の因果的効果$\overline{\delta}$とのくい違いについて次の式が成り立ちます。

$$(\overline{Y}_{T|X=1}-\overline{Y}_{C|X=0})-\overline{\delta}=(\overline{Y}_{T|X=0}-\overline{Y}_{C|X=0})+(1-\pi)(\delta_T-\delta_C)$$

この式はXの選択バイアスにはXを経験したグループと経験しなかったグループがXの経験以前に持っていたであろう違いの影響、$\overline{Y}_{T|X=0}-\overline{Y}_{C|X=0}$の他に$X$の影響自体がTグループとCグループで異なることから生じるバイアス、$(1-\pi)(\delta_T-\delta_C)$、が存在しうるということを示しています。

重要なことは 実際にパネルデータを用いて推定できるのは、δ_Tであって$\overline{\delta}$でないことが多いということです。このことは例えば離婚や転職がそれを実際に経験した者に不利益をもたらしたかということの過去の評価に答えは得られても、離婚や転職はもし経験すれば不利益をもたらすかどうかという、いまだ実現せずかつTグループへの選択メカニズムが異なる場合への答えはデータからは得にくいことを意味します。また同様の理由で統計的因果分析は公共の職業訓練の拡大など実際に行われた政策が成功したかどうかを評価できるけれども、これから採用される、特に政策に影響される人々の選択メカニズムが異なるような政

策が成功するかどうかは判断できないことが多いということをも意味しています。

神話の６－１についてパネルデータによる回帰分析でY_tの予測をY_{t-1}を制御して（説明変数に加えて）行えば、Yの変化の予測の分析をすることになるか否か、という点ですが、通常そういう理解があるように思うのですが、これは完全に誤りです。より厳密に述べた神話６－２も同様に誤りです。

いま説明の簡単化のためXをダミー変数と仮定します。すると

$$y_t = a + by_{t-1} + cx_t + \varepsilon$$

の回帰式にOLS法を当てはめると、Xの効果cについて

$$c = \overline{Y}_{t,T} - \overline{Y}_{t,C} - b\,(\overline{Y}_{t-1,T} - \overline{Y}_{t-1,C})$$

が成り立ちます。ここでTとCはそれぞれX=1およびX=0グループを意味し、$\overline{Y}_t$と$\overline{Y}_{t-1}$はそれぞれのグループ内でのYの時点tと時点t-1での標本平均です。

この解に見られるXの効果cは「すでに時点t-1で存在している個人差の影響を排除してXの影響を見る」というのが$(\overline{Y}_{t,T} - \overline{Y}_{t,C}) - (\overline{Y}_{t-1,T} - \overline{Y}_{t-1,C})$という数量で計測されると考えるなら、本来関係のないY_{t-1}の回帰係数bが関わっていることが、注目すべき点です。実際、特殊な場合として「事後」のグループ間格差が「事前」のグループ間格差に等しい場合、

つまり$\overline{Y}_{t,T} - \overline{Y}_{t,C} = \overline{Y}_{t-1,T} - \overline{Y}_{t-1,C} = D \neq 0$の場合でも（つまり$Y$の変化がTグループとCグループで平行線となり明らかにXの効果がないと考えられるときでも）、$c = D(1 - b)$となり回帰係数bが１でない限りXの効果があるという結論を得てしまいます（これはLord's Paradoxと呼ばれています）。さらに極端な場合Y_{t-1}のY_tへの影響がもし0であれば、時点tでのYのグループ間格差は、時点t-1でのYのグループ間格差にかかわらず、すべてXの影響の結果となってしまいます。ではこれらの結果はなぜ起こるのでしょうか？

実は「Y_tの予測にY_{t-1}を制御する」ということは「Yの変化を説明しようとする」ことでは全くなく「時点t-1でのYの個人差の影響を除外すること」とも異なるのです。Y_{t-1}を制御することは「時点t-1におけるYの違いがあり、かつY_{t-1}がY_tに影響するという、その二つのことの組み合わせから起こる、**時点t-1のYの差が時点tのYの差として生じる持ちこみ効果**を除外する」ことを意味します。ですからY_{t-1}がY_tに全く影響しなければ、いくら時点t-1でYに差があろうと時点tに持ち込まれる差は０となり、したがって時点tでYに差があれば、別の理由に帰せられるわけです。一般に係数bが小さくなればなるほど、持ち込まれる効果は少なくなり、その分がXの効果に帰せられるため、$c = (\overline{Y}_{t,T} - \overline{Y}_{t,C}) - b\,(\overline{Y}_{t-1,T} - \overline{Y}_{t-1,C})$というような結果を生じるわけです。

一方Yの変化を説明したい場合には変化スコア$Y_t - Y_{t-1}$を従属変数としYの差分を分析することが考えられます。実際Xがダミー変数のとき、

回帰式$y_t - y_{t-1} = a + b\,(x_t - x_{t-1}) + \varepsilon$からは

$b = (\overline{Y}_{t,T} - \overline{Y}_{t,C}) - (\overline{Y}_{t-1,T} - \overline{Y}_{t-1,C})$という、$Y$の時点$t$-1での差の影響を完全に除外した$X$の効果の推定値が得られます。しかし$Y$の差分を分析する方法には別の面で制限があり、それに後述の「神話７」に関係します。

また「時点t-1で存在している個人差の影響を排除してXの影響を見る」という表現には暗黙のうちに「時点t-1とtの間でXの変化を経験することになるグループとXの変化を経験しないことになるグループ間の事前の差」の制御という含意があります。すなわち「観察されない（あるいは制御されない）個人差」があり、それがY_{t-1}に影響を与えている場合、その影響も含めて排除するという含意ですが、単にY_{t-1}を独立変数として制御したのでは、Yの持ちこみ効果だけの除外となりますから、そういったより一般的な個人差の影響は全く除外できません。より一般的な個人差の影響の除外の問題は、以下に述べるXの状態への選択バイアスと、それを取り除こうとするfixed effects modelの利用に関係しています。

神話７については再掲すると「$Y_t - Y_{t-1}$を従属変数とする回帰分析は観察されない個人の異質性を制御できる点で長所があるが、Y_tのY_{t-1}からの独立を仮定する上に『平均値への回帰（regression

to the mean)』の問題もあるので利用すべきでない」という主張ですが、これは誤りです。

「平均値への回帰」というのはY_{t-1}が$\Delta_t = Y_t - Y_{t-1}$に影響を与えると「問題が起こる」という議論なのですが実際に問題なのはパラメーターの推定値の一致性（consistency）でそこで問題になるのはY_{t-1}のY_tへの影響であり、それがなければ、当然Y_{t-1}とΔ_tの相関係数は－1でY_{t-1}はΔ_tに強く影響しますが、これは全く問題が起きません。誤解は広く浸透しているようで、初等社会統計学の教科書出版では定評のあるSAGE出版の1995年のFinkelの著書*Causal Analysis with Panel Data*でも

"[The change score model] contains one highly restrictive assumption: that the lagged dependent variable Y_{t-1} does not have an influence either on Y_t or Δ_t. This assumption is likely to be incorrect, and for this reason, [it] usually fails as a structural model for analyzing change (page 6)."

などと、全く誤った記述をしています。

一般に変化スコア$Y_t - Y_{t-1}$についてはその平均値の差の検定は大いに使用すべきです。その長所は「時間とともに変化しない個人の異質性（fixed effectsという）」の影響を完全に除外できる点で、その点でYの差をXの差$X_t - X_{t-1}$の値の異なるグループ間で比較する（例えば収入の差の平均を転職者と職の継続者で比較する）ことはXの値の変化への選択バイアス（例えば収入の低いものが転職しやすい傾向）を取り除いて分析できることを意味し、極めて有効な手段です。Yの差の分散や分布の差を見ることも有用です。

$Y_t - Y_{t-1}$の平均を教育や年齢など、時間で変化しない変数や時間差が定数である変数Xの値について異なるグループ間で比較することも有用です。ただし、これはXのYへの影響を見るのでなく（その影響はfixed effectsで完全に除外されている）、Xの時間tとの交互作用X_tのYへの影響を見ていることになります。

Yがダミー変数のときは、Yの差の平均値の差のグループ間比較は、Y_{t-1}の値が0の場合の0から1への変化の割合と、Y_{t-1}の値が1の場合の1から0への変化の割合とは分けて比較をすべきです。両者をあわせた分析には通常はあまり意味がありません。

$Y_t - Y_{t-1}$について注意を要するのはそれを回帰分析の従属変数として用いるときで、Y_{t-1}がY_tに影響すると仮定するか否かに大きく依存します。

差分を用いる回帰分析はfixed-effects modelに現れます。Fixed effects modelは、いまY_{t-1}のY_tへの影響がないと仮定とするとYの回帰式は個人i時点tにおいて

$$y_{it} = a_t + c_i + \sum_k b_k x_{itk} + \varepsilon_{it}, t = 1,..,T; a_1 = 0 \qquad (1)$$

と表現できます。通常の回帰式と異なるのは、回帰切片が個人iによって変わる点で、標本数が1000なら1000異なるパラメーターc_iを切片の推定に用いることと同等です。このモデルの推定には、各個人内のT時点間の平均を引いた

$$y_{it} - \bar{y}_i = (a_t - \bar{a}) + \sum_k b_k (x_{itk} - \bar{x}_{itk}) + (\varepsilon_{it} - \bar{\varepsilon}_i) \qquad (2)$$

ただし、

$\bar{y}_i = (1/T)\sum_{t=1}^{T} y_{it}, \bar{a} = (1/T)\sum_{t=1}^{T} a_t, \bar{x}_{ik} = (1/T)\sum_{t=1}^{T} x_{itk}, \bar{\varepsilon}_i = (1/T)\sum_{t=1}^{T} \varepsilon_{it}$ 、

に対し式（1）の誤差項がi.i.d.条件を満たすならばOLSを当てはめて推定します（他の場合は通常GLSを用いるがこれについてはWooldridge (2002)を参照）。

Fixed effects modelの長所は、時間とともに変化する説明変数について（時間によって変化しない）個人差に基づく状態への選択バイアスを完全に取り除くことができる点です。したがってXの変化の影響は、**その状態変化の経験者の間でのXのYへの因果的影響**を表すと解釈できます。逆にfixed effects modelの短所は、Yの観察時に時間とともに変化しない変数Xの影響は測定不能であることです。このことはYの変化のリスクのある時間以外で変化するX（例えばYが所得の時、初職就業後変化しない教育程度）の因果的影響はfixed effects modelでは計れないことを意味します。

Y_{t-1}のY_tへの影響がある時は、fixed effects modelに問題が起こります。まず前提としてtime-

lagged effectsで、Y_tがY_{t-1}やY_{t-2}だけでなくY_{t-3}やY_{t-4}にも依存するなどの場合は、従属変数が離散的な変数の場合、状態継続時間依存をより一般的に取り扱うイベントヒストリーモデルを用いるべきです。（イベントヒストリーデータのパネル調査での効率的調査方法については山口（2001-2002）の『統計』連載のうち区間調査法によるイベントヒストリーデータの収集とその利用を参照）。仮にYが社会経済的地位のように連続変数でも時間的変化が連続的でなく変化をイベントとして捉えることができるならば、継続時間依存がある場合はイベントヒストリー分析が望ましくそれは可能です（Petersen（1988）を参照）。またイベントヒストリー分析におけるfixed effects modelの利用についてはYamaguchi（1986）を参照してください。以下の議論では Y_tの過去の状態への依存はY_{t-1}やY_{t-2}への依存として表現でき、より一般的なtime-lagged effectsはないと仮定します。

いま以下のモデルを仮定します。

$$y_{it} = a_t + c_i + \sum_k b_k x_{itk} + d y_{i(t-1)} + \varepsilon_{it}, t=2,..,T \qquad (3)$$

すると、この差分をとると

$$y_{it} - y_{i(t-1)} = a_t - a_{t-1} + \sum_k b_k (x_{itk} - x_{i(t-1)k}) + d(y_{i(t-1)} - y_{i(t-2)}) + \varepsilon_{it} - \varepsilon_{i(t-1)}, t=3,..,T \qquad (4)$$

となりますが、このモデルは$y_{i(t-1)} - y_{i(t-2)}$と$\varepsilon_{it} - \varepsilon_{i(t-1)}$が相関を持つため、OLSでもGLSでもパラメーターの推定値は一致性を持ちません。差分でなく個人の平均式との差をとっても結果は同様です。したがって通常の方法ではパラメーターを推定できないのです。しかしinstrumental variableを用いれば上記の差分の式のパラメーターの一致性を持つ推定値を得られます（Hsiao 1986）。特に$Y_t - Y_{t-1}$の回帰分析に$Y_{t-1} - Y_{t-2}$のinstrumentとしてY_{t-2}を用いる方法は利用に注意を払えば有効な方法と考えられます。

一般にinstrumental variableの利用には注意が必要です（Heckman 1997）。ここでの絶対条件は$y_{i(t-2)}$と$\varepsilon_{it} - \varepsilon_{i(t-1)}$との独立ですが、これは式(3)で誤差項が独立ならそう仮定されているものの、実際に仮定が成り立つか否かは別問題です。特にもし$y_{i(t-2)}$がy_{it}に直接影響を持つなら、$y_{i(t-2)}$とε_{it}が相関を持ちますが、この場合は元々の式（3）が誤りで、式（3）を

$$y_{it} = a_t + c_i + \sum_k b_k x_{itk} + d y_{i(t-1)} + e y_{i(t-2)} + \varepsilon_{it}, t=3,..,T \qquad (5)$$

と修正し、その差分の式

$$y_{it} - y_{i(t-1)} = a_t - a_{t-1} + \sum_k b_k (x_{itk} - x_{i(t-1)k}) + d(y_{i(t-1)} - y_{i(t-2)}) + e(y_{i(t-2)} - y_{i(t-3)}) + \varepsilon_{it} - \varepsilon_{i(t-1)}, t=4,..,T \qquad (6)$$

の説明変数$y_{i(t-1)} - y_{i(t-2)}$にinstrumentとして$y_{i(t-2)}$を用いればよいことになります[1]。またここで$y_{i(t-2)} - y_{i(t-3)}$は$\varepsilon_{it} - \varepsilon_{i(t-1)}$と独立と考えられるのでinstrumentはいりません。他に式(3)または式(5)で誤差項間にserial 相関がある場合も問題が起こりますが、fixed effectsを制御しているので時間に依存する説明変数で重要な変数が省略されていない限り誤差項間の独立の仮定は通常問題はないでしょう。

Instrumental variableを用いる方法の有効性には他にも条件があって一つは小標本でないことと、他はinstrumentが「強い」instrumentであることですが、後者については(a) Y_{t-1}とY_{t-2}の相関があまり高くなく（例えば0.7以下）かつ(b)Yの分散が時間 t とともに大きく変動しないこと、の2条件が満たされれば$y_{i(t-2)}$と$y_{i(t-1)} - y_{i(t-2)}$の共分散は安定的に負で、$y_{i(t-2)}$は強いinstrumentとみなしてよいと思います。Y_{t-1}とY_{t-2}の相関が高い場合はむしろfixed-effects modelでなくY_tの予測にY_{t-1}を説明変数として用いる通常のモデルが適切でしょう。以上のようにY_{t-1}がY_tに影響する場合のfixed effects modelには利用に注意が必要ですが、「不可能」では全くなく、利用条件が満たされれば強力な分析方法といえます。

Y_{t-1}がY_tに影響する場合fixed effects modelsは使いにくいという制約の理由は、より一般的にはいわゆるincidental parameter問題からくる最尤推定値の不一致性の問題から派生し[2]、問題はやや

異なりますが、従属変数が離散的な場合にも生じます。そうでない場合（Y_{t1}がY_tに影響しない場合）は条件最尤推定法を用いて2値をとるYにfixed effects logit modelを用いるChamberlain（1980）の方法や、事象件数データのポワソン回帰分析にfixed-effect modelを用いる方法（Cameron & Trivedi 1998）などが開発されています。また共変動する2値をとる二つの変数Y_AとY_Bの因果関係の分析にfixed effects modelを用いる方法はDuncan（1985）が開発しました。Y_{t1}がY_tに影響する場合のfixed effects modelについてはYが2値をとる場合についてYamaguchi（1996）が対数線形モデルを開発しましたが仮定が強く利用に制限があります。

因果分析上fixed effects modelsについて特に留意すべきことは、このモデルはXの値の変化を経験した者についての個人内のXとYの関係からXの効果を測定しているのであくまでトリートメントグループ内のXの効果δ_Tであって、全体での平均的効果$\overline{\delta}$ではないということです。

回帰モデルでの「観察されない異質性」の制御にはfixed effects modelの他にrandom effects modelがあります。後者は観察されない異質性を一定の分布を持つランダム変数で表しますが、このモデルはY_{t1}はY_tに対する影響の過大評価を修正する機能があり、また時間で変化しない説明変数をモデルで用いることができるという利点もあるけれど、その異質性のランダム変数はXとの独立を仮定しているので、状態Xへの選択バイアスは取り除けません。

しかしrandom effects modelを拡張して従属変数Yとそれとの因果関係が問題になる特定の説明変数Xの決定の同時モデルを考え、その誤差項間の相関の有無によって結果が異なるかどうか見る方法があります。イベントヒストリー分析の例ですが、Lillard等（1995）は婚前 同棲に関するBecker理論のテストを選択バイアスの除去の問題と結びつけてうまく分析しています。Becker理論（1981）によれば同棲は試験結婚で、相手に対する情報をもたらすため、その情報を持って結婚すればそうでないより結婚のミスマッチの可能性は減り、従って離婚率は減るはずであるという仮説が成り立ちます。しかし単純な分析をすると通常同棲相手と結婚した夫婦は離婚率が高いとのBecker理論と矛盾する結果を得てしまいます。しかし同棲者と結婚した者がそうでない者より離婚率が高いのは、同棲から結婚へという推移自体が因果的に離婚のリスクを高めるのではなく、婚前同棲傾向の潜在的に高い者が離婚の潜在リスクも高い（婚前同棲と離婚の回帰モデルの残差項間に有意な正の相関がある）ことからきている可能性があり、実際にこの同棲経験への選択バイアスをYの予測式とXの予測式の誤差項間の相関で制御すると、Becker理論と矛盾しない結果が得られるというのがLillard等の分析です。同様のやり方は、より簡単には例えばbivariate probitモデルを用いるなどしてN年内の離婚確率と婚前同棲の確率の同時モデルなどをパネルデータを用いてできます。またこの方法は上記の例のようにXの変化（婚前同棲）がYの変化（離婚）のリスク時外で起こっても、XのYへの因果的影響を計れるという点にfixed effects modelsにはない長所を見い出せます。

またここで注意すべきは因果関係についてXに対するε_yの影響（即ちε_xとε_yとの相関）というXの状態への選択バイアスを制御してXのYへの効果を計ることは、 概念的にはXの係数として$\delta_T=\delta_C$となる状態を仮定してδを推定しているというのに近いということです。ここではfixed effects modelと異なりXはランダムな内生変数なので、Xの状態への選択モデルは擬似実験的なXの状態への標本配分を表し、 ε_xとε_yとの相関はYについてXの値の実現以前に存在するTグループ間とCグループ間の観察されないYの値の違いを制御すると考えられるので、残るXのYへの効果は$\delta_T=\delta_C$となるような状況でのXの影響$\overline{\delta}$の推定値と解釈できるからです。しかしこのYの予測とXの状態の選択との同時モデルは、Xの状態への選択モデルが正しくなければ結果にバイアスをもたらしますから、理論的かつ経験的に選択プロセスの適切な知識を必要とする点で、fixed effects model より知識の要求度の高いモデルといえるでしょう。

5. むすび

このようにパネルデータは分析を複雑にもしましたが、因果関係の解明など、他の調査データではできない様々な分析を可能にし、またこの点で統計データ分析の社会・経済の実態の解明や予測、さらには政策評価などへの利用を大きく前進させたといえます。以上パネルデータ分析についての「常識の誤り」についてでした。後になって私の議論にも誤りがあったと指摘され、新たな神話の題材にならないとよいのですが。

†この講演の準備は経済産業研究所の客員フェローシップにより一部サポートされたので記してここに謝したい。また今回の講演の直接のきっかけを作っていただいた永井暁子氏等家計経済研究所の方々に感謝する。

注

1）回帰式（4）や（6）の回帰係数の標準誤差は回帰式（3）または（5）で ε_{it} にi.i.d.を仮定し $V(\varepsilon_{it})=\sigma^2$ とすると、$V(\varepsilon_{it}-\varepsilon_{i(t-1)})=2\sigma^2$ で

$$COV(\varepsilon_{it}-\varepsilon_{i(t-1)}, \varepsilon_{i(t+u)}-\varepsilon_{i(t+u-1)})=\begin{cases}-\sigma^2 & \text{if } u=1\\ 0 & \text{if } u>1\end{cases}$$

なので、式（4）または（6）を $y^*=X'\beta+\varepsilon^*$ で表しxの $(y_{i(t-1)}-y_{i(t-2)})$ の列をそのinstrumental variable $y_{i(t-2)}$ で置き換えた行列をZとし、Ωを対角成分が2で隣接する準対角成分が−1で、その他の成分が0の行列とすると、パラメーターの推定値は $(Z'X)^{-1}Z'y^*$ でその共分散行列は、より仮定を弱めても有効なGMM法を用いる推定方法もあるが（Wooldridge 2002）、簡単には、$\sigma^2(Z'X)^{-1}(Z'\Omega Z)(X'Z)^{-1}$ で与えられる。

2）Fixed effectsのパラメーター c_i の推定値は各個人の観察値の数は有限なので標本数が増えても一致性を満たさないが、最尤推定法では他のパラメーターの推定値が c_i の推定値に依存してしまうため一致性を持たなくなってしまう問題。

文献

Baltagi, B. H., 1995, *Econometric Analysis of Panel Data,* Chichester: Wiley.

Becker, G. S., 1981, *A Treatise on the Family*, Cambridge: Harvard University Press.

Cameron, A. C., and P. K. Trivedi, 1998, *Regression Analysis of Count Data,* Cambridge: Cambridge University Press.

Chamberlain, G., 1980, "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47: 225-38.

Coleman, J. S., 1964, *Models of Change and Response Uncertainty,* Englewood Cliffs: Prentice-Hall.

Duncan, O. D., 1985, "New Light on the 16-Fold Table," *American Journal of Sociology*, 19: 88-128.

Finkel, S. E., 1995, *Causal Analysis with Panel Data*, Thousand Oaks: Sage.

Hagenaars, J. A., 1990, *Categorical Longitudinal Data,* Newbury Park: Sage.

Heckman, J. J., 1997. "Instrumental Variables," *Journal of Human Resources*, 32 : 442-62.

樋口美雄，2001，『雇用と失業の経済学』日本経済新報社.

Hsiao, C., 1986, *Analysis of Panel Data*, Cambridge: Cambridge University Press.

Lillard, L. A., M. J. Obrien, and L. J. Waite, 1995, "Premarital Cohabitation and Subsequent Marital Dissolution," *Demography*, 32: 437-57.

Petersen, T., 1988, "Analyzing Change over Time on a Continuous Dependent Variable," *Sociological Methodology*, 18: 137-64.

Winship, C. and S. L. Morgan, 1999, "The Estimation of Causal Effects from Observational Data," *Annual Review of Sociology,* 25: 639-706.

Wooldridge, J. M., 2002, *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press.

Yamaguchi, K., 1986, "Alternative Approaches to Unobserved Heterogeneity in the Analysis of Repeatable Events," *Sociological Methodology*, 16: 213-49.

————, 1996, "Some Log-linear Fixed-Effect Latent-Trait Markov-Chain Models," *Sociological Methodology*, 26: 39-78.

山口一男，2001-2002，「イベントヒストリー分析（1）−（15）」『統計』52(9)−53(11).

————，2002，「先代ブッシュ政権崩壊の謎を世論調査のパネルデータ分析から解く」『世論』90: 30-37.

やまぐち・かずお　シカゴ大学社会学部教授。主な著書に*Event History Analysis*（Sage，1991）。社会統計学専攻。