JGSS 統計分析セミナー2013

ー傾向スコア・ウェイティング法を用いた応用モデルー

曺　成虎
大阪商業大学 JGSS 研究センター*

JGSS Statistical Analysis Seminar:
Applied Model Using Propensity Score Weighting

Sung-ho CHO
JGSS Research Center
Osaka University of Commerce

This paper is a summary of the lecture hosted by JGSS Research Center in September, 2013. The topic of the seminar was casual models using propensity scores. We can obtain efficient parameter estimations using RCM with propensity scores of inverse probability by removing correlations with confounding factors.

This article discusses applied models using propensity scores and focuses mainly on DFL and HP-IPW methods. Because DFL methods can decompose factors in non-linear models, it is a complementary model of Blinder-Oaxaca model, which has widely been used in Economics. HP-IPW has advantage of having unbiased estimators because it added advantages of RCM on Heckman model so that it mitigates assumption of correlation between endogenous and dependent variables. Little research has been conducted using HP-IPW up to now. According to the results we found in this paper, however, it could be a valuable method.

Key Words: JGSS, propensity score, causality analysis

本稿は 2013 年 9 月 2 日、3 日の両日間行われた JGSS 研究センター主催の統計分析セミナーの講義内容をまとめたものである。セミナーのテーマは 2009 年と 2011 年に引き続き、「傾向スコア」を用いた因果分析であった。RCM は従来の OLS において仮定された説明変数間の独立性仮定が緩和されたモデルであり、傾向スコアの逆確率を用いて、主に説明したい変数が交絡要因と相関があってもその相関を除去することで、効率的な推定量を得ることができるモデルである。

本稿では特に傾向スコア・ウェイティング法を用いて応用したモデルについて解説した。ここで取り上げた DFL 法は非線形モデルをベースにした場合の要因分解ができることで、経済学において幅広く使われてきた B-O 法を補う形のモデルともいえよう。そして、HP-IPW 法は RCM のメリットを Heckman 法に適用することで、従来の仮定を緩和させ交絡変数が内生変数や従属変数間に相関がある場合にも一致推定量が得られる方法である。この推定方法を用いて書かれた論文は、現在それほど見当たらず、今後の活用が期待される。

キーワード：JGSS、傾向スコア、因果分析

---

*現所属：韓国保健社会研究院人口政策研究本部

1. はじめに

2013 年 9 月 2 日と 3 日の 2 日間、JGSS 研究センター主催の統計分析セミナーが開催された。このセミナーは 2007 年から始まっており、これまで数多くの大学院生・研究者に統計分析のスキル向上に貢献してきた。講師はシカゴ大学社会学部の山口一男教授が担当されており、2007 年の初回のセミナーから引き続きセミナーの講師担当をしていただいた。この場を借りて感謝の意を表する。

2013 年の統計分析セミナーのテーマは「傾向スコア（propensity score)」を用いた因果分析であった。このテーマは 2009 年、2011 年に引き続き今年も行われた。このことは傾向スコアを用いた分析の重要性がますます増えてきたことを反映していると考えられる。傾向スコアを用いた因果分析の基本的な説明は過去の文献を参照していただくことにし（三輪・菅澤 2010、林 2012)、今回は上記の論文で扱っていなかった内容を中心に説明する。

本稿では DiNardo, Fortin, Lemieux 法（DiNardo et al. 1996、以下 DFL 法）と、Heckman 法と傾向スコアの逆確率のウェート（Inverse Probability Weight: 以下 IPW）を併用した分析について解説する。DFL 法は経済学において多く使われている Blinder-Oaxaca 法（Blinder 1973; Oaxaca 1973、以下 B-O 法）とともに要因分解をする代表的な分析手法であるが、B-O 法はパラメトリックな回帰分析に基づいた分析である反面、DFL 法は傾向スコアのウェーティング法に基づくことに相違点がある。たとえば、B-O 法は線形回帰モデルの場合にのみ有効に要因分解できるが、DFL 法は非線形回帰モデルの場合にも有効に要因分解できるメリットがある。

図 1　DFL 法の基本的な模式図（1）

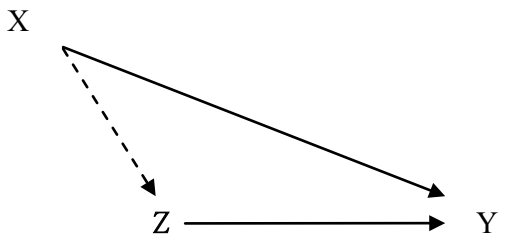


図 1 は DFL 法の基本的な模式図を表している。X は人種、Y は従属変数、Z は仲介変数（mediating variable）であり、Z を介さない X から Y への影響を推定することができる。すなわち、DFL 法を図 1 から直感的に説明すると、傾向スコアを用いて Z から X を予測し(1)、逆にその予測ができなくする。それにより Z と X との相関を取り除くことができ、X から Y への純粋な効果が推定できる。なお、X 変数による格差も推定でき、たとえば、Y を主観的な階層意識、X を人種、Z を収入、教育水準、職業威信（occupation prestige)、年齢であるとすると、主観的な階層意識への影響は Z の違いにより説明できるのか、それとも X の独自の影響が実際あるのかどうかを推定できるということであり、さらに、それらの変数による人種（X）の格差があるかどうかを推定できることである。

Heckman 法と傾向スコアの IPW を併用した分析（以下、HP-IPW 法）は従来の回帰分析の仮定をより緩和させ、ATT（Average Treatment effect for the Treated）を測定する方法である。ここで採用する Heckman 法は、サンプルの選択バイアスを取り除く方法ではなく（Heckman 1979)、処理効果の選択バイアスを取り除く方法である（Heckman and Robb 1986)。HP-IPW 法のメリットは 2 つあり、①IPW 法において仮定された SITA（Strong Ignorablility of Treatment Assignment）条件を緩和できること、②Heckman 法において仮定された内生変数以外の説明変数と誤差項とは独立であるという条件を緩和できることである。

本稿の構成は次の 2 節では DFL 法について、基本的なモデルや DFL 法を用いた実証分析の演習をし､3 節では 2 節と同様に HP-IPW 法の基本的なモデルや HP-IPW 法を用いた実証分析の演習をする。そして、4 節のおわりには、本稿の内容を簡潔にまとめる。

2. DiNardo, Fortin, Lemieux 法（DFL 法）

2.1 基本的なモデル

X を 2 つのグループ、たとえば、男性と女性あるいは 2 つの人種を表す変数とし、次のような式を想定する。

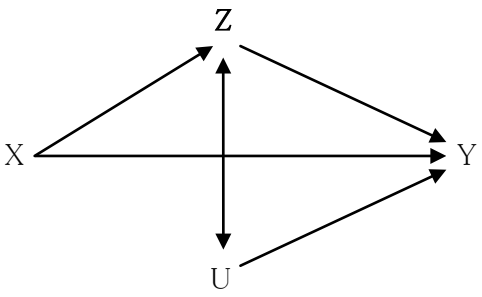
$$\begin{aligned} y_i &= \phi(\mathbf{v_i}, \mathbf{z_i}, \boldsymbol{\theta_1}) + \varepsilon \quad \text{for person } i \text{ in the group with X=1, and} \\ y_i &= \phi(\mathbf{v_i}, \mathbf{z_i}, \boldsymbol{\theta_0}) + \varepsilon \quad \text{for person } i \text{ in the group with X=0} \end{aligned} \tag{1}$$

ここで、ϕ は特定していない関数形、z は媒介変数（mediating variable）、u は観察されない交絡変数、$\boldsymbol{\theta_1}$ と $\boldsymbol{\theta_0}$ は係数を表す。(1) 式はパラメトリックな回帰分析の仮定より弱い仮定を置くことで、より有効な推定量を得ることができる。したがって、次のような仮定を置く。

$$\mathbf{U} \perp X|\mathbf{Z} \tag{2}$$

これは因果分析の処理変数（treatment variable）における外生性の仮定と近似しているものである。DFL 法で想定している因果関係を表すと図 2 のようになる。図 2 は図 1 に U が加わっているものであり、一般的な因果モデルは観察されない交絡要因（confounding factor）がないと仮定しているのに比べ、DFL 法は Z をコントロールした X は U と独立であることを仮定しており、一般的な因果モデルより制約が緩和されていることがわかる。

図 2　DFL 法の基本的な模式図（2）



ここからは Y の平均差を説明される要因と説明されない要因に分解することを考えることにする。Y_1 はパラメータ $\boldsymbol{\theta_1}$ から推定された結果、Y_0 はパラメータ $\boldsymbol{\theta_0}$ から推定された結果であり、X をコントロールした結果ともいえる。したがって、各グループにおける Y の平均は次のように表すことができる。

$$\begin{aligned} \mathrm{E}(\mathrm{Y}_1|x{=}1) &= \int_{\mathbf{v}} \int_{\mathbf{u}} E(Y|\mathbf{z},\mathbf{u},\boldsymbol{\theta_1}) f(\mathbf{z},\mathbf{u}|x{=}1) d\mathbf{u} d\mathbf{z} \\ &= \int_{\mathbf{v}} \left[\int_{\mathbf{u}} E(Y|\mathbf{z},\mathbf{u},\boldsymbol{\theta_1}) f(\mathbf{u}|\mathbf{z},x{=}1) d\mathbf{u} \right] f(\mathbf{z}|x{=}1) d\mathbf{z} \\ &= \int_{\mathbf{v}} E(Y|\mathbf{z},\boldsymbol{\theta_1}) f(\mathbf{z}|x{=}1) d\mathbf{z} \end{aligned} \tag{3}$$

and、

$$\begin{aligned} \mathrm{E}(Y_1|x{=}0) &= \int_{\mathbf{v}} \int_{\mathbf{u}} E(Y|\mathbf{z},\mathbf{u},\boldsymbol{\theta}_0) f(\mathbf{z},\mathbf{u}|x{=}0) d\mathbf{u} d\mathbf{z} \\ &= \int_{\mathbf{v}} \left[\int_{\mathbf{u}} E(Y|\mathbf{z},\mathbf{u},\boldsymbol{\theta}_0) f(\mathbf{u}|\mathbf{z},x{=}0) d\mathbf{u} \right] f(\mathbf{z}|x{=}0) d\mathbf{z} \\ &= \int_{\mathbf{v}} E(Y|\mathbf{z},\boldsymbol{\theta}_0) f(\mathbf{z}|x{=}0) d\mathbf{z} \end{aligned} \tag{4}$$

$\mathrm{E}(Y_1|x{=}1) - \mathrm{E}(Y_1|x{=}0)$を分解するためには、$x$=0 の人々が x=1 の人々と同様の分散を持つような、Y の反事実的な平均（counterfactual mean）を考慮する必要がある。これを式に表すと、

$$\begin{aligned} \mathrm{E}(Y_1|x{=}1) &= \int_{\mathbf{v}} \int_{\mathbf{u}} E(Y|\mathbf{z},\mathbf{u},\boldsymbol{\theta}_0) f(\mathbf{z},\mathbf{u}|x{=}1) d\mathbf{u} d\mathbf{z} \\ &= \int_{\mathbf{v}} \left[\int_{\mathbf{u}} E(Y|\mathbf{z},\mathbf{u},\boldsymbol{\theta}_0) f(\mathbf{u}|\mathbf{z},x{=}1) d\mathbf{u} \right] f(\mathbf{z}|x{=}1) d\mathbf{z} \quad \text{(due to the ignorability assumption)} \\ &= \int_{\mathbf{v}} \left[\int_{\mathbf{u}} E(Y|\mathbf{z},\mathbf{u},\boldsymbol{\theta}_0) f(\mathbf{u}|\mathbf{z},x{=}0) d\mathbf{u} \right] f(\mathbf{z}|x{=}1) d\mathbf{z} \\ &= \int_{\mathbf{v}} E(Y|\mathbf{z},\boldsymbol{\theta}_0) f(\mathbf{z}|x{=}0) d\mathbf{z} \end{aligned} \tag{5}$$

where $$\omega(\mathbf{z}) \equiv \frac{f(\mathbf{z}|{=}1)}{f(\mathbf{z}|\mathrm{x}{=}0)} = \frac{p(x{=}0)p(x{=}1|\mathbf{z})}{p(x{=}1)p(x{=}0|\mathbf{z})}$$

のようになる。$\mathrm{E}(Y_0|x{=}1)$の推定は ATT（average treatment effect for the treated）の推定とかなり近似していることがわかる。唯一の相違点は Z が交絡変数ではなく、仲介変数であることである。$\mathrm{E}(Y_1|x{=}1) - E(Y_0|x{=}0)$はまた次のように 2 つの要因の合計として表すことができる。

$$\mathrm{E}(Y_1|x{=}1) - E(Y_0|x{=}0) = \{E(Y_1|x{=}1) - E(Y_0|x{=}1) + \{E(Y_0|x{=}1) - E(Y_0|x{=}0)\}$$

この分解は X の Y への影響を直接効果と間接効果に分解すると理解することができる。右辺の第 1 項は仲介変数が同様である場合に 2 つのグループ間の差を表しており、これを直接効果といえる。そ

して、第 2 項は仲介変数の分布から説明される結果の差であり、これは間接効果である。しかしながら、DFL 法は若干の制限がある。B-O 法と同様に DFL 法もグループ間における説明変数の分布を合わせることにより分解される。Z が仲介変数であるため、周辺分布は仲介変数により変化する。したがって、あるグループの分布をその他の分布に合わせることは、非現実的であるため、分布を標準化する方法ではなく、Z の分布が変わらないようするために Z の周辺分布を標準化する方法を用いて総効果を直接効果と間接効果に分解することにする。これは RCM（Rubin's Causal Model）をベースにした因果分析から$\mathrm{E}(Y_1)$と$\mathrm{E}(Y_1)$を推定した結果とかなり近似していることになる。直接効果が観察される場合、2 つのグループにおける Z の周辺分布を用いた Y の反事実的な平均は次のようになる。

$$\mathrm{E}(Y_{1,\,direct}) \equiv \int_{\mathbf{v}} E(Y_1|\mathbf{z},\boldsymbol{\theta}_1)\, f(\mathbf{z})d\mathbf{z} = \int_{\mathbf{v}} \omega_1(\mathbf{z})E(Y_1|\mathbf{z},\boldsymbol{\theta}_1)\, f(\mathbf{z}|x=1)d\mathbf{z} \qquad (6)$$

where

$$\omega_1(\mathbf{z}) \equiv \frac{f(\mathbf{z})}{f(\mathbf{z}|\mathrm{x}=1)} = \frac{p(x=1)}{p(x=1|\mathbf{z})}$$

Similarly,

$$\mathrm{E}(Y_{0,\,direct}) \equiv \int_{\mathbf{v}} E(Y_0|\mathbf{z},\boldsymbol{\theta}_0)\, f(\mathbf{z})d\mathbf{z} = \int_{\mathbf{v}} \omega_0(\mathbf{z})E(Y_0|\mathbf{z},\boldsymbol{\theta}_0)\, f(\mathbf{z}|x=0)d\mathbf{z} \qquad (7)$$

where

$$\omega_0(\mathbf{z}) \equiv \frac{f(\mathbf{z})}{f(\mathbf{z}|\mathrm{x}=0)} = \frac{p(x=0)}{p(x=0|\mathbf{z})}$$

分析の手順は IPW と同様に、ロジットモデルによる回帰分析を行うことによりウェートを推定することができ、そこから y の平均ウェートの$\mathrm{E}(Y_{1,\,direct})$と$\mathrm{E}(Y_{0,\,direct})$が得られる。直接効果は$\mathrm{E}(Y_{1,\,direct}) - \mathrm{E}(Y_{0,\,direct})$、間接効果は$(\mathrm{E}(Y_1|x=1) - E(Y_0|x=0)) - (\mathrm{E}(Y_{1,\,direct}) - E(Y_{0,\,direct}))$を計算することで得られる。

2.2 DFL 法を用いた実証分析の演習例

ここからは DFL 法を用いた実証分析の練習を行う。データはアメリカの General Social Survey (GSS) であり、1990 年のデータを用いる。分析に用いる変数の記述統計量は表 1 に表しており、図 1 の Y にあたる主観的な階層意識変数は、4 段階で測っており、1 は下位層、2 は労働者層、3 は中産層、4 は上位層である。X にあたる人種変数は黒人であれば 1、その他の人種（白人・黄色人等）は 0 である。Z にあたるその他の変数は表 1 に表している通りである。

冒頭で言及したように、主観的な階層意識が人種あるいはその他の要因によって、それぞれどの程度説明されるかについて分析を行う。まず、上述したように X と Z との相関を無くす必要があり、その作業は IPW と同様にロジットモデルによりウェートを推定する（表 2）。

表１　記述統計量（GSS90）

	平均	標準偏差
被説明変数		
主観的な階層意識	2.460	0.615
説明変数		
人種	0.108	0.310
媒介変数		
世帯所得		
5,000ドル未満	0.042	0.201
5,000～7,000ドル未満	0.037	0.189
7,000～10,000ドル未満	0.046	0.210
10,000～15,000ドル未満	0.084	0.278
15,000～20,000ドル未満	0.099	0.299
20,000～25,000ドル未満	0.096	0.295
25,000ドル以上†	0.596	0.491
職業威信の指数		
20未満†	0.080	0.272
20～30未満	0.125	0.331
30～40未満	0.284	0.451
40～50未満	0.217	0.413
50～60未満	0.179	0.384
60～70未満	0.069	0.254
70以上	0.045	0.207
年齢		
20代†	0.215	0.411
30代	0.302	0.460
40代	0.225	0.418
50代	0.114	0.319
60代	0.143	0.350
教育水準		
高校中退以下	0.170	0.376
高校卒†	0.329	0.470
短大卒・大学中退	0.265	0.442
大学卒	0.129	0.336
大学院以上	0.107	0.309
Observations	1022	

注）†はレファレンス変数を表す。

表 2　説明変数（人種）への影響（ロジットモデル）

	ウェートなし			ウエート付き		
	係数		標準誤差	係数		標準誤差
世帯所得（Ref. 25,000ドル以上）						
5,000ドル未満	0.302		0.526	0.049		0.532
5,000～7,000ドル未満	1.220	**	0.423	−0.004		0.557
7,000～10,000ドル未満	0.714		0.437	0.034		0.506
10,000～15,000ドル未満	1.026	**	0.325	0.020		0.393
15,000～20,000ドル未満	0.093		0.394	−0.082		0.364
20,000～25,000ドル未満	0.679	**	0.348	−0.004		0.355
職業威信の指数（Ref. 20未満）						
20～30未満	0.120		0.370	0.201		0.482
30～40未満	−0.587	#	0.355	0.010		0.446
40～50未満	−0.969	*	0.406	0.382		0.451
50～60未満	−0.907	*	0.444	0.061		0.486
60～70未満	−0.497		0.569	0.187		0.578
70以上	−1.119		0.832	−0.696		0.814
年齢（Ref. 20代）						
30代	0.724	*	0.333	−0.047		0.284
40代	0.623	#	0.362	0.083		0.302
50代	1.375	**	0.378	−0.184		0.389
60代	0.667	#	0.391	−0.211		0.369
教育水準（Ref. 高校卒）						
高校中退以下	−0.060		0.299	−0.137		0.331
短大卒（中退含）・大学中退	0.298		0.268	0.025		0.262
大学卒	−0.249		0.420	−0.118		0.362
大学院以上	−0.499		0.546	0.199		0.383
定数項	−2.571	**	0.447	−2.190	**	0.483
Observations	1022					

注）** <0.01, * <0.05, # <0.1

表 2 のウェートを付けていない結果から、次の式でウェートを算出する。

$$WT = \frac{X}{P(X=1|\mathbf{Z})} + \frac{1-X}{1-P(X=1|\mathbf{Z})}$$

なお、WT に各々の確率をかける必要がある。すなわち、WT*X を算出することであるが、X=1 の場合と X=0 の場合の両方を算出しなければならない。算出方法は次の式から求められる。

$$WT1 = \frac{X}{P(X=1|\mathbf{Z})}$$
$$WT2 = \frac{1-X}{1-P(X=1|\mathbf{Z})} \quad (8)$$

これらの式から求められた統計量を表しているのが表 3 である。このウェートが正しく作られたかどうか確認する必要があり、IPW の平均が 1 に近ければ正しいことになる。表 3 によると、その平均はほぼ 1 であり、算出方式は正しいことが証明されている。

表 3　ウェート付き説明変数の記述統計

	最小値	最大値	合計	平均値	標準偏差
人種	0	1.000	110.000	0.108	0.310
WT1	0	32.410	1043.340	1.021	3.744
WT2	0	2.088	1021.763	1.000	0.362
Observations			1022		

ただし、ここでは X と（1-X）をかけていない状態であり、次の式のようにそれらをかけると合計はサンプルサイズと同様になる（表 4）。

$$WT_{\mathrm{DFL}}=\frac{110}{1043.34}WT_1+\frac{(1022\text{-}110)}{1021.76}WT_2 \tag{9}$$

表 4　再調整されたウェートの記述統計

	最小値	最大値	合計	平均値	標準偏差
WT_{DFL}	0.238	3.417	1022.003	1.000	0.261
Observations			1022		

表 4 をみると、合計がサンプルサイズとほぼ同様であり、平均はちょうど 1 になっていることがわかる。DFL 法は IPW を用いて仲介変数との相関を無くすことが目的であり、実際このウェートを付けることで、両方の相関がなくなっているかどうか確認する必要がある。

表 5　仲介変数（Z）と説明変数（X）の予測値の相関

	ウェート調整前	ウェート調整後
相関係数	0.257**	-0.008
Observations	1022	

注）** <0.01, * <0.05, # <0.1

表 5 は仲介変数（Z）と説明変数（X）の予測値の相関を表しており、ウェートを付ける前と後を比較している。ウェートを付ける前の相関係数は 1%の有意水準で有意になっているが、ウェート調整後は有意性を無くなっており、両変数が独立になっていることを示す。すなわち、図 1 の X から Y への効果を、仲介変数を介さない純粋な効果として推定できる。

表 6　主観的な階層意識に対する回帰分析（OLS モデル）

	ウェートなし			ウエート付き		
	係数		標準誤差	係数		標準誤差
人種	−0.081		0.057	−0.104	#	0.056
世帯所得（Ref. 25,000ドル以上）						
5,000ドル未満	−0.357	**	0.089	−0.350	**	0.090
5,000～7,000ドル未満	−0.185	*	0.094	−0.213	*	0.093
7,000～10,000ドル未満	−0.185	*	0.085	−0.190	*	0.085
10,000～15,000ドル未満	−0.261	**	0.065	−0.270	**	0.066
15,000～20,000ドル未満	−0.280	**	0.060	−0.304	**	0.060
20,000～25,000ドル未満	−0.249	**	0.061	−0.259	**	0.061
職業威信の指数（Ref. 20未満）						
20～30未満	−0.120		0.078	−0.122		0.079
30～40未満	−0.035		0.070	−0.016		0.071
40～50未満	0.060		0.074	0.066		0.074
50～60未満	0.133	#	0.078	0.149	#	0.079
60～70未満	0.158		0.096	0.153		0.097
70以上	0.315	**	0.111	0.299	**	0.113
年齢（Ref. 20代）						
30代	−0.141	**	0.049	−0.140	**	0.049
40代	−0.043		0.053	−0.060		0.053
50代	0.091		0.064	0.088		0.064
60代	0.146	*	0.060	0.126	*	0.060
教育水準（Ref. 高校卒）						
高校中退以下	−0.064		0.054	−0.062		0.054
短大卒（中退含）・大学中退	0.063		0.045	0.068		0.045
大学卒	0.283	**	0.060	0.268	**	0.060
大学院以上	0.376	**	0.068	0.411	**	0.068
定数項	2.474		0.078	2.474		0.079
Observations	1022					

注）** <0.01, * <0.05, # <0.1

表 6 は X（人種）から Y（主観的な階層意識）への効果を推定した結果である。人種の効果をみると、ウェートを付ける前は有意ではなかったが、仲介要因の影響を取り除く作業を行ってからの結果は有意水準が 10%ではあるが、有意に変わっていることがわかる。これらの結果から、直接効果の $\mathrm{E}(Y_{1,\,direct}) - \mathrm{E}(Y_{0,\,direct})$、間接効果の $(\mathrm{E}(Y_1|x{=}1) - E(Y_0|x{=}0)) - (\mathrm{E}(Y_{1,\,direct}) - E(Y_{0,\,direct}))$ を計算すると、総効果は 0.2 で直接効果（説明されない差）は 0.104（52%）、間接効果（説明される差）は 0.96（48%）である。

3. Heckman 法と傾向スコアの IPW の併用（HP-IPW 法）

3.1 従来における仮定の緩和

HP-IPW 法は冒頭にも言及したように、2 つのメリットがある。①RCM のように交絡変数は Y の誤差項と相関があっても一致推定量を得られる、②説明変数と誤差項との相関があっても一致推定量を得られる。

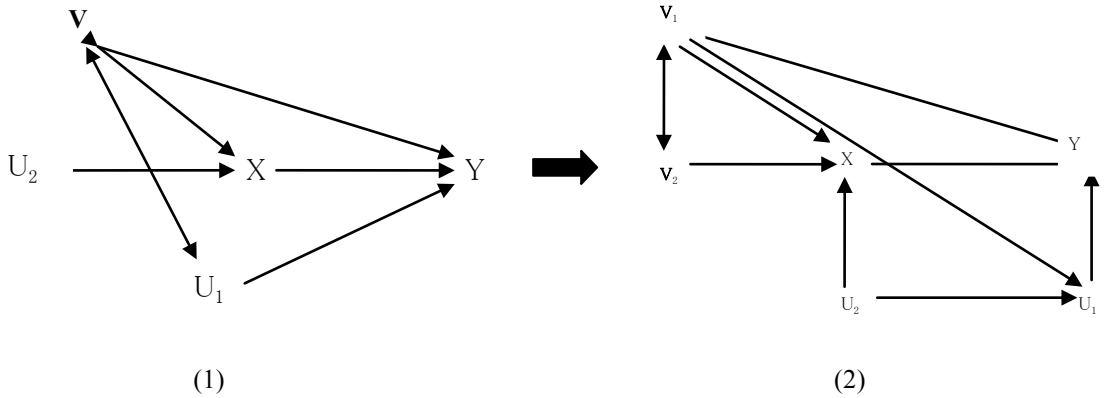


図 3　RCM と HP-IPW 法の模式図

図 3（1）は RCM の基本的な模式図を描いている。交絡変数 V は説明変数 X と被説明変数 Y と相関しても、X と V との相関を IPW 法で除去できれば、一致推定量を得られることが前節において見てきた通りである。これを X が内生変数の場合に適用したのが、図 3（2）である。V_1 は図 3（1）の V と同様な変数であるが、V_2 は X と Y との相関を除去するために用いる操作変数（instrument variable）である。しかし、本来操作変数を用いた Heckman 法で推定するためには、X と V_1 が独立でなければならないが、RCM の仮定を加えることで、両変数が独立でなくとも Heckman 法を用いることができる。したがって、IPW 法を用いて V_1 と X との相関を除去し、通常の Heckman 法を用いることになる。ところが、図 3（2）は一見 IPW 法を拡大しているように思われるが、ここでは U_1 と U_2 が正規分布であるという新たな仮定を加えなければならないため、単に RCM の拡大とはいえない。この仮定は Heckman 法と RCM を融合させるためであり、本来 RCM の仮定とは異なるものである。

3.2 HP-IPW 法の基本的なモデル

HP-IPW 法の基本的なモデルは次のように表すことができる。

$$Y_i=t_iY_i+\left(1-t_i\right)Y_0=\alpha(\mathbf{v}_1)+\beta(\mathbf{v}_1)t_i+u_1 \tag{10}$$

$$Y_{i,\,obs}=x_iY_i+\left(1-x_i\right)Y_0=\alpha(\mathbf{v}_1)+\beta(\mathbf{v}_1)x_i+u_1 \tag{11}$$

（10）式は観察されない場合であり、（11）式は観察される場合を表す関数である。ここで、t は処理を施しているか否かを表しており、X は処理の割り当て（treatment assignment）を表す。説明変数 V_1 は観察されるすべての交絡要因を含むと仮定する。加えて、潜在連続変数 $X^*>0$ であれば X は 1 を取ると仮定する。X^*は次のように定式化できる。

$$X_i^*=\gamma(\mathbf{v}_1,\mathbf{v}_2)+u_2 \tag{12}$$

V_2 は上述したように操作変数を表しており、V_1 をコントロールした場合に、Y に直接影響を与えていない変数である。なお、u_1 とu_2は正規分布であり、パラメータと相関していると仮定する。

$\mathrm{V}(u_1)=\sigma_1^2,\ V(u_2)=\sigma_2^2,\ COV(u_1,u_2)=\sigma_{12}$

そう仮定すると、次のような式が得られる。

$$\mathrm{E}(Y_1|x=1)=E_{\mathrm{v}|\mathrm{x}=1}\big(\alpha(\mathbf{v}_1)\big)+E_{\mathrm{v}|\mathrm{x}=1}\big(\beta(\mathbf{v}_1)\big)+E(u_1|u_2>-\gamma(\mathbf{v}_1,\mathbf{v}_2))$$

$$\mathrm{E}(Y_0|x=1)=E_{\mathrm{v}|\mathrm{x}=1}\big(\alpha(\mathbf{v}_1)\big)+E(u_1|u_2>-\gamma(\mathbf{v}_1,\mathbf{v}_2))$$

$$\mathrm{E}(Y_1|x=0)=E_{\mathrm{v}|\mathrm{x}=0}\big(\alpha(\mathbf{v}_1)\big)+E_{\mathrm{v}|\mathrm{x}=0}\big(\beta(\mathbf{v}_1)\big)+E(u_1|u_2<-\gamma(\mathbf{v}_1,\mathbf{v}_2))$$

$$\mathrm{E}(Y_0|x=0)=E_{\mathrm{v}|\mathrm{x}=0}\big(\alpha(\mathbf{v}_1)\big)+E(u_1|u_2<-\gamma(\mathbf{v}_1,\mathbf{v}_2))$$

そこから、ATT を次のような式から計算する。

$$\mathrm{ATT}=E(Y_1|x=1)-E(\mathrm{Y}_0|x=1)=E_{\mathrm{v}|\mathrm{x}=1}\big(\beta(\mathbf{v}_1)\big),$$

なお、措置前群の選択バイアス（pre-treatment selection bias）は次のようになる。

$$\begin{aligned}\mathrm{E}(Y_0|x=1)-E(\mathrm{Y}_0|x=1)=&\big\{E_{\mathrm{v}|\mathrm{x}=1}\big(\alpha(\mathbf{v}_1)\big)+E_{\mathrm{v}|\mathrm{x}=0}\big(\beta(\mathbf{v}_1)\big)\big\}\\&+E(u_1|u_2>-\gamma(\mathbf{v}_1,\mathbf{v}_2)-E(u_1|u_2<-\gamma(\mathbf{v}_1,\mathbf{v}_2)\end{aligned}\qquad(13)$$

（13）式の右辺の第 1 項の選択バイアスは IPW 法、すなわち V_1 をコントロールした傾向スコアの予測値を用いて除去することができる。そこから得られる処理群と制御群との差は次のように表すことができる。

$$\begin{aligned}\mathrm{E}^*(Y_1|x=1)-E^*(Y_0|x=1)=&E_{\mathrm{v}|\mathrm{x}=1}\big(\beta(\mathbf{v}_1)\big)\\&+\{E(u_1|u_2>-\gamma(\mathbf{v}_1,\mathbf{v}_2)-E(u_1|u_2<-\gamma(\mathbf{v}_1,\mathbf{v}_2)\}\end{aligned}\qquad(14)$$

ここで、E^*は IPW の期待値を表す。Johnson and Kotz（1972）によると(2)、

$$\mathrm{E}(u_1|u_2>-\gamma(\mathbf{v}_1,\mathbf{v}_2))=(\sigma_{12}/\sigma_2)\frac{\phi(Z_\mathrm{i})}{1-\Phi(Z_\mathrm{i})}$$

where $Z_\mathrm{i}=-\gamma(\mathbf{v}_1,\mathbf{v}_2)/\sigma_2$

なお、

$$\mathrm{E}(u_1|u_2>-\gamma(\mathbf{v}))p\big(u_2>-\gamma(\mathbf{v})\big)+E(u_1|u_2>-\gamma(\mathbf{v}))p\big(u_2<-\gamma(\mathbf{v})\big)=E(u_1)=0$$

であるため、次のような式が得られる。

$$\mathrm{E}(u_1|u_2> -\gamma(\mathbf{v})) = E(u_1|u_2> -\gamma(\mathbf{v}))p\big(u_2> -\gamma(\mathbf{v})\big)/p\big(u_2< -\gamma(\mathbf{v})\big)$$

$$= -(\sigma_{12}/\sigma_2)\frac{\phi(Z_\mathrm{i})}{1-\mathbf{\Phi}(Z_\mathrm{i})}\frac{1-\mathbf{\Phi}(Z_\mathrm{i})}{\mathbf{\Phi}(Z_\mathrm{i})} = -(\sigma_{12}/\sigma_2)\frac{\phi(Z_\mathrm{i})}{\mathbf{\Phi}(Z_\mathrm{i})}$$

したがって、

$$\mathrm{E}^*(Y_1|x{=}1) - E^*(Y_0|x{=}0) = E_{\mathrm{v}|\mathrm{x}=1}\big(\beta(\mathbf{v}_1)\big) + E(u_1|u_2> -\gamma(\mathbf{v})) - E(u_1|u_2< -\gamma(\mathbf{v}))$$

$$= E_{\mathrm{v}|\mathrm{x}=1}\big(\beta(\mathbf{v}_1)\big) + \frac{\sigma_{12}}{\sigma_2}\,\phi(Z_\mathrm{i})W_\mathrm{i}$$

where

$$W_\mathrm{i} = \frac{1}{1-\mathbf{\Phi}(Z_\mathrm{i})} + \frac{1}{\mathbf{\Phi}(Z_\mathrm{i})}$$

IPW をかけたデータから、Y を予測するための説明変数として、X と$\phi(Z_\mathrm{i})W_\mathrm{i}$を加えることにより、ATT の一致推定量を得られる。$\phi(Z_\mathrm{i})W_\mathrm{i}$は説明変数 V_1 と V_2 を、X を被説明変数とするプロビットモデルで回帰することにより得られる。V_1 は IPW を適用すれば X と独立になるが、プロビットモデルで回帰してから得られる$\phi(Z_\mathrm{i})W_\mathrm{i}$が有意でなければ除外することができる。

しかしながら、Heckman 法は Little（1985）が指摘しているように、X に対する有意な操作変数を見つけなければ、効率的な推定量を得ることは難しい。仮に観察される交絡要因がすべて X に含まれているとすれば、IPW 法を用いてから X の予測値は有意でなくなる。そこから、プロビットモデルで Heckman の調整項、$\phi(Z_\mathrm{i})W_\mathrm{i}$を推定すれば、ほぼ不変に近い値になるだろう。これが操作変数、V_2 が必要な理由である。ただし、上述のように良い操作変数を見つけなければ、有効な推定量を得ることは難しい。

分析の手順は以下のようなステップである。

1 段階：IPW を用いて X と V_1 との相関を除去する IPW 法を採用する。
2 段階：サンプルに IPW をかける。
3 段階：2 段階で作成されたデータに操作変数のみを導入したプロビットモデルで推定する。
4 段階：$\phi(Z_\mathrm{i})W_\mathrm{i}$を作成する。
5 段階：$\phi(Z_\mathrm{i})W_\mathrm{i}$を説明変数に導入し推定を行う。

3.3 HP-IPW 法を用いた実証分析の演習例

ここでは HP-IPW 法を用いて、実際実証分析の演習を行う。データは GSS の 1983 年と 91 年の調査を用いる。分析に用いる変数の記述統計量は表 7 に表している。被説明変数（Y）は年間教会出席頻度であり、「皆無＝0」、「年 1 回未満＝1」、「年 1 回＝2」、「年数回＝3」、「月 1 回＝4」、「月 2, 3 回＝5」、「ほぼ毎週＝6」、「毎週＝7」、「週 1 回以上＝8」に分けている。そして、内生変数（X）である離婚ダミーは「離別、別居＝1」、「有配偶、死別＝0」に分けている。なお、回帰分析に用いる変数の記述統計量は表 7 に表している。

まず、V_1 と X との相関を除去する作業は 2.2 節で行われた方法と同様である。表 8 ではウェートを付ける前後の推定結果を表している。ウェートを付ける前は性別、黒人、里親ダミーが有意になっている。この推定結果を基にウェートを計算する。

表 7　記述統計量（GSS83－91）

	平均	標準偏差
被説明変数		
年間教会出席頻度[1)]	3.965	2.612
内生変数		
離婚ダミー[2)]	0.211	0.408
仲介変数		
教育水準		
高校中退以下	0.167	0.373
高校卒†	0.371	0.483
短大卒・大学中退	0.238	0.426
大学卒	0.131	0.338
大学院以上	0.093	0.290
性別（女性＝１）	0.586	0.493
年齢		
20代†	0.245	0.430
30代	0.430	0.495
40代	0.325	0.469
人種		
白人†	0.876	0.330
黒人	0.101	0.301
その他	0.023	0.151
10代結婚ダミー[3)]	0.339	0.473
里親ダミー[4)]	0.065	0.246
Observations	1420	

注1）年間教会出席頻度は「皆無＝0」、「年1回未満＝1」、「年1回＝2」、「年数回＝3」、「月1回＝4」、「月2, 3回＝5」、「ほぼ毎週＝6」、「毎週＝7」、「週1回以上＝8」である。
注2）離婚ダミーは「離別、別居＝1」、「有配偶、死別＝0」である。
注3）初婚の経験が12～19歳の場合に1を取る。
注4）16歳時に親のどちらかが里親である場合に1を取る。
注5）†はレファレンス変数を表す。

表 8　内生変数（離婚ダミー）への影響（ロジットモデル）

	ウェートなし			ウエート付き	
	係数		標準誤差	係数	標準誤差
性別（女性＝１）	0.525	**	0.140	−0.005	0.141
人種（Ref. 白人）					
黒人	0.828	**	0.190	0.019	0.172
その他	0.419		0.400	0.008	0.383
里親ダミー	0.533	*	0.239	0.015	0.221
定数項	−1.798	**	0.118	−1.319	0.121
Observations	1420				

注）** <0.01, * <0.05, # <0.1

ウェートは（8）式から算出することができるが、ATT を求めなければならないため、X=1 の場合のみのウェートを算出する。その後、それが正しいかどうか診断する。表 9 をみると、WT1 と離婚ダミーとの合計と平均値が一致するため、ウェートの計算は正しいことがわかる。なお、(9) 式からサンプルサイズを調整することができる（表 10)。実際サンプルにウェートを付けて推定をすると、表 8 の右側にみられるように、有意性がすべてなくっており、相関が除去されたことがわかる。

表 9　ウェート付き説明変数の記述統計

	最小値	最大値	合計	平均値	標準偏差
離婚ダミー	0	1.000	300.000	0.211	0.408
WT1	0	1.000	300.000	0.211	0.408
WT2	0	4.074	1118.896	0.788	0.583
Observations			1420		

表 10　再調整されたウェートの記述統計

	最小値	最大値	合計	平均値	標準偏差
WT_{HP-IPW}	0.619	4.078	1419.996	1.000	0.417
Observations			1022		

表 11　Heckman の 2 段階推定

	被説明変数					
	離婚ダミー(Probit)			教会出席(OLS)		
	係数		標準誤差	係数		標準誤差
離婚ダミー				−0.756	**	0.169
性別（女性＝１）				0.623	**	0.148
年齢（Ref. 20代）						
30代	0.408	**	0.103			
40代	0.406	**	0.108			
人種（Ref. 白人）						
黒人				0.725	**	0.182
その他				0.260		0.404
10代結婚ダミー	0.275	**	0.078			
里親ダミー				−0.929	**	0.235
$\phi(Z_i)W_i$				1.117		0.848
定数項	−1.229	**	0.094	1.754		1.508
Observations	1420					

注１）** <0.01, * <0.05,

注２）ただし、$W_i = \frac{1}{1-\Phi(Z_i)} + \frac{1}{\Phi(Z_i)}$

次は Heckman の 2 段階推定を行う。表 11 は Heckman の手法から推定した結果を表している。まず、内生変数である離婚ダミーを被説明変数としてプロビットモデルによる回帰分析を行う。この回帰分析に用いた変数はすべて操作変数である(3)。この結果から$\phi(Z_i)W_i$を計算し、年間教会出席頻度関数に導入する。$\phi(Z_i)W_i$の係数が有意であれば、OLS では選択バイアスにより一致推定量が得られなくなることを意味するため、必ず 2 段階推定を行わなければならない。表 11 の結果ではそれが有意でないため、選択バイアスはそれほど大きな問題ではないと判断できる。

4. おわりに

本稿は 2013 年 9 月 2 日、3 日の両日間行われた JGSS 研究センター主催の統計分析セミナーの講義内容をまとめたものである。セミナーのテーマは 2009 年と 2011 年に引き続き、「傾向スコア」を用いた因果分析であった。RCM は従来の OLS において仮定された説明変数間の独立性仮定が緩和されたモデルであり、傾向スコアの逆確率を用いて、主に説明したい変数が交絡要因と相関があってもその相関を除去することで、効率的な推定量を得ることができるモデルである。

本稿では特に傾向スコア・ウェイティング法を用いて応用したモデルについて解説した。ここで取り上げた DFL 法は非線形モデルをベースにした場合の要因分解ができることで、経済学において幅広く使われてきた B-O 法を補う形のモデルともいえよう。そして、HP-IPW 法は RCM のメリットを Heckman 法に適用することで、従来の仮定を緩和させ交絡変数が内生変数や従属変数間に相関がある場合にも一致推定量が得られる方法である。この推定方法を用いて書かれた論文は、現在それほど見当たらず、今後の活用が期待される。

[Acknowledgement]

日本版 General Social Surveys（JGSS）は、大阪商業大学 JGSS 研究センター（文部科学大臣認定日本版総合的社会調査共同研究拠点）が、東京大学社会科学研究所の協力を受けて実施している研究プロジェクトである。

[注]

（1）X と Z の間には相関があるため、その予測は可能である。

（2）合せて、Heckman（1979）を参照されたい。

（3）操作変数は X と強い相関があり、Y との相関がないことが重要な条件である。これを簡易に行うには両変数にすべての説明変数を導入し、その影響を調べる方法がある。付表では、離婚ダミーに強い影響があり、年間教会出席頻度変数と相関がない変数は、年齢と 10 代結婚ダミーであり、この 2 つの変数を操作変数に用いたのである。

[参考文献]

Blinder, A., 1973, “Wage Discrimination: Reduced Form and Structural Variables,” *Journal of Human Resources* 8:436-55.

DiNardo, J., N. Fortin, and T. Lemieux, 1996, “Labor Market Institution and the Distribution of Wage,” *Econometrica* 64:1001-44.

Heckman, J. 1979, “Sample Selection Bias as a Specification Error,” *Econometrica* 47(1), 153-61.

Heckman, J., and Richard Robb, 1985, “Alternative Methods for Estimating the Impact of Interventions,” Heckman, James J., and Burton Singer [eds.], *Longitudinal Analysis of Labor Market Data*, Cambridge: Cambridge University Press.

Johnson, N., and S. Kotz, 1972, *Distribution in Statistics: Continuous Multivariate Distributions*, New York: John Wiley & Sons.

Little, R. J. A., 1985, “A Note about Models for Selectivity Bias,” *Econometrica* 53:1469-74.

Oaxaca, R., 1973, “Male-Female Wage Differentials in Urban Labor Markets,” *International Economic Review* 14:693-709.

林光, 2012「JGSS 統計分析セミナー2011－傾向スコア・ウェイティング法を用いた因果分析－」大阪商業大学 JGSS 研究センター編『日本総合的社会調査共同研究拠点 研究論文集』12:107-127.

三輪加奈・菅澤貴之, 2010「JGSS 統計分析セミナー2009－傾向スコアを用いた因果分析－」大阪商業大学 JGSS 研究センター編『日本総合的社会調査共同研究拠点 研究論文集』10:285-296.