

# 傾向スコアを用いた調査研究からの因果効果の推定について

星野崇宏（慶應義塾大学経済学部・大学院経済学研究科教授）

## 1 導入

傾向スコア (propensity score) は 2000 年代から社会科学（とくに経済学と社会学）や医学研究において頻繁に利用されるようになった手法である。2000 年以降に限定して Web of Science で検索を行うと、タイトルやキーワードまたはアブストラクト中に記載のある論文数は 9,588 件と、分散構造分析や構造方程式モデル（経済学研究での潜在変数のない構造方程式を除外しないで 1 万 1,123 件）とほぼ同等、パス解析の 5,369 件や操作変数法の 1,910 件よりも多く利用されていることがわかる。

傾向スコアはそもそも因果効果 (causal effect, または平均処遇効果: average treatment effect) を推定するために Rosenbaum and Rubin (1983) によって提案された概念であるが、その理論的な背景には欠測データモデルが存在するため、欠測として理解できる様々なバイアスへの対処に利用される。具体的には項目単位の欠測の処理、選択バイアスの除去やパネル調査での脱落の影響の補正、さらには未回収標本によるバイアスの補正や有意抽出標本から無作為抽出標本への結果の調整など、調査において生じる様々なバイアスへの対処への利用が行われている。

本稿はスペースの制約があるため、傾向スコアを用いた因果効果の推定について説明を行い、共変量選択、操作変数法との比較、異質的な因果効果に焦点を当てて説明する。詳細については星野 (2009)、Imbens and Rubin (2015) を、さらに最新のいくつかの方法論的トピックや偏りのある調査への対応への利用については近々発刊される数理社会学会『理論と方法』ワンステップアップ講座 (星野, 印刷中) を参照いただきたい。

## 2 Rubin の因果効果

以下の 2 つの例をみてみよう。いずれも本誌の読者を考慮し、*American Sociological Review* に掲載された研究による。

例 1: 大卒学歴の取得の経済的なリターン  
(Brand and Xie, 2010)

大卒者と高卒（や中退）者の学歴を単純に比較しても、もともと社会経済的地位が高く文化資本が豊かな家庭に育ち、高校時点での能力が高い生徒が大学に進学するため、単純に 2 群の就職後の賃金を比較しても大学進学のリターンとはいえない。

例 2: シングルマザーになることで中年での健康状態は悪化するか (Williams et al., 2011)

アメリカの大規模追跡調査 NLSY を用いて、最初の子どもを産んだ時点でシングルマザーであったグループとそうでないグループで 40 歳時点での健康状態を比較したところ、シングルマザーの健康状態のほうが悪かった。しかしこれは 2 群でそもそも社会経済的地位やソーシャルサポートの質が異なることによるものか、「シングルマザーである」という社会的経験により周囲の偏見などから高いストレスを受けることによるものなのかは単純な比較ではわからない。

これらの研究では進学やシングルマザーになるか否かなど「研究者が容易に操作できない独立変数」が賃金や健康状態などの結果変数にどのような影響を与えるかに関心がある。社会科学全般において、このように独立変数の操作を伴う実験研究が行えない場合は多い。その場合には上記に記載したように育った家庭環境や社会経済的地位な

どといった独立変数にも結果変数にも影響を与える変数（以後、共変量と呼ぶ）の分布が独立変数の水準間（たとえば大卒群と高卒群の間）で異なるため、結果変数の分布を水準間で比較しても「独立変数が結果変数に単独で影響を与える効果」（=因果効果）とはいえない。

社会科学でも伝統的には実験心理学や、また近年では実験経済学や実験政治学で無作為割当を伴う実験研究が盛んに行われている。経済実験や心理実験などで実験中のみ条件操作をすればよいようなインセンティブや教示、提示刺激などの短期的な効果が研究対象であればよいが、教育や生育環境、職種など長期間継続する特定の要因が従属変数に長期的な影響を与えるような研究関心では介入操作が倫理的にも不可能である場合が多い。また、たとえ同意を得て実施しても途中で研究者側の割当に従わなくなる不服従や追跡調査からの脱落が生じる可能性が高い。たとえば前述の解析例2では「シングルマザーであること」の長期的な効果であったが、研究者が家族構成を操作することは倫理的にも現実的にも不可能である。

さらにホーソン効果やピグマリオン効果などいわゆる介入的な実験研究の生態学的妥当性の問題、実験対象者の代表性などからも、実験研究が可能な研究関心であっても代表性のある標本抽出に基づく調査研究から統計的な因果効果推定を行う有用性は高い。

独立変数の操作を伴う実験研究ができない場合、つまり調査研究から因果効果を推定するための統計学的な枠組みにRubin因果モデルがある（Rubin, 1974）。これを説明するために、まず潜在的結果変数という概念の説明を行う。

簡単のためまずは独立変数の水準が介入群と対照群の2水準の場合で説明する。

$Z$ を独立変数の水準を表すダミー変数とし、介入群なら $Z=1$ 、対照群なら $Z=0$ とする。ここで観測される結果変数を $Y$ とするとき、その背後に「もし介入群に対象者が割り当てられたときに得られるであろう変数」 $Y_1$ と「もし対照群に対象者が割り当てられたときに得られるであろう変数」 $Y_0$ の2つの潜在的結果変数が存在する、と考える。ただし、実際に観測されるのはそのどちらかであり、 $Y$ と $Y_1$ 、 $Y_0$ の関係は被験者がどちらの群に割り当てられたかを表すダミー変数 $Z$

を用いて

$$Y = ZY_1 + (1 - Z)Y_0$$

と表現できる。たとえば、大卒を $Z=1$ 、高卒を $Z=0$ とすると、大卒者ではその人がもし大卒だった場合に得られたであろう賃金 $Y_1$ が実際に賃金 $Y$ として観測され、高卒者では同様に高卒だったら得られたであろう賃金 $Y_0$ が観測される。このように、 $Z=1$ ならば $Y_1$ は観測されるが $Y_0$ は欠測となり、 $Z=0$ ならば逆に $Y_0$ は観測されるが $Y_1$ は欠測となる。

さてここでRubinの因果効果、または平均介入効果（average treatment effect: ATE）は $Y_1 - Y_0$ の期待値 $E(Y_1 - Y_0)$ として定義される。これは、2つの潜在的結果変数の差の母集団平均であり、「個人ごとに $Y_1 - Y_0$ を計算して母集団で平均したもの」である。これは個人内で差をとっているため個人差が除去されており、2つの条件の違いだけで結果変数の平均がどの程度異なるか、を示すものであるため、個人によって値の異なる共変量 $X$ には依存しない量であるため、その期待値を因果効果と呼んでよい。

この因果効果を推定する最も単純な方法は、個人ごとの差変数 $Y_1 - Y_0$ の対象者全体での標本平均、つまり $(\overline{Y_1 - Y_0}) = \bar{Y}_1 - \bar{Y}_0$ を求めることであるが、実際には潜在的結果変数は一方が観測されれば他方は欠測するので、この量は求めることはできない。一方、通常の群間平均値差は各群の標本平均の差 $\bar{Y}_{1(\text{介入群})} - \bar{Y}_{0(\text{対照群})}$ であり、その期待値は $E(Y_1|Z=1) - E(Y_0|Z=0)$ であるため、 $ATE = E(Y_1 - Y_0)$ の推定量として考えるとバイアスが生じる。

また、ここでもし条件の無作為割当（実験研究）が可能ならば、 $Z$ と $Y_1$ 、 $Y_0$ が独立であるため、 $E(Y_1|Z=1) - E(Y_0|Z=0) = E(Y_1) - E(Y_0) = ATE$ となり、群間平均値差は因果効果ATEの不偏推定量となり、単純な群間比較を行えばよいことがわかる。

### 3 因果効果の推定と傾向スコア

ここで、もし2群で分布が異なる共変量 $X$ に対して

$$\begin{aligned} E(Y_1|X) - E(Y_0|X) \\ = E(Y_1|X, Z=1) - E(Y_0|X, Z=0) \quad (1) \end{aligned}$$

が成立する場合、つまり共変量が同じ値であれば、どちらの群に限定して期待値をとってもよい（各群に限定してよい。ということは欠測を無視してよい）、という条件が成立すれば、式（1）の右辺を  $X$  について期待値をとると  $E_X E(Y_1 - Y_0 | X) = E(Y_1 - Y_0) = ATE$  となる。したがってもし式（1）が成立すれば、介入群と対照群で  $X$  が同じ2人をマッチングして結果変数の差をとってから  $X$  について平均をとる、あるいは  $X$  が同じ集団に層別して、層内の群間差を出してから  $X$  について平均をとることで因果効果がバイアスなく推定できることになる。

この式（1）の条件をベイズの定理を用いて表現し直すと

$$\begin{aligned} Pr(Z=1 | Y_1, Y_0, X) &= Pr(Z=1 | X) = e(X) \\ 0 < e(X) < 1 \end{aligned} \quad (2)$$

つまり、割当は潜在的結果変数には依存せず、傾向スコア推定に用いた共変量  $X$  にのみ依存する、という条件で表すことができる。これを「強く無視できる割当条件（Strongly Ignorable Treatment Assignment）」と呼ぶ。ここで  $e(X)$  を傾向スコアと呼ぶ。この強く無視できる割当条件が成立する場合には、マッチングや層別分析をすれば因果効果が原理上は推定できるが、通常は上記の条件が満たされるような  $X$  は多変数であり、複数変数について単純にマッチングや層別を行うことは難しい。

そこで、共変量  $X$  の情報を「共変量  $X$  をもった対象者が介入群へ割り当てられる確率」である傾向スコアに一元化することで、因果効果を推定しようというのが傾向スコア解析である。傾向スコア  $e(X)$  は実際にはその真値はわからないため、2値の群別変数  $Z$  を  $X$  で予測するロジスティック回帰分析やプロビット回帰分析で計算される各対象者の  $Z=1$  の予測確率とする。

独立変数が3水準以上では、その水準に割り当てられる予測確率（一般化傾向スコアと呼ばれる）を利用すればよい。

## 4 傾向スコアの具体的な利用方法

傾向スコアを用いた因果効果の推定法として Rosenbaum and Rubin (1983) の元論文では、傾向スコアの値がなるべく近くなるようなマッ

グや、傾向スコアを用いた層化、傾向スコアを共変量とした共分散分析が提案されている。このうち最もよく利用されてきたのはマッチングである。もともと傾向スコアが考え出されるきっかけとなったのは、共変量の数が多い場合に共変量をそのまま利用したマッチングは実行が難しいという問題であった。

ただし複数の共変量を傾向スコアに一元化しても、そもそもマッチングは「なるべく値の近い対照群と介入群をペアにする」という方法であるため、「近さの基準＝距離の選択」や「どれだけ遠いとマッチングしないかの基準＝caliper」などを決める際の恣意性がある。また、単純に1対1のマッチングを行う場合、介入群のサンプルサイズ > 対照群のサンプルサイズ、であればマッチングによって得られる結果は「介入群での因果効果（average treatment on the treated）」  $E(Y_1 - Y_0 | Z=1)$  になる。

マッチングで ATE を推定したければ復元抽出を許すかたちでのマッチング、あるいは1対多・多対多のマッチングが必要になるが、その場合も数を決める際の恣意性があるため、たとえば標本全体での共変量の分布とマッチング後の集団の共変量の分布が同じになるような重みを付けた制約付きマッチングが行われることも多い。層化の場合には5層に分けることが多いが、層別の基準や層の数などで結果が異なるという恣意性がある。

一方、共変量として利用する際には0から1までの値をとる傾向スコアを用いた線形回帰は望ましくないことから、カーネルマッチングや局所線形回帰といった手法が Heckman らによって提案されている (Heckman et al., 1998)。これは一種のノンパラメトリック回帰であるが、たとえばある介入群に所属する個人とのペアを対照群に所属する1人や数人の個人とマッチさせるのではなく、対照群に所属するすべての人の値を、カーネル関数の重みで利用するという特殊なマッチングとして理解することもできる。

また、上記の3つの方法以外によく利用される方法として、傾向スコアの逆数を重みとする逆確率重み付け (inverse probability weighting) 推定がある。これは標本調査論における Horowitz-Thompson 推定量が抽出確率の逆数で重みを付けるように、「全員が介入を受けた場合の潜在的

結果変数」を「介入群への傾向スコア=その対象者が介入群として抽出され観測される確率」の逆数で、「全員が対照条件にいた場合の潜在的結果変数」を「対照群への傾向スコア」の逆数で、それぞれ重み付けて平均をとる、という明快な方法である。標本抽出の場合と異なり、傾向スコアを推定したことに伴う重みの変動を考慮する必要があるが、ここでの重みの推定の影響は通常の段階推定とは異なり、推定量の分散を一般に小さくする方向に働くという特殊な性質がある(星野, 2009)。

さらに、「傾向スコア推定の際の割当のモデル」と「共変量で従属変数を説明する回帰モデル」のどちらも利用して解析を行う二重にロバストな推定法も提案されている。これは2つのモデルのどちらか一方が正しくモデル指定されていれば、正しい因果効果を与えるものであり、モデルの誤設定を避けながら効率的な推定法を与えるものとして注目されている。

## 5 共変量選択

傾向スコアを利用するにせよしないにせよ、共変量調整による因果効果推定的前提条件は式(2)であり、この条件を満たすように共変量を選べばよい。これは「傾向スコア推定に用いた共変量によって割当を説明するモデルに潜在的結果変数を説明変数として投入しても、割当の説明力は変化しない」という条件であるが、 $Y_1$ ,  $Y_0$ は同時に観測されないで、データからこのことを示すことはできない。そこで式(2)を間接的に示す共変量選択の妥当性のチェックの方法として以下の2つの方法が利用されることが多い。

- (1) 共変量  $X$  が割当  $Z$  を説明するロジスティック回帰等のモデルの予測力を示す。

式(2)から、 $X$ が十分 $Z$ を説明していれば、 $Y_1$ ,  $Y_0$ は $Z$ の予測のための説明変数として不要であることが推測されるからである。具体的な予測力の判定方法としてはロジスティック回帰の正判別率よりもc統計量など感度と特異度を同時に考慮した指標を利用することが推奨される。正判別率を利用すると、たとえば介入群のサンプルサイズが全体の5%などと小さいときには、全対象

者を閾値に「対照群」と判別すれば正判別率は95%ときわめて高くなるためである。医学分野などではc統計量は0.8以上などという基準が確立しているようである。

- (2) 調整に利用した共変量についても、傾向スコアを用いて調整を行い、調整後の群間差が小さくなることを示す。

式(2)が成立するならば、共変量  $X$  の分布の群間差が消失しているはずであるので、そのことを示せばよい。

これら2点のチェックはあくまで式(2)を間接的に示すものであり、直接的なチェックになってはいないため、共変量の分布が条件間で等しくなることを重視して傾向スコアを選択することや、割当を最も説明するような傾向スコアを選択することが必ずしも式(2)の成立の可能性を高めることにはつながらない。たとえば割当の説明力だけを重視するならば、「結果変数に関連の強い変数」と「割当に関連の強い変数」であれば後者を選んで傾向スコアを推定すればよいが、実際にはそれを行うと因果効果の推定において二乗誤差が大きくなるのがシミュレーションによって示されている(Brookhart et al., 2006)。これは重回帰分析において「本来従属変数への偏回帰係数がゼロである」説明変数を投入しても他の偏回帰係数の推定のバイアスにはつながらず、推定値の標準誤差が大きくなるのと同様であり、傾向スコアを用いた解析において従属変数と直接関係のない変数を共変量として投入することで推定の二乗誤差が増えるためである。

さて、Imbens and Rubin (2015) は以下の共変量選択法を推奨している。これは、まず共変量候補  $K$  個のうち

- (1) 先行研究の知見等から必ず入れる  $Kb$  個を指定する
- (2)  $K - Kb$  個についてロジスティック回帰分析で割当に関連するかという観点から  $Kl - Kb$  個を変数増加法を用いて選択する
- (3)  $Kl$  個の変数の2次項とその交互作用項から  $Kq$  個選択し、1次項を含めて  $Kq + Kl$  個を変数増加法を用いて選択する
- (4) 推定した傾向スコアで層別して共変量の分布の違いがなくなっているかをチェックする

というものである。とくに(1)は先行研究の知見から従属変数に関連するといわれる変数を網羅的に利用することで結果変数との関連が高い共変量を選択し、そのうえで割当を説明する精度を高めるという方法であり、上に記載した「結果変数との関連」を考慮した変数選択となっている。

## 6 操作変数法との関係

傾向スコアは観測されていない共変量の影響を排除できないため、経済学の実証研究でよく利用される操作変数法を用いたほうがよい、とする意見があるが、本当であろうか。

確かに適切な操作変数ももしみつきそれが測定できているのであれば、操作変数法は内生性バイアスを排除できる非常に強力な手法である。このため、しばしば傾向スコアよりも操作変数法のほうがよいといわれることがある。しかし操作変数は「独立変数と相関があるが、独立変数で従属変数を説明した際の残差とは相関がない」という非常に強い条件を満たす変数であり、この仮定が成立しない場合には推定に大きなバイアスが生じることは様々な研究で指摘されている。

実際に、ある変数が自分の解析状況において操作変数であることを示すことは困難である。ではなぜ経済学でしばしば使われるかといえば、経済政策の変更という自然実験(Cameron and Trivedi, 2005)においては政策変更以外とは関連しないと仮定できる変数が存在する場合があることによる。

ここで、いくつか適切な操作変数の例を挙げよう。教育年数(説明変数)の賃金(従属変数)への効果について考える際に、残差には教育年数以外で賃金に影響を与えるであろう知能や非認知能力が含まれており、残差と説明変数に相関があるので回帰分析での最小二乗法や最尤法といった通常の解析での仮定が満たされず、推定量にバイアスが生じる(賃金のすべての決定要因が説明変数として利用されていないことによる除外変数バイアスと呼ばれる)。ここで、「生まれた四半期」変数は知能など教育以外の決定要因と関連がなく、教育年数と関連が高いと考えられる変数であるため、操作変数として適切であると考えられる(Angrist and Kruger, 1991)。

別の例として、服役囚が生まれ育った地域はつるむ仲間がいたり地域での失業率の高さなど、もともと犯罪を犯しやすい地域である場合が多い。では服役囚が釈放後生まれ育った地域に戻ることによって再犯率が高まるという因果関係が議論できるかということ、釈放後どの地域に住みたいかという嗜好と再犯の犯しやすさどちらにも影響を与える要因(たとえばリスク志向や親族などとの関係など)が影響している可能性があるため、育った地域に戻った群と戻らない群での再犯率の単純な比較では因果効果はわからない。そこでKirk(2009)は、ハリケーン・カトリーナの被害によって外生的に育った地域以外の地域に戻らざるをえない人が多かった時期の前後のデータを利用し、釈放がハリケーン前か後かという変数を操作変数とすることで、「育った地域に戻るか戻らないか」以外で再犯率に影響を与える変数の影響を除外した因果効果(正確には局所平均介入効果)の推定を行っている。

前者は社会制度として外生的に規定される生まれた月による教育年数の違いを利用した説得的な操作変数法の利用例であり、後者は甚大な被害を与えたハリケーンを一種の自然実験状況と見立てている。

これらの例からもわかるように、理論的な見地からそれぞれ操作変数としての正当性が主張されるものであって、あくまでも研究者間での同意によるものである。社会学などの研究の場合、なかなか操作変数の条件に従う変数を見つめることは難しいであろう。

また、近年では操作変数法は経済学でもその仮定の強さから、さらには操作変数法を使わずに本来背後に存在するメカニズムをモデル化して解析するほうがよい(いわゆる構造推定と呼ばれる考え方)という議論からしばしば批判されることも多い(たとえばHeckman and Urzúa, 2010など)。

したがって社会学などでは、経済学同様の政策変更に伴う自然実験的な状況でなければ、操作変数法を使うよりは社会調査の事前の設計時点で様々な共変量候補を積極的に測定するような調査を企画実施し、それらの情報を利用した共変量調整を行うほうがよいと思われる。

さらに、操作変数法による解析で基本的には因果効果の推定は行うことはできず、局所平均介入

効果 (local average treatment effect) という部分母集団に対する因果効果の推定を行っていることにも注意が必要である (Imbens and Angrist, 1994; 星野, 2009)。

また、条件の違いを単純にダミー変数 (条件ダミー) として説明変数にした回帰モデルの推定において、条件ダミーと誤差が独立ではない場合には操作変数法を用いて偏回帰係数を求めることがある。このとき、条件ダミーの偏回帰係数が因果効果となるための条件は「背後に存在する潜在的結果変数の2つの回帰モデルでの誤差が全く同一であること」、つまり「2つの条件は平均レベルでのみ違いがあり、それ以外では同じ」という非現実的な仮定になる (通常の因果効果推定では誤差には相関があるが同一ではないといった程度の仮定が置かれるためより仮定が緩い)。

## 7 異質的 (heterogeneous) な因果効果

多くの研究関心では母集団全体、あるいは介入対象の部分母集団における因果効果だけではなく、「介入群に割り当てられる確率と因果効果の大きさの関係」などに関心がある場合がある。冒頭の解析例1で大卒群を介入群とすると、「大学進学しやすい=大学進学の傾向スコアが高い人と、しにくい=傾向スコアが低い人で大卒による賃金上昇効果がどれくらい違うのか」は、特定の大学進学率が過剰かそうでないかを判断するなど、教育政策的に大きな意味をもつ。実際、解析例1は「大学進学はもともと大学に行く可能性がある人にとって得か、あるいは逆か?」という因果効果の異質性を調べている研究である。

進学行動と賃金の関係については、経済学分野の研究では進学行動にも経済的な効用最大化、つまり「リターンが得られる確率が高いと考える人ほど進学する」といういわゆる positive selection 仮説が想定されることが多い。一方、進学と職業達成や再生産の関係は教育社会学において重要な研究テーマであり、経済的な効用最大化に限定されずに家庭の経済力や文化資本、ピアの効果などが進学行動の意思決定要因として想定される。この観点からは、家庭が裕福でクラスメートも皆大学に進学するのが当然だという高校に在籍したもともと大学に進学しやすい学生は経済的な効用最

大化を考慮せずに「当然のこととして」大学進学を行う一方、進学しにくい学生ほど経済的な効用最大化を考慮して実際に賃金上昇が大きく見込めるような場合にしか進学せず、結果として大卒による賃金上昇効果は進学しにくい学生において高いという negative selection 仮説が想定される。Brand and Xie (2010) は階層線形モデルと傾向スコアによる層別を行い、傾向スコアの低い=大学進学を行いにくい環境にいる学生のほうが大卒による賃金上昇効果が高いという解析結果を示している。

このような「因果効果の異質性」を明示的に考慮した解析法として近年利用されつつあるのが周辺介入効果 (marginal treatment effect) を用いた解析である (Heckman and Vytlacil, 2005)。周辺介入効果自体は ATE や局所平均介入効果など様々な因果効果を一般化した量であるが、とくに傾向スコアを引数とする周辺介入効果  $E(Y_1 - Y_0 | e(X))$  は、「傾向スコアが高いほうから上位何% までには効果があるか」といったことを推論するのに非常に有効な量である。たとえば Carneiro et al. (2010) による解析結果の図4は周辺介入効果を図示しているが、横軸の  $U_s$  は観測されない「介入群へのなりにくさ」を示す値であり、1から傾向スコアを引いた値であって、横軸の左から大学進学をしやすい確率の高い群のパーセンタイルの人たちに対応した賃金の上昇幅 (を用いて収益率を計算したもの) が縦軸となっており、ここでは positive selection 仮説が示されていることがわかる。

注  
・1 朝倉書店から2016年初頭をめどに翻訳が刊行される予定である。

文献  
Angrist, J. D. and A. B. Krueger, 1991, "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106(4): 979-1014.  
Brand, J. E. and Y. Xie, 2010, "Who Benefits Most from College? Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education," *American Sociological Review*, 75(2): 273-302.  
Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn and T. Stürmer, 2006, "Variable Se-

- lection for Propensity Score Models," *American Journal of Epidemiology*, 163(12): 1149-56.
- Cameron, A. C. and P. K. Trivedi, 2005, *Microeconometrics: Methods and Applications*, NY: Cambridge University Press.
- Carneiro, P., J. J. Heckman and E. Vytlačil, 2010, "Estimating Marginal Returns to Education," *American Economic Review*, 101(6): 2754-81.
- Heckman, J. J., H. Ichimura and P. Todd, 1998, "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65: 261-94.
- Heckman, J. J. and E. Vytlačil, 2005, "Structural Equation, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73: 669-738.
- Heckman, J. J. and S. Urzúa, 2010, "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify," *Journal of Econometrics*, 156(1): 27-37.
- 星野崇宏, 2009, 『調査観察データの統計科学——因果推論・選択バイアス・データ融合』岩波書店。  
—— (印刷中) 『傾向スコア解析の基礎と応用』  
『理論と方法』31(2)。
- Imbens, G. and D. B. Rubin, 2015, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, NY: Cambridge University Press.
- Imbens, G. and J. Angrist, 1994, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2): 467-75.
- Kirk, D. S., 2009, "A Natural Experiment on Residential Change and Recidivism: Lessons from Hurricane Katrina," *American Sociological Review*, 74(3): 484-505.
- Rosenbaum, P. R. and D. B. Rubin, 1983, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1): 41-55.
- Rubin, D. B., 1974, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66(5): 688-701.
- Williams, K., S. Sassler, A. Frech, F. Addo and E. Cooksey, 2011, "Nonmarital Childbearing, Union History, and Women's Health at Midlife," *American Sociological Review*, 76(3): 465-86.

