Stata+α「データの整理」

2013 年 12 月

前回はテキストファイルをインポートし、条件に見合うデータだけを残すために keep コマンドを利用しました。keep コマンドと同じ様にデータの選別を行う時に利用するコマンドの中で重要なものが drop コマンドです。この 2 つのコマンドについては Getting Started Manual 日本語版の中でも例題を用いて詳しく解説されています。

http://www.lightstone.co.jp/stata/manual_ja.html

ここでは Stata のオンラインヘルプの例を用いてその用法を簡単に確認しておきます。以下は、

.help drop

とした時に表示されるオンラインヘルプの Examples の情報です。

DROP コマンドについて

最初に drop コマンドの用法を説明します。これは、条件を満たすデータを削除する場合に利用するコマンドです。

セットアップ。census.dta というサンプルファイルを開きます。

.sysuse census,clear

データの内容(変数名、データの種類など)を確認します。

.describe

変数名の先頭 3 文字が pop で始まる変数を列ごと、すべてメモリー上から削除します。

. drop pop*

削除後のデータセットの内容を確認します。

. des

変数 marriage と divorce を削除します。

. drop marriage divorce

変数 medage が 32 よりも大きなデータについて、行単位でデータを削除します。

. drop if medage > 32

この時点でデータエディタには 5 つの変数、47 行のデータが残っています。

	state	state2	region	medage	death
1	Alabama	AL	South	29.30	35,305
2	Alaska	AK	West	26.10	1,604
3	Arizona	AZ	West	29.20	21,226
4	Arkansas	AR	South	30.60	22,676
5	California	CA	West	29.90	186,428
6	Colorado	CO	West	28.60	18,925
7	Connecticut	CT	NE	32.00	26,005
8	Delaware	DE	South	29.80	5,123
9	Georgia	GA	South	28.70	44,230
10	Hawaii	HI	West	28.40	4,849
11	Idaho	ID	West	27.60	6,753
12	Illinois	IL	N Cntrl	29.90	102,230
13	Indiana	IN	N Cntrl	29.20	47,300

図 1.並べ替えの前のデータ

次のコマンドは region(地域)ごとにデータを並べ替えます。この時点でデータセットには 5 つの変数が残っています。region で並べ替えを行ったときに、例えばその値が NE 同士であるものの中で、どれを一番上にするか、という問題があります。ここでは、単純に元のデータセットの順番(行番号)の大小関係を使って並べ替えを行います。つまり、

.sort region, stable

stable オプションは行番号の大小関係を維持して、全体を並べ替えるオプションです。

	state	state2	region	medage	death
1	Connecticut	CT	NE	32.00	26,005
2	Maine	ME	NE	30.40	10,768
3	Massachusetts	MA	NE	31.20	54,919
4	New Hampshire	NH	NE	30.10	7,594
5	New York	NY	NE	31.90	171,769
6	Rhode Island	RI	NE	31.80	9,300
7	Vermont	VT	NE	29.40	4,587
8	Illinois	IL	N Cntrl	29.90	102,230
9	Indiana	IN	N Cntrl	29.20	47,300
10	Iowa	IA	N Cntrl	30.00	26,348
11	Kansas	KS	N Cntrl	30.10	21,910
12	Michigan	MI	N Cntrl	28.80	75,102
13	Minnesota	MN	N Cntrl	29.20	33,412
14	Missouri	MO	N Cntrl	30.90	49,329
15	Nebraska	NE	N Cntrl	29.70	14,465

図 2.stable オプションで並べ替えたデータ

ここで、各 region の先頭のデータだけを削除してみましょう。

. by region: drop if _n == 1

	state	state2	region	medage	death
1	Maine	ME	NE	30.40	10,768
2	Massachusetts	MA	NE	31.20	54,919
3	New Hampshire	NH	NE	30.10	7,594
4	New York	NY	NE	31.90	171,769
5	Rhode Island	RI	NE	31.80	9,300
6	Vermont	VT	NE	29.40	4,587
7	Indiana	IN	N Cntrl	29.20	47,300
8	Iowa	IA	N Cntrl	30.00	26,348
9	Kansas	KS	N Cntrl	30.10	21,910
10	Michigan	MI	N Cntrl	28.80	75,102

図 2.並べ替えて 1 行目のデータを削除したもの

データが 4 つ少なくなっています。これらの操作におけるポイントは次の 2 つです。

1. by カテゴリ変数とすることで、カテゴリ変数ごとのデータ処理が可能になる
2. 行番号は_n で指定できる

1 番の機能は特に重要です。Stata でこのデータ処理機能が利用できるようになれば、もう初心者ではありません。2 番の_n の事をシステム変数と呼びます。システム変数にはこれ以外にもいくつか用意されていますが、敢えて覚えておくとすれば_N です。これはデータセットにおけるデータの個数(行数)です。ですから、仮に次のようなコマンドを実行すると、

```
.sysuse auto,clear
.drop if _n != _N
```

auto.dta のデータの個数は 74 個ですから、1 行目から 73 行目までのデータは削除され、74 行目のデータだけが残ることになります。

最も単純な利用方法として、変数を削除する場合には次のように利用します。ここでは複数の変数を削除するので、変数名をスペース区切りで入力します。

```
.drop mpg weight
```

注意) drop や次の keep において、いわゆる UNDO 機能はありませんのでご注意ください。

KEEP コマンドについて
これは drop コマンドとは逆に、条件を満たすデータだけを残す場合に利用するコマンドです。

```
.sysuse auto,clear
.keep in 1/2
.describe
```

サンプルデータを読み込み、最初の 2 行だけのデータを残します。そして、ファイルに含まれるデータの情報をみると、obs(観測値)の個数が 2 になっていることが分かります。

```
. drop _all
```

clear コマンドと同じことになります。

最後にパネルデータにおける欠損値の確認方法を紹介します。ここでも利用するコマンドは 10 月に紹介した misstable コマンドです。

.webuse abdata,clear
.xtset
.codebook year
.xtdescribe
.misstable summarize

最初にサンプルデータ abdata をインターネット経由で読み込みます。次の xtset はパネルデータとして設定情報を確認するコマンドです。

```
panel variable:  id (unbalanced)
 time variable:  year, 1976 to 1984
         delta:  1 unit
```

1976 年から 1984 年までの 9 年間の調査データであることが分かります。個体を識別する ID が変数 id で、時間を示す変数が year です。途中、調査年に抜けがないか、codebook コマンドで確認できます。
1976 年から連続して 84 年まで調査が実施されています。
xtdescribe はデータセットがパネルデータとして設定されている場合に利用できるコマンドで、つぎのような情報を出力します。

```
          range:  [1976,1984]
 unique values:  9

    tabulation:  Freq.  Value
                    80  1976
                   138  1977
                   140  1978
                   140  1979
                   140  1980
                   140  1981
                   140  1982
                    78  1983
                    35  1984
```

右側の 1 は当該年に全てのデータが揃っており、欠損値が存在しないことを示しています。逆に「.」は欠損値の存在を示しています。1 行目の意味は 1 が 7 個、点が 2 個並んでいます。つまり、1983 年と 84 年には欠損値が存在することが分かります。

Freq.	Percent	Cum.	Pattern
62	44.29	44.29	1111111..
39	27.86	72.14	.1111111.
19	13.57	85.71	.11111111
14	10.00	95.71	111111111
4	2.86	98.57	11111111.
2	1.43	100.00	..1111111
140	100.00		XXXXXXXXX

次回は取り込んだデータに対して統計量を求めるコマンドと、その分布を視覚的に確認するためのグラフ作成機能についてご案内します。

■