

STATAを用いた離散型選択分析の手順と解説

2017年 12月
経済産業研・東京大学 戒能

1. 使用データ; 米国の州別社会統計、死刑制度有無及び死刑執行率

出典: McManus W. (1985) Journal of American Political Science

時点 1950年の横断面試料

対象 米国44州

目的 死刑制度の有無や運用はどのような社会状況の影響を受けて決定されていたか?
(死刑制度の執行率などの運用状況は死刑制度のない州では観察できない)

変数

(被説明変数)

x 州別死刑執行率

d 州別死刑制度有無ダミー ("有"=1, "無"=0)

(説明変数, ダミー変数以外は全て対数化し "l#"に変換済)

m 人口10万人当殺人犯罪認知件数

p 州別法廷殺人事件有罪率

t 殺人容疑者平均拘留日数

y 州別平均所得 (\$1,000-)

f 平均有業率 (= 1 - [失業率])

n 非白人人口比率

s 南部州ダミー

2. 適用する分析手法

- 離散型選択解析法

Logit/Probit モデル、Tobitモデル及び Heckman2段推計型モデルを用いた解法

- (上記に関連する検定手法)

3. 予備処理 - 説明変数の相関観察

(時系列解析などと異なり STATA上での特段の設定は不要)

(1) STATAの起動・更新

(2) データ投入

○ STATAのデータエディタを開く

> Data > Data Editor > Data Editor (Edit)

○ エクセルファイル(DISCSLEC01.xls) の L5~T49 を上記エディターに貼付け
"Treat first row as variable names" を選択

○ STATAのデータエディタを閉じる

(3) 変数間の相関の観察

> Statistics > Summaries, tables and tests > Summary and descriptive statistics > Summary statistics

> Statistics > Summaries, tables and tests > Summary and descriptive statistics > Correlations and covariances

← summary ; 変数の記述統計が表示される

← correlate ; 変数間の相関係数が表示される

相関の高い説明変数を複数用いることはできない(多重共線性)
; n(非白人人口率), s(南部州)は m(殺人件数) と高い相関有、
かつ相互にも高い相関有

← n,s は使用を留保

原典では両方用いた分析を行っているが明らかに不適切

← EXCELなどの表計算ソフトで変数の分布を直接観察することも有効

(記述統計に依存した分析は"危険")

○ (STATA ファイルの仮保存 ; "***.dat" ファイルが生成される)

4. 分析(1) - 死刑制度有無の選択分析

○ 二項選択型モデルを適用した分析を実施し死刑制度の有無を説明する社会変数を識別・特定する

← 一般に相関は因果を意味しないが、本件の場合「逆因果」の可能性はほぼ皆無
ex. 死刑制度があるから平均所得が高い/低いという可能性はあり得るか?

(1) 死刑制度ダミーの Logit/Probitモデル分析

○ Logit,Probitモデルを当てはめた結果を比較する(参考として線形モデルを加える)

> Statistics > Binary outcomes > Logistic regression

> Statistics > Binary outcomes > Probit regression

← logit d lm lp lt ly lf

← probit d lm lp lt ly lf

← reg d lm lp lt ly lf, robust

(当該データでは一段選択だが多段の場合 Multilevel mixed-effects models)

← 推計後 estat ic で AIC/BICを表示

	Logit	Probit	(参考) 線形
Log likelihood	-10.651	-10.406	
AIC	33.30	32.81	30.91
BIC	44.01	43.52	41.62

← 説明変数の有意性は logit/probit ではほぼ同じだが線形とは大きな差異あり
(AICなどの指標は低いが原理的に線型モデルの結果は用いるべきではない)

← 次の段階でのHeckman2段推計による分析に備え Probitモデルで有意であった変数を記録しておく

5. 分析(2) - 死刑執行率の分析

○ 不連続(切断)型・内生的選択型モデルを適用した分析を実施し死刑制度の運用状況を説明する社会変数を識別・特定する

(1) 死刑執行率の Tobit(Type-2)モデル分析

○ 不連続(切断)型モデルであるTobit(Type-2)モデルを当てはめた結果を考察する

> Statistics > Linear models and related > Censored regression >
tobit regression

← tobit x lm lp lt ly lf, ll(0) (ll(#)で切断点#を指定可)

← 推計後 estat ic で AIC/BICを表示 AIC = -54.55

← 結果が分析(1)の線型モデルと類似することを確認 (∵原理的にTobitは線形)

(2) 死刑執行率の Heckman2段推計型モデル分析

○ 内生的選択型モデルである Heckman2段推計モデルを当てはめた結果を考察する

> Statistics > Sample selection models > Heckman selection model(ML)

← heckman x lm ly lf, select(d= lm lt)

← heckman x lt ly lf, select(d= lm lt)

...

← 推計後 estat ic で AIC/BICを表示

・ x の回帰の説明変数を変更して試行錯誤する必要あり

(AIC最小化によるモデル選択手法が有効)

・ x の回帰の説明変数と d の選択の説明変数(d= ##)を複数重複させると多重共線性により解けない場合あり(縄田(1997)他)

← heckman x lm lf, select(d= lm lt) AIC = -67.75

		係数	p値	
(x 回帰説明変数)				
lm	(犯罪件数)	-0.007	0.587	--
lf	(有業率)	+0.795	0.000	***
const.	(定数項)	+0.603	0.000	***
(d 選択説明変数)				
lm	(犯罪件数)	+2.361	0.000	***
lt	(拘留日数)	+1.918	0.000	***
const.	(定数項)	-10.88	0.000	***

6. 結果解釈と考察

- ・ 死刑制度の有無の選択は人口当犯罪発生件数や平均拘留日数(凶悪度の指標)に強く影響されており、執行の選択は有業率の高さ(失業率の低さ)に強く影響されて決定されていると考えられる。
- ・ 死刑制度の有無の選択と死刑制度の執行の原因となる社会的背景は異なっていると考えられ、両者が人口当犯罪発生件数や平均拘留日数などの表面的な社会的指標によって内生的に選択されている可能性は低いと考えられる。
- ・ 但し不連続(切断)型モデルより内生的選択モデルの適合度から見て両者が完全に無関係とは言えず、人口当犯罪発生件数と平均所得など説明変数相互での相関の高さの影響や、説明変数に含まれていない他の社会的要因(マフィアなど組織的犯罪の有無や再犯者比率の相違など)の影響の可能性は否定できない。
- ・ 不連続(切断)型モデルより内生的選択モデルの適合度が高く、死刑制度のない州でも潜在的に執行の要因が存在する可能性は低い。つまり有業率が執行に影響するのは死刑制度が存在する州に固有の現象であると考えられる。
- ・ 法廷有罪率は死刑制度の有無や運用に全く影響していない。
- ・ これら一連の結果の背後には、凶悪犯の警官による現場射殺案件の存在、殺人容疑者特に再犯者の死刑判決率の高さなど他の要因が存在している点を考慮する必要がある。