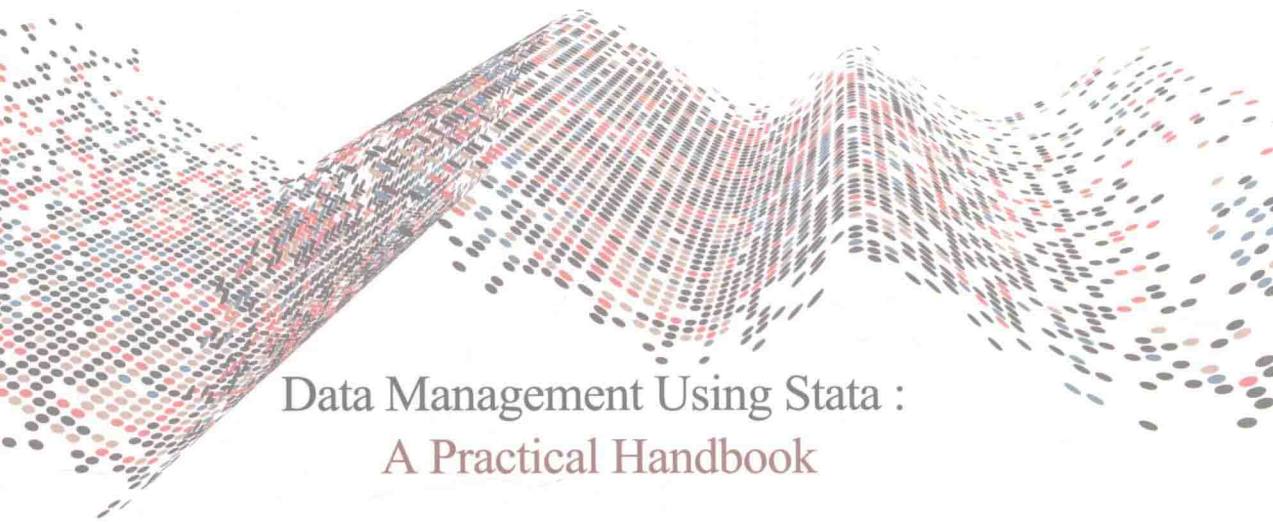
数据管理实务译丛

中国人民大学中国调查与数据中心组编

STATA 环境下的数据管理实务手册

【美】迈克尔·N·米歇尔（Michael N. Mitchell）/ 著

唐丽娜 / 译



Data Management Using Stata :
A Practical Handbook

中国人民大学出版社

数据管理实务译丛
中国人民大学中国调查与数据中心组编

STATA 环境下的数据管理实务手册

【美】迈克尔·N·米歇尔（Michael N. Mitchell）/ 著
唐丽娜 / 译

Data Management Using Stata :
A Practical Handbook

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

Stata环境下的数据管理实务手册/（美）米歇尔著；唐丽娜译．—北京：中国人民大学出版社，2013.11

ISBN 978-7-300-18239-1

Ⅰ.①S… Ⅱ.①米…②唐… Ⅲ.①数据管理-手册 Ⅳ.①TP274-62

中国版本图书馆CIP数据核字（2016）第068994号

Stata环境下的数据管理实务手册

[美] 迈克尔·N·米歇尔 著

唐丽娜 译

Stata Huanjingxia de Shuju Guanli Shiwu Shouce

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010-62511242（总编室）	010-62511770（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京昌联印刷有限公司		
规　　格	185 mm×235 mm　16开本	**版　　次**	2016年5月第1版
印　　张	27插页1	**印　　次**	2016年5月第1次印刷
字　　数	522 000	**定　　价**	68.00元

前　言

原始数据和统计分析之间有一道沟，这道沟就是数据管理，数据管理通常包含大量繁琐费劲的工作，这些工作就摆在你和你的数据分析之间。我发现数据管理经常包含了数据分析中那些最具挑战性的方面。我一直想写一本关于如何用 Stata 来解决数据管理工作中这些繁琐费力工作的书。

我想写这样一本书的其中一个原因是：由此能够展示一下 Stata 在数据管理中的作用。有时人们认为 Stata 的优势仅在于其统计上的功能。我个人使用 Stata 已经有 10 多年了，给学生教授 Stata 这门课程也有 10 多年了，其在数据管理方面的强大功能和简单用法总是让我印象深刻。拿命令 reshape 来说，用这个简单的命令很容易就可以把一个宽数据文件转变为长数据文件（更多示例参阅 8.3 节），反之亦然。而且，从某种程度上说，reshape 命令是 Stata 的一名使用者发明的，这表明：Stata 用户编写的程序进一步增强了 Stata 在数据管理方面的功能，而且用户编写的程序很容易下载（10.2 节中有说明）。

本书的每部分几乎都自成一体，每部分讲的都是如何在 Stata 中完成一项特定的数据管理任务。以 2.4 节为例，这节介绍了怎样把一个逗号分隔的数据文件读入 Stata 里。读者不必逐页阅读本书，我鼓励你在最相关的主题间来回跳读。

数据管理是一项重要的（有时甚至是令人生畏的）工作。我以一种非正式的形式来写这本书，就像是我们一起坐在电脑前，我在向你展示关于数据管理的一些小知识。本书的目的是帮助读者轻松快速地学到读者在数据管理中需要的 Stata 知识和技巧。如果读者需要进一步的帮助指导来解决自己遇到的问题，那么 10.3 节讲了很多读者可用的在线 Stata 资源。特别推荐的是 Statalist Listserver，通过它，读者可以接触到来自世界各地的 Stata 用户提供的知识资源。

如果读者想给我一些意见或建议，我非常愿意倾听。可以给我发邮件，我的电子邮箱地址是：MichaelNormanMitchell@gmail.com，也可以访问我的网站，地址是：http://www.MichaelNormanMitchell.com。于我而言，写这本书既是一种挑战也是一种乐趣，希望读者会喜欢它。

加利福尼亚州西米谷市　Michael N. Mitchell

2010 年 8 月

目　录

第一章　入门介绍

有人说收集数据就像收垃圾一样：收集之前就应该想好怎么处理它。

——罗素·福克斯，马克思·哥白尼和罗伯特·虎克

1.1 本书的使用

书如其名，这是一本关于用 Stata 来管理数据的操作手册。作为一本操作手册，也就没有必要一定遵循某种顺序来阅读每个章节。书中不仅各章各自独立，各章中的大多数小节也相互独立。书中每一部分都关注某一个特定的数据管理任务，且提供了相应的示例来展示如何在 Stata 中实现这一特定的数据管理任务。我认为本书至少有两种使用方式。

读者可挑选其中一章，比如第三章“数据清理”，通过阅读这一章来掌握一些有关如何清理和准备数据的新知识点或小技巧。这样，当下次需要清理数据时，就可以直接使用之前学到的这些相关知识点，如果需要的话，也可以再快速浏览一下相关章节。

或者，面对之前从来没有做过的数据任务（也许之前做过，但是已经很长时间没有操作过了），希望能够快速获得帮助。例如，要读入一个用逗号作为分隔符的数据文件。这时候，拿起这本书直接翻到第二章“读入数据”的 2.4 节，这节介绍了如何读入以逗号和制表符作分隔符的数据文件。根据这节中的示例，就能把逗号分隔的数据文件读入 Stata，然后继续你的数据处理工作。

当阅读这本书的时候，读者会发现本书的每个章节都是为解决某个具体问题而设计的，但千万不要迷失在一些附属或难懂的细节之中。如果发现自己需要了解一些更深的知识，本书的每个小节也列出了一些 Stata 帮助文件中的相关参考文件，这些参考中包含了更多的知识。如果读者用的是 Stata 11.0 版本，那么这些帮助文件中还包含了在线参考手册的链接。由于本书是按照实际数据管理中会遇到的不同任务来组织的，而 Stata 的参考手册是根据命令来组织的，因此我希望本书能够帮助读者将手头上要处理的数据管理任务和手册中与这些任务相关的对应条目联系起来。从这个角度来看，本书并不是 Stata 参考手册的竞争者，相反是它们的使用指南。

建议读者自己去操作和运行书中的示例。和被动学习（比如仅阅读本书）相比，实际操作让你进入一种主动学习的状态。如果读者主动在 Stata 中敲入命令，查看运行结果，自己试验同一命令的变体，那么相信这时你对知识的理解和被动学习相比，会更好且更深入。

为了方便读者重复操作书中列出的示例，书中所有的数据都可以从网络上直接下载。通过在 Stata 中键入下面的命令，将书中涉及的所有数据直接读入 Stata 的当前工作目录下：

```
. net from http://www.Stata-press.com/data/dmus①
. net get dmus1
. net get dmus2
```

执行完这些命令后，就可以使用这些数据了，比如：要用数据 wws.data，只需键入如下命令即可：

```
. use wws
```

书中每个小节都是独立的，因此可以在每个小节开始时键入相关命令，直接重复运行本节中的示例。有时甚至可以在某个小节的中间重复运行一个示例，但并不是在所有的小节中都能这么操作。此时，需要重新回到这一小节的开头来重复这些示例。尽管大部分的章节是独立的，但有些部分仍是建立在之前章节的基础上。即使在这种情况下，数据也是可用的，以便读者能从任何一个给定小节开头部分来运行这些示例。

尽管书中讲的所有示例都可以通过点击 Stata 菜单中的相关条目来实现，但本书的重点是使用 Stata 的命令行进行操作。但有一点需要说明：Stata 里有两个非常方便的交互界面/点击的功能，即使一些以写命令为主的用户（包括我自己）也会发现这些功能很有用。数据编辑器（Data Editor，2.10 节会介绍）是一个非常有用的用来把数据录入 Stata 的交互界面。这节中还介绍了变量管理器（Variable Manager）的使用。虽然这是在给一个新创建的数据添加标签的背景中介绍变量管理器，但它对修改（或增加）一个既存数据的标签同样非常有用。

需要说明的是本书是在 Stata 11.0 下写成的。书中大部分示例在 11.0 之前的版本中也同样有效。但是，有些示例在 11.0 之前的版本下是无效的，最明显的是第六章中那些用来讲解数据合并的例子。

这就提出了一个问题，读者要一直保持自己所用的 Stata 是最新的，这也是一个不错的练习。要想证实你的 Stata 是最新版并获取所有免费更新，输入下面这个命令：

```
. update query
```

然后根据提示操作。升级完成后，可键入命令 help whatsnew 来查看刚刚都更新了些什么以及此前 Stata 的更新记录。

在下载完所需数据并实现 Stata 的全面升级后，便可投入到本书的学习中，并亲自操作书中的所有示例。在此之前，希望读者能看完下面一节，它是对本书的总体介绍，能够帮助读者选择可能是你想最先阅读的章节。

---

① 本书英文版初版于 2010 年，原始随书数据发布在 Stata 的官方网站上。但作者后来对数据做了更新，所以本书中有些输出结果和原书会有不同。——译者注

1.2 本书的概要

本书每一章都包含了一个不同的、和数据管理有关的主题，每一章都非常独立。本书各章之间的先后顺序和传统书本中的不一样，传统的书需要从头读到尾。也许读者学习本书的大部分内容都不是按照书中呈现的顺序进行，而是按一种不同顺序来学习。我想让读者先快速对本书有整体的了解，以便能以自己喜欢的顺序来学完书中的大部分内容。

本书共 11 章，包括介绍章节（第一章），主体章节第二至十章，以及一个附录。

接下来的四章，第二至五章，讨论的是基本主题，其中包含了所有数据管理项目中都会遇到的问题：读取和录入数据、数据清理、给数据加标签以及创建变量。之所以将这些主题放在前面来讲，是因为我认为它们都是数据管理中最常见的主题；把它们放在前面还有一个原因：它们都是最明确且最具体的主题。

后面的三章，第六至八章，讨论的是在很多（但不是所有）数据管理项目中都会出现的问题：合并数据、处理分组的观测值以及改变数据形状。

第九章讲的是数据管理编程。虽然这章中涉及的主题对很多（不是所有的）数据管理项目而言很常见，但相对前面的二至五章中讨论的主题而言，它们更深入、更高级。这章讲的是如何构建数据分析以使其能够被循环使用，并介绍了很多用来处理重复性任务的快捷编程方法。

第十章主要是一些扩展内容，介绍了怎样为本书获取一些在线资源，如何寻找和安装其他 Stata 用户编写的程序，并推荐了一系列作为补充的在线资源。如果更早地看完这一章，或许你会发现这些信息很有用。

附录 A 列出了 Stata 操作中的一些基本要素。和前面的章节不一样，这些要素是分散的，并且不是关于某一个特定的数据管理任务的，但是它们无处不在，本书通篇都会经常涉及。前面的几个章节会经常涉及附录中的各节，在附录中给每一个要素提供一个解释，这样就不需要在它们每次出现的时候都重复这些解释了。附录包含的主题有：注释、逻辑表达式、函数、if 和 in、缺失值以及变量列表。把这一章放在最后，是为了方便读者在需要时能够快速翻到这里。也许你会发现，和反复回到附录相比，先读完附录并让自己熟悉附录中的这些要素会更容易些。

下一节介绍并解释了一些选项，这些选项可以和贯穿全书的命令 list 一起使用。

1.3 列出书中的观测值

本书主要用各种示例向读者展示 Stata 中数据管理命令的工作原理。我更倾向于用一个简单的示例向读者展示怎样使用一条命令，而不是用很多文字来解释这条命令。为此，我会经常使用 list 命令来讲解其他命令的作用。命令 list 默认的输出结果并不总是如我们希望的那样清楚明了。有时我会在 list 命令的后面加入一些选项让运行结果尽量清晰。我用这节来讲解这些选项并解释为什么全书都会用到它们，而不是每次这些选项出现时都去解释它们。

在第一组示例中，使用数据 wws. dta，它包含 2 246 个虚构的、关于女性及其工作的观测值。

```
. use wws
(Working Women Survey)
```

对包含了很多观测值的文件，列出其中一部分观测值非常有用。我经常用 in 来显示在一个数据中选中的观测值。在下面的示例中，列出了第 1～5 个观测值，显示了变量 idcode，age，hours 和 wage。

```
. list idcode age hours wage in 1/5
```

	idcode	age	hours	wage
1.	5159	38	38	7.15781
2.	5157	24	35	2.447664
3.	5156	26	40	3.824476
4.	5154	32	40	14.32367
5.	5153	35	35	5.517124

有时变量名太长，命令 list 就会把变量名进行缩写。这样列表就会更紧凑，但这样也会让被缩写的标题更难以理解。例如，下面的列表显示了前 5 个观测值中的变量 idcode、married、marriedyrs 和 nevermarried。请注意变量 marriedyrs 和 nevermarried 是如何被缩写的。

```
. list idcode married marriedyrs nevermarried in 1/5
```

	idcode	married	marrie~s	neverm~d
1.	5159	0	0	0
2.	5157	1	0	0
3.	5156	1	3	0
4.	5154	1	2	0
5.	5153	0	0	1

在缩写变量时，可以用选项 abbreviate()来指定命令 list 可用的最少字符数。例如，指定 abbreviate(20)表示所有的变量都不能被缩写到小于 20 个字符。在本书中，这个选项缩写为 abb()［例如下面的 abb(20)］。这里用这个选项就能让所有变量名都完整地显示出来。

```
. list idcode married marriedyrs nevermarried in 1/5, abb(20)
```

	idcode	married	marriedyrs	nevermarried
1.	5159	0	0	0
2.	5157	1	0	0
3.	5156	1	3	0
4.	5154	1	2	0
5.	5153	0	0	1

如果在一行中要显示的变量列表太长，列表会在页内自动换行。如下所示，这种列表很难读懂，因此在本书中会避免这种情况的出现。

```
. list idcode ccity hours uniondues married marriedyrs nevermarried in 1/3,
>  abb(20)
```

	idcode	ccity	hours	uniondues	married	marriedyrs
1.	5159	1	38	29	0	0
	nevermarried					
	0					
2.	5157	0	35	0	1	0
	nevermarried					
	0					
3.	5156	0	40	0	1	3
	nevermarried					
	0					

有时用选项 noobs 来避免出现这种自动换行的情况。选项 noobs 禁止显示观测值的序号，这样偶尔会节省出足够的空间以防止变量列表在该页内自动换行。

在上面示例的命令中加入选项 noobs，再重新运行该命令，现在就节省出足够的空间防止列表在页内被换行显示。

```
. list idcode ccity hours uniondues married marriedyrs nevermarried in 1/3,
>  abb(20) noobs
```

idcode	ccity	hours	uniondues	married	marriedyrs	nevermarried
5159	1	38	29	0	0	0
5157	0	35	0	1	0	0
5156	0	40	0	1	3	0

接下来的示例用的是数据 tv1. dta，它包含了 10 个观测值，内容是 4 个不同孩子的看电视的习惯。

```
. use tv1
```

可以用命令 list 来查看整个数据。

```
. list
```

	kidid	dt	female	wt	tv	vac
1.	1	07jan2002	1	53	1	1
2.	1	08jan2002	1	55	3	1
3.	2	16jan2002	1	58	8	1
4.	3	18jan2002	0	60	2	0
5.	3	19jan2002	0	63	5	1
6.	3	21jan2002	0	66	1	1
7.	3	22jan2002	0	64	6	0
8.	4	10jan2002	1	62	7	0
9.	4	11jan2002	1	58	1	0
10.	4	13jan2002	1	55	4	0

注意每 5 个观测值间的后面会显示一条分隔线。这样更容易阅读输出结果。有时囿于空间，会禁止显示分隔线以确保列表能显示在同一页上。选项 seperator(0)［这个选项的缩写为 sep(0)］不显示这些分隔线。

```
. list, sep(0)
```

	kidid	dt	female	wt	tv	vac
1.	1	07jan2002	1	53	1	1
2.	1	08jan2002	1	55	3	1
3.	2	16jan2002	1	58	8	1
4.	3	18jan2002	0	60	2	0
5.	3	19jan2002	0	63	5	1
6.	3	21jan2002	0	66	1	1
7.	3	22jan2002	0	64	6	0
8.	4	10jan2002	1	62	7	0
9.	4	11jan2002	1	58	1	0
10.	4	13jan2002	1	55	4	0

在其他情况下，分隔符在区分观测值组群时特别有用。在这个数据中，每个孩子都有多个观测值，通过加入选项 sepby(kidid)要求在每两个 kidid 之间加入一条分隔线。这便于清楚地看到不同孩子的观测值组。

```
. list, sepby(kidid)
```

	kidid	dt	female	wt	tv	vac
1.	1	07jan2002	1	53	1	1
2.	1	08jan2002	1	55	3	1
3.	2	16jan2002	1	58	8	1
4.	3	18jan2002	0	60	2	0
5.	3	19jan2002	0	63	5	1
6.	3	21jan2002	0	66	1	1
7.	3	22jan2002	0	64	6	0
8.	4	10jan2002	1	62	7	0
9.	4	11jan2002	1	58	1	0
10.	4	13jan2002	1	55	4	0

本节对书中使用命令 list 时会用到的选项的介绍到此为止。在使用这些选项遇到疑惑，且没有关于这些选项是什么以及为什么用这些选项的解释时，希望本节能帮助你解除这些疑惑。

第二章　读取和录入数据

数据！数据！数据！巧妇难为无米之炊。

——夏洛克·福尔摩斯