Stata+α「データ分析の前に」

2013 年 10 月

はじめて Stata を利用するお客様のために Stata の操作方法と、設計思想をご理解いただくための講習会「入門コース」をご用意しております。4 時間という短い時間の中で基本的な操作方法を習得していただき、Excel をはじめとするソフトウェアと比べ、何が違うのか、という点をご理解いただくような講習会の設計となっております。

ただ、実際にデータの分析を行う場合、まずはデータの内容をよく理解する必要があります。いわばデータ分析の下準備とも言える作業がありますが、「入門コース」ではあまり触れることができませんので、何回かに分けてデータの下準備で利用するコマンドを紹介いたします。

■データの構成を理解するための基本的なコマンド

describe
codebook
misstable

Stata の入門コースをまだ受講していない方でも、次に示すコマンドの通り、コマンドウィンドウに入力していただければ上記のコマンドの機能についてご理解いただけます。

■describe コマンド

まずはデータを読み込んで describe コマンドを実行してみましょう。ここでは Stata がインターネットに接続できる環境にあるものとします(先頭のピリオドを入力する必要はありません)。

```
.webuse  studentsurvey
.describe
```

サンプルデータ studentsurvey.dta をインターネット上のサーバから取得します。そして describe コマンドにより変数のデータ形式を Result ウィンドウに表示します。describe コマンドの d に下線をつけているのはコマンドの省略形であることを示すものです。つまり、1 文字、d だけを入力しても describe と同じ出力を得ます。

```
Contains data from http://www.stata-press.com/data/r13/studentsurvey.dta
  obs:           125                          Student Survey
 vars:             7                          10 Mar 2013 08:31
 size:         4,375
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
m1              float   %9.0g                 teaching
m2              float   %9.0g                 academics
age             float   %9.0g
female          float   %9.0g
dept            float   %9.0g
offcampus       float   %9.0g
comment         str11   %11s
-------------------------------------------------------------------------------
Sorted by:
```

読み込んだデータの変数名(variable name)と表示するケタ数(%の後ろの数字)が分かります。display format の最後の 1 文字が s の場合、変数 comment は文字列変数であることを示しています。その他の変数は%9.0g なっており、9 桁(小数点を含む)表示で、小数点以下のケタ数はゼロに設定されています。%g の場合、数字が設定した桁数に比べ大きすぎる(または小さすぎる)場合、自動的に E(10 の累乗)を使った指数表記に切り替えます。

■codebook コマンド

次に codebook コマンドを実行してみましょう。

.codebook

```
female
-------------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  3/125

            tabulation:  Freq.  Value
                           58  0
                           64  1
                            3  .
```

codebook コマンドは個々の変数の情報を表示します。変数 female は 0 と 1 の 2 つの値からなり、125 個のデータの中に 3 つの欠損値が存在します。

describe コマンドと codebook コマンドでデータの形式と、個別の変数の情報を読み取ることができます。次は misstable コマンドで欠損値について調べましょう。

■misstable コマンド

.misstable summarize

			Obs<.			
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
age	3		122	5	17	21
female	3		122	2	0	1
dept	9		116	4	1	4

Stata ではピリオドで欠損値を示します。7 個の変数のうち、age と female には各 3 つ、dept には 9 個の欠損値のあることが分かります。

欠損値を有する 3 つの変数に絞って、欠損値のパターンを調べてみましょう。

.misstable patterns

Missing-value patterns
(1 means complete)

Percent	Pattern 1	2	3
93%	1	1	1
5	1	1	0
2	0	0	0
100%			

Variables are (1) age (2) female (3) dept

125 個のデータのうち、3 つとも揃っているのが 93%、age と female にはデータあるが、dept だけが欠損値になっているもの(行)が 5 つ、そして 3 変数とも欠損値に成っている行が 2%あることが分かります。

実線で囲んだ 5%部分のデータを表示させる場合には次のようにします。

.list age female dept if age<. & female<. & dept==.

	age	female	dept
7.	18	1	.
12.	19	1	.
14.	20	0	.
16.	20	0	.
45.	18	0	.
52.	19	0	.

6 個(6/125=0.048)のデータセットが先のコマンドの条件に該当します。

□欠損値に値を代入する

Female の値によって dept はどのような値をとるのか、次のコマンドで調べてみましょう。

.tabstat dept,by(female)

female	mean
0	2.351852
1	2.612903
Total	2.491379

そこで、female=0 なら dept には 2 を、female=1 なら 3 を dept の欠損値に代入することにします。

.gen dept2=dept

.replace dept2 = cond(female,3,2) if dept2==.

最初にオリジナルデータ dept を直接操作することを避けるために、dept2 を複製します。そして、replace コマンドを使って dept2 の欠損値に 3 または 2 を代入します。

cond(female,3,2)は female の値が真(つまり 1)なら 3 を dept2 に代入し、偽(0)なら 2 を代入するというコマンドです。If 文を使えば同じ命令を複数行で書くこともできます。

最後に代入がうまくいったか、確認しましょう。

.list age female dept dept2 if age<. & female<. & dept==.

	age	female	dept	dept2
7.	18	1	.	3
12.	19	1	.	3
14.	20	0	.	2
16.	20	0	.	2
45.	18	0	.	2
52.	19	0	.	2

今回は取り込んだデータの内容を確認する基本的なコマンドについて紹介しました。次回は、テキストデータを Stata に取り込む際に利用するコマンドについてご紹介します。