mwp-392

Intro 1 - 選択モデルの解釈　【 評価版 】

選択モデル (choice models) というのはアウトカム変数が選択肢を表すようなデータを対象にしたモデルのことです。その場合、選択肢は意志決定者 (decision maker) — 個人、あるいは事業者 — によりいくつかの候補の中から選択されます。例えば朝食用シリアル (breakfast cereal) をいくつかのブランド品の中から消費者が選択するといったケースや、事業者が広告媒体としてテレビ、ラジオ、インタネット、新聞のいずれかを選択するといったケースがそれに該当します。

選択データに対するモデルとしては離散的な選択肢 (discrete choices) に対するモデルとランク付けされた選択肢 (rank-ordered alternatives) に対するモデルの 2 種類があります。各個人が単一の候補を選択する — 例えばシリアルを 1 箱購入する — といった場合、そのデータは離散的な選択肢に関するデータとなります。これに対し各個人が選択肢のランク付けを行う — 例えばシリアルを好きなものから順に並べる — 場合、そのデータはランク付けされたデータとなります。Stata はこれら 2 種類のモデル — 離散的選択モデルとランク順選択モデル — のフィットを行うためのコマンドを用意しています。

選択モデルは解釈が難しいという定評があります。モデルをフィットさせても、その係数値から直接関心の対象である検定の答えが得られることは稀です。係数値の符号から効果の方向は判別できます。また条件付きロジスティック回帰の場合であればオッズ比を求めることもできます。しかしそれ以上の情報を係数値から読み取ることはほとんど不可能であると言えます。

しかし心配には及びません。Stata には `margins` コマンド（`[R]` **margins** (*mwp-029*) 参照）が用意されているからです。ここでは `cmclogit` コマンドを例にとって `margins` の種々の用例を紹介しますが、同様の操作は他の `cm` コマンドの場合にも行えます。

1. 係数値の解釈

ここでは Example データセット travel.dta を用いて用例の紹介を行います。

```
. use http://www.stata-press.com/data/r16/travel.dta
. describe
```

[*1]

```
. describe

Contains data from http://www.stata-press.com/data/r16/travel.dta
  obs:           840
 vars:             9                          2 Dec 2018 13:27
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
choice          byte    %8.0g                 travel mode chosen
termtime        byte    %8.0g                 terminal time (0 for car)
invehiclecost   int     %8.0g                 in-vehicle cost
traveltime      int     %8.0g                 travel time
travelcost      int     %8.0g                 generalized cost of travel
income          byte    %8.0g                 household income
partysize       byte    %8.0g                 party size in mode chosen
id              int     %9.0g                 case identifier
mode            byte    %8.0g      travel     travel mode alternatives
-------------------------------------------------------------------------------
Sorted by: id
```

このデータセット中には 210 人の個人（変数 id によって識別される）が 2 つの都市間の移動手段として何を選択したかが記録されています。選択肢としては航空機、列車、バス、自動車の 4 種類があります。この移動手段は変数 mode の値によって、1: air、2: train、3: bus、4: car のように識別されます。参考までに id = 1 の個人についてのデータをリスト出力してみます。

```
. list choice termtime traveltime travelcost income partysize id mode
> if id == 1, abbreviate(10)
```

	choice	termtime	traveltime	travelcost	income	partysize	id	mode
1.	0	69	100	70	35	1	1	air
2.	0	34	372	71	35	1	1	train
3.	0	35	417	70	35	1	1	bus
4.	1	0	180	30	35	1	1	car

[*1] メニュー操作：File ▷ Example Datasets ▷ Stata 16 manual datasets と操作、Choice Models Reference Manual [CM] の Intro 1 の項よりダウンロードする。

データの構成としては mode の値ごとに別個の観測データ (observation) が用意されている点に注意してください。各個人がどの移動手段を選択したかは変数 choice によって特定できます。それぞれの個人は移動に要する時間やコスト、同行者の人数等を勘案して移動手段の選択を行っているわけです。

データセット上、移動時間に関連した変数は traveltime と termtime の 2 つに分割されているわけですが、ここではそれらを合算した値をもって移動時間と定義することにします。

```
. generate time = traveltime + termtime
```

選択モデルのフィットを行うためには最初に cmset を実行する必要があります。

- Statistics ▷ Choice models ▷ Setup and utilities ▷ Declare data to be choice model data と操作
- cmset ダイアログ: Cross-sectional choice model data: ● （デフォルト）
  Case ID variable: id
  Alternatives variable: mode

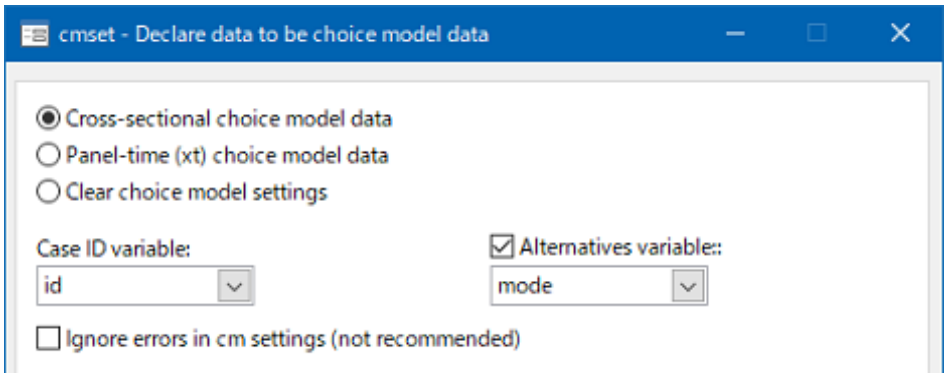

図 1 cmset ダイアログ

```
. cmset id mode

           caseid variable:  id
     alternatives variable:  mode
```

☞ Cross-sectional なデータ構成の場合、それぞれの case は複数の alternatives（選択肢）から構成されます。このとき個々の alternative が 1 つの observation（観測データ）を構成する形となる点に注意してください。

`cmset` が済んだところで `cmclogit` コマンド — `cm` データ用の `clogit` (conditional logistic) — を実行します。

> ☞ `clogit` と通常の `logistic/logit` との違いは、`clogit`（`[R]` **clogit** (*mwp-182*) 参照）がグループングされたデータを操作対象とする点にあります。

`cmclogit` の従属変数となる 2 値変数は `choice` です。一方、選択を左右する共変量としては移動に要する所要時間を表す `time`、年収を表す `income`、移動人数を表す `partysize` の 3 つを想定します。その際、`time` は alternative ごとに異なる値を取るのに対し、`income` と `partysize` は alternative によらず一定である点に注意する必要があります。後者のような case レベルの変数は構文上、`casevars()` オプションを使って指定することになります。

- Statistics ▷ Choice models ▷ Conditional logit (McFadden's choice) model と操作
- Model タブ: Dependent variable: `choice`
  Alternative-specific independent variables: `time`
  Case-specific independent variables: `income partysize`

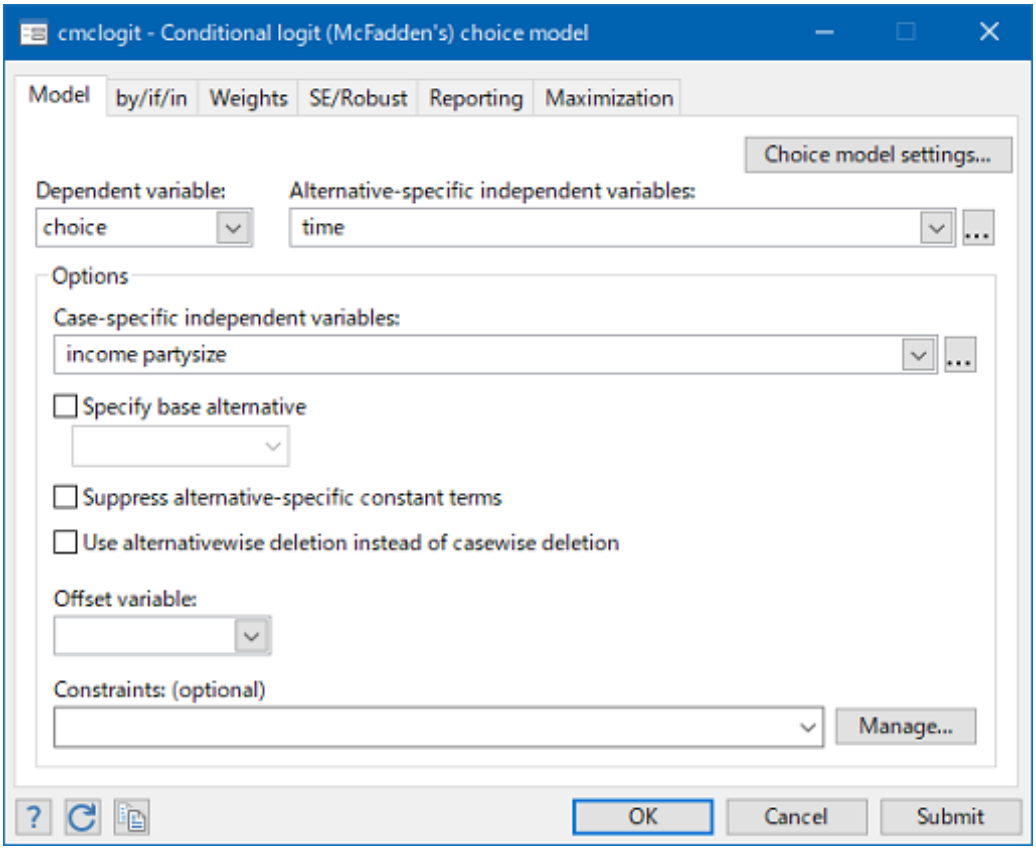

図 2 `cmclogit` ダイアログ - Model タブ

```
. cmclogit choice time, casevars(income partysize)

Iteration 0:   log likelihood = -249.36629
Iteration 1:   log likelihood = -236.01608
Iteration 2:   log likelihood = -235.65162
Iteration 3:   log likelihood = -235.65065
Iteration 4:   log likelihood = -235.65065

Conditional logit choice model                 Number of obs      =        840
Case ID variable: id                           Number of cases    =        210

Alternatives variable: mode                    Alts per case: min =          4
                                                              avg =        4.0
                                                              max =          4

                                                  Wald chi2(7)    =      71.14
Log likelihood = -235.65065                       Prob > chi2     =     0.0000
```

choice	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mode						
time	-.0041641	.0007588	-5.49	0.000	-.0056512	-.002677
air	(base alternative)					
train						
income	-.0613414	.0122637	-5.00	0.000	-.0853778	-.0373051
partysize	.4123606	.2406358	1.71	0.087	-.0592769	.883998
_cons	3.39349	.6579166	5.16	0.000	2.103997	4.682982
bus						
income	-.0363345	.0134318	-2.71	0.007	-.0626605	-.0100086
partysize	-.1370778	.3437092	-0.40	0.690	-.8107354	.5365798
_cons	2.919314	.7658496	3.81	0.000	1.418276	4.420351
car						
income	-.0096347	.0111377	-0.87	0.387	-.0314641	.0121947
partysize	.7350802	.2184636	3.36	0.001	.3068993	1.163261
_cons	.7471042	.6732971	1.11	0.267	-.5725338	2.066742

この結果から何が導けるでしょうか？ time の係数値は負であるため、ある移動手段が選択される確率は移動時間の増加に伴い減少することがわかります。列車という選択肢の場合、income の係数値は負の値です。従って年収が増すにつれて、ベースの選択肢である航空機ではなく列車が選択される確率は下がることを意味します。自動車という選択肢の場合、partysize の係数値は正の値です。従って移動人数が多くなるほど航空機ではなく自動車が選択されやすくなることがわかります。

2. margins を用いた推論

cmclogit や他の cm コマンドからの出力はわずかな情報をもたらしてくれます。しかしそれは通常、研究者が抱く種々の問いに答えられるものではありません。上記の例について言えば次のようなことを知りたいわけです。

Q1: 航空機を選択する個人の割合はどれくらいか？

Q2: 年収による効果はどれほどか？年収が$30,000 から$40,000 に、あるいは$40,000 から$50,000 に増えたときに、自動車を選択する確率はどれだけ変化するであろうか？列車を選択する確率についてはどうか？

Q3: セキュリティが強化され、空港での待ち時間が 1 時間増えたとするなら、それぞれの移動手段を選択する確率にどのような影響が及ぶであろうか？

margins を使うとこれらやその他の問いに対する答えを得ることができます。

2.1 選択確率の期待値

評価版では割愛しています。

2.2 連続的な共変量の効果

評価版では割愛しています。

2.3 離散的な共変量の効果

評価版では割愛しています。

2.4 選択肢に固有な共変量の効果

評価版では割愛しています。

3. margins によるその他の推論

評価版では割愛しています。

■