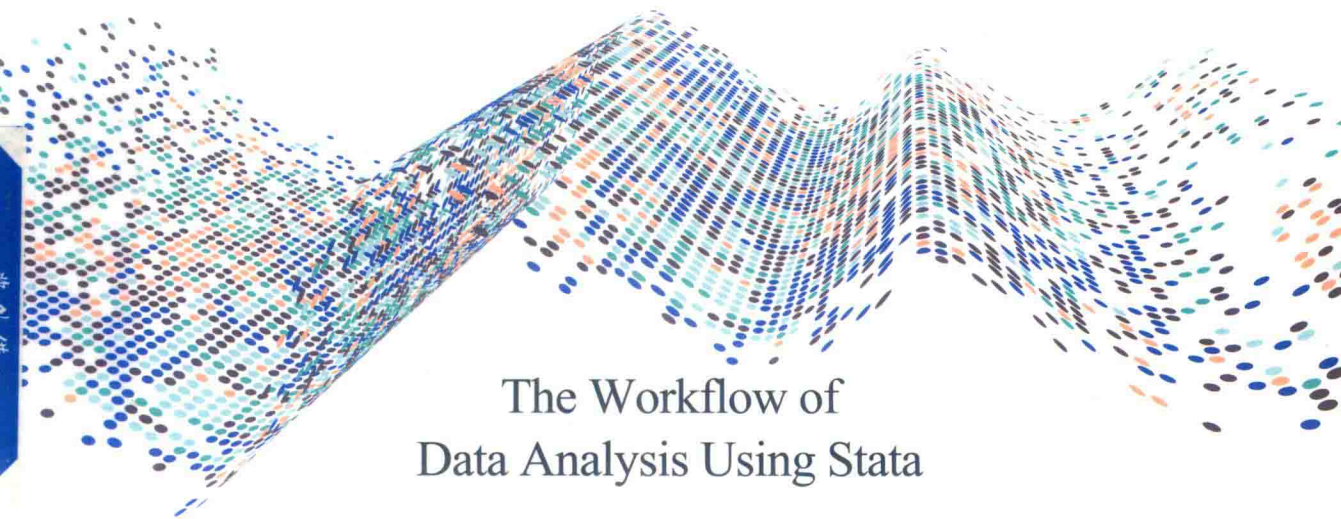
数据管理实务译丛
中国人民大学中国调查与数据中心组编

基于STATA的数据分析流程

[美]斯考特·隆恩（J. Scott Long）/ 著
唐丽娜　王卫东 / 译



The Workflow of
Data Analysis Using Stata

中国人民大学出版社

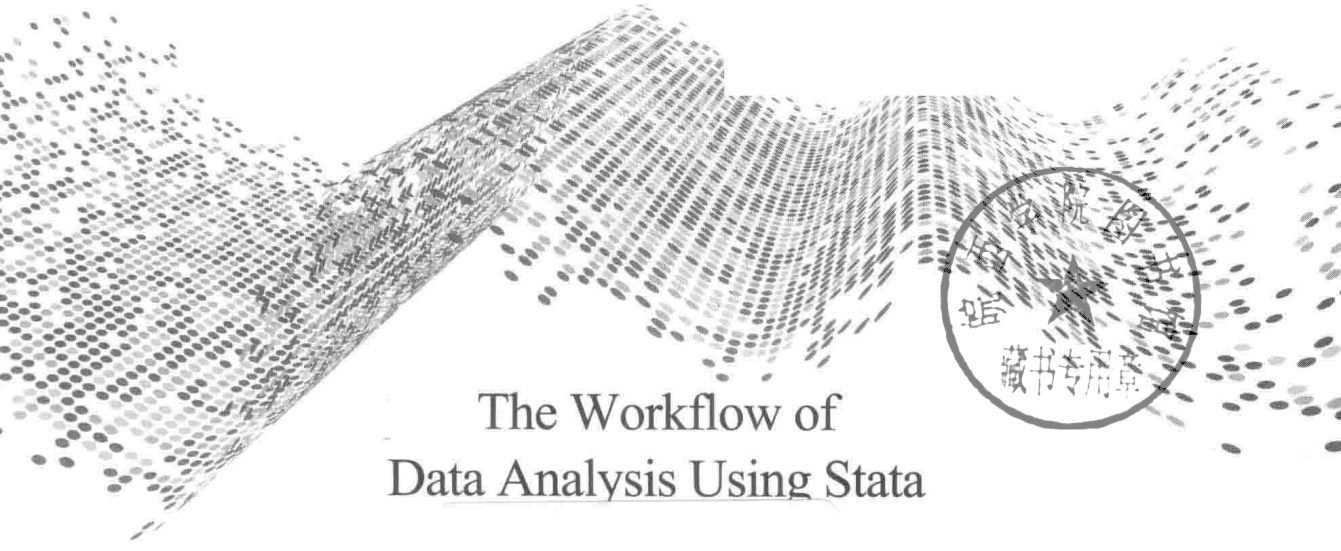
数据管理实务译丛

中国人民大学中国调查与数据中心组编

基于STATA的数据分析流程

[美]斯考特·隆恩（J. Scott Long）/ 著

唐丽娜　王卫东 / 译



The Workflow of
Data Analysis Using Stata

中国人民大学出版社

· 北京 ·

图书在版编目（CIP）数据

基于 Stata 的数据分析流程/（美）斯考特·隆恩（J. Scott Long）著；唐丽娜，王卫东译．—北京：中国人民大学出版社，2019.5
（数据管理实务译丛）
ISBN 978-7-300-26876-7

Ⅰ.①基… Ⅱ.①斯…②唐…③王… Ⅲ.①统计分析-应用软件 Ⅳ.①C819

中国版本图书馆 CIP 数据核字（2019）第 066305 号

数据管理实务译丛
基于 Stata 的数据分析流程
[美] 斯考特·隆恩 著
唐丽娜 王卫东 译
Jiyu Stata de Shuju Fenxi Liucheng

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		
经　　销	新华书店		
印　　刷	北京东君印刷有限公司		
规　　格	185 mm×235 mm　16 开本	**版　　次**	2019 年 5 月第 1 版
印　　张	22 插页 1	**印　　次**	2019 年 5 月第 1 次印刷
字　　数	435 000	**定　　价**	75.00 元

前 言

本书旨在介绍若干种方法，以便读者能够更加高效精准地分析数据。本书并不涉及具体的统计分析技巧，而是探讨所有数据分析过程中都要有的步骤。这些步骤包括制订工作计划、保存工作内容、构建和检查变量、进行统计分析、展示分析结果、复制研究发现以及对所有做过的工作进行存档。我将这些事项统称为数据分析的工作流程。对于研究结果的可复制性来说，一个好的流程是必不可少的。而可复制性，恰为科学研究所不可或缺的性质。

在多年的教学、科研、咨询和合作过程中，我逐渐萌生了写这本书的想法。我看到越来越多的人被淹没在数据中。廉价的数据计算和存储让创建新的文件和新的变量比管理它们更加容易。随着数据文件变得越来越复杂，数据管理的过程也变得越来越有挑战性。在接受他人咨询时，我的大部分时间花在了这些事情上：数据管理问题、一组分析结果是如何得出的。在与他人合作时，我发现了成倍的与工作流程有关的问题。另一个动力来自我与杰里米·弗里兹（Jeremy Freese）编写 Stata 程序包 SPost 的工作。去年，这个程序包的下载次数超过了 20 000 次，数以百计的用户与我们联系。在回复这些用户提出的问题时，我了解到不同领域的研究者们如何管理他们的数据分析工作，以及他们的管理工作是怎么崩溃的。我在帮助别人解决一些看似与 SPost 命令有关的问题时，经常发现这些问题与他们的工作流程有关。当大家问我有没有与工作流程有关的资料可供参考时，我没有任何可建议的。

写作这本书的最后一个动力来自布鲁斯·弗雷泽（Bruce Fraser）的《Adobe Photoshop CS2 的真实世界相机》（*Real World Camera Raw with Adobe Photoshop CS2*）这本书（2005）。数码摄影的一大诱人优势在于可以拍摄大量照片。其隐藏的问题在于需要掌握成千上万张照片的信息。影像专家们已经意识到这个问题很长时间了，并把它叫作“工作流程”——追踪工作的整个过程，从各阶段到最终产品。渐渐地，当我花在寻找照片上的时间超过了拍照的时间时，显然我需要接受弗雷泽的建议，为数码摄影建立一套流程了。弗雷泽的这本书启发我从工作流程的角度出发，重新审视数据分析。

在酝酿多年之后，我花了两年时间完成了本书的写作。起初，我认为我的工作流程非常好，写书很简单，就是把自己的经验记录下来。但随着写作的深入，我发现自己的工作中还存在着诸多不完善、不方便、不一致之处。有时这些不足之处来自我明知某种流程存在着缺陷，但从来没有花时间去寻找一种更好的解决办法。有些问题是由于我的疏忽，我没有意识到完成它或不能完成它可能产生的后果。在一些情况下，

我发现自己用多种方法完成了同一项工作，却没有从中选出一种最佳方案。本书的写作过程促使我在工作中更加前后一致、更加高效。在修改两篇已被接收发表的论文时，改善后的工作流程的优点显而易见。这两篇文章的数据分析过程，一个是在写作本书之前完成的，另一个是在本书初稿已基本成型时完成的。让我高兴的是，用本书中的工作流程来修改文章中的数据分析是那么容易。这种提高部分归因于找到了一种更好的做事方法。同样重要的是我用了一种前后一致且有记录的做事方法。

我从未幻想过本书建议的流程是最好的或者唯一的做事方法。事实上，我希望能从读者那里听到有关更好的工作流程的建议。读者的建议会被放在本书提供的网站上。但是，我提供的这些方法是行之有效的，并且能够避开一些隐患。就一个有效的工作流程而言，重要的一点是找到一种做事方法并坚持使用。在工作开始之初，统一的工作程序能够让你提高工作效率，而且在工作后期，如果你想回头看前面的工作，这种一致性能够帮助你清楚地知道已经完成的工作。一致性也能让研究团队中的工作更容易开展，因为合作者能更容易地跟进他人的工作进度。建立一套成文的工作流程的好处还有很多，与使用同样流程的人一起工作的优点也有很多。希望读者能够从本书中发现这样的流程。

虽然本书应该对所有从事数据分析的人员都有帮助，但写作还是有一些局限。首先，书中主要的计算机语言是 Stata，因为我发现 Stata 是最好的、通用的数据管理和统计分析软件。虽然几乎所有用 Stata 完成的工作，用其他软件也能实现，但本书不使用其他软件包。其次，书中大部分示例使用的数据来自社会科学领域，这是因为我本人主要从事社会科学研究。但是，书中讨论的原则同样适用于其他领域。最后，本书的分析主要在 Windows 系统中进行。这不是因为我认为 Windows 系统比 Mac 系统、Linux 系统更好，而是因为我工作时主要用的就是 Windows 系统。书中提供的方法和程序在其他操作系统中也可顺利运行，当程序在不同的操作系统中有差异时，我会尽可能地提醒读者注意。

我想感谢很多人，他们要么给本书的初稿提出了意见，要么解答了与工作流程有关的一些问题。我要特别感谢泰特·莱恩费德·密第那（Tait Runfeldt Medina），科蒂斯·查尔德（Curtis Child），纳丁·瑞博令（Nadine Reibling）和肖那·L. 柔曼（Shawna L. Rohrman），他们详尽的意见和建议使本书的内容更加完善。我还要感谢艾伦·阿科克（Alan Acock），麦伦·古特曼（Myron Gutmann），帕特里夏·麦克马纳斯（Patricia McManus），杰克·托马斯（Jack Thomas），利亚·万威（Leah Van Wey），里奇·沃特（Rich Watson），特瑞·怀特（Terry White）和里奇·威廉斯（Rich Williams），我在与他们探讨有关工作流程的问题时获益良多。StataCorp 公司的许多同人也通过不同方式向我提供了许多帮助。我尤其要感谢的是本书的出版人丽萨·吉尔摩（Lisa Gilmore）、编辑詹妮佛·内韦（Jennifer Neve）以及封面设计者安妮

特·费特（Annette Fett）。Stata-Corp 公司的大卫·M. 德鲁克（David M. Drukker）先生帮助我解答了许多问题，使本书得以日臻完善，而且与他的友谊也为我的写作增添了乐趣。书中部分资料来自基金号是 R01TW006374 的研究，该研究由福格蒂国际中心、美国国家精神健康研究院、印第安纳大学布鲁明顿分校行为与社会科学研究办公室提供资助。其他工作由一个匿名基金会和拜耳集团（The Bayer Group）提供支持。我由衷地感谢印第安纳大学艺术与科学学院提供的支持和帮助。

如果没有我亲爱的朋友弗雷德（Fred）无意中的鼓励，我可能不会开始写作本书。如果没有我亲爱的妻子瓦莱丽（Valerie）的大力支持，我更是无法将本书完成。我谨将本书献给亲爱的瓦莱丽，作为一份姗姗来迟的礼物。

斯考特·隆恩

2008 年 10 月于印第安纳州布鲁明顿

本书体例说明

本书的排版印刷采用的是 Stata 标准字体格式。以打字机字体输出显示的条目是 Stata 的命令语句和命令选项。例如，use mydata, clear。斜体字代表的是需要用户自行添加的信息。例如，use *dataset-name*, clear 表示用户需要指定数据文件的文件名。当我提供某一个命令的语法时，通常只列出该命令的某些选项。要想了解该命令的所有信息，可以通过在命令窗口输入 help *command-name* 或者查阅参考手册来获取。参考手册的名称采用 Stata 软件通用的符号来表示。例如，[R] logit 指的是《基础参考手册》(*Base Reference Manual*) 中的 logit 条目，[D] sort 指的是《数据管理参考手册》(*Data Management Reference Manual*) 中的 sort 条目。

在书中的某些地方，部分示例的命令语句或者输出结果的右端会超出页面无法正常显示，此乃我有意为之，目的在于向读者阐明不控制命令语句和输出结果的列宽会带来的后果。

书中包含了大量示例，我建议读者在阅读的同时要试着运行一下这些示例。所有以 wf 开头的文件均可下载。书中使用（文件 *filename*. do）字样来表示与示例相对应的 do 文件名。有极少的例外情况（例如，一些 ado 文件），如果文件名不是以 wf 开头（如 science2. dta），则无法下载这类文件。读者可通过查阅本书提供的文件包中的索引来查找下载的某一文件在书中出现的位置。

要下载示例文件，必须在联网的 Stata 环境中进行。Stata 10 和 Stata 9 各有两个可供下载的工作流程软件包。Stata 10 的两个软件包是 wf10-part1 和 wf10-part2，Stata 9 的两个软件包是 wf9-part1 和 wf9-part2。在 Stata 命令窗口中输入 findit workflow 来查找并安装这些软件包。在选择所需软件包后，可根据提示完成安装。由于本书示例文件过多，我将其分成两个文件包以供下载，但在书中会将其统称为工作流程软件包。在尝试运行这些示例文件之前，请确保已按照 [GS] 20 Updating and extending Stata-Internet functionality 中的提示更新了自己的 Stata 软件。读者可登录网站 http://www. indiana. edu/~jslsoc/workflow. htm 获取更多与本书内容相关的信息。

目　录

第1章　引言

本书旨在介绍若干种方法，以便读者能够更加有效、高效、精准地分析数据。这些方法被统称为*数据分析的工作流程*。工作流程涵盖了数据分析的整个过程，包括制订工作计划、记录工作内容、清理数据、创建变量、进行统计分析、实现分析过程的可复制性、展示研究发现以及工作归档。其实你已经有了一个工作流程，只是你尚未把它看成工作流程。这些工作流程可能是经过精心设计的，也可能只是临时建立的。由于很难找到专门探讨数据分析工作流程的书籍，也没有正式讲授这项技术的课程，所以研究者通常只有在遇到难题时才想到建立工作流程，听从的都是同事们的非正式的建议。举例来说，当你发现自己有两个同名但内容不同的文件时，就想建立文件命名的规程（例如，一个工作流程）。更普遍的是，一种好的数据分析方法经常是通过低效率的反复试错法来习得的。因此，希望本书能够帮助读者缩短学习过程，从而能够把更多的时间用于自己真正想做的事情上。

对本书初稿的反馈使我坚信，无论是初学者还是数据分析专家，都应该更加正式地思考一下自己是如何进行数据分析的，这个思考的过程会使他们受益良多。实际上，当我开始写书时，曾经认为自己的工作流程很好，只需要把自己平常的工作流程写出来即可。但当我把这些问题都系统地思考一番，并与其他研究者交流之后，又惊又喜地发现自己的工作流程水平有了很大的提高。每个人都可以轻而易举地改进自己的工作流程。虽然更改流程意味着时间的投入，但是这些投入会得到回报，那就是在日后工作中节省的时间和规避数据分析过程中的很多错误。

虽然书中提出了很多和工作流程有关的具体建议，但大部分我建议的事情可以用其他方法完成。我对某一特定问题的最好解决方案的建议基于我与数以百计的研究者和学生的合作，他们供职于不同的产业部门，涉及的领域从化学到历史学。这些建议对我个人的工作而言是行之有效的，而且大部分在广泛的实践应用中也得到了进一步的完善。但这并不意味着完成指定任务的方式只有一种，也不是说我有最好的办法。和任何一种复杂的统计软件一样，在 Stata 里有多种方法可以用于完成同一件任务。有些方法只能在有限的条件下完成任务，而且这些方法要么容易出错，要么效率低。在诸多行之有效的方法中，你就需要选择自己喜欢的方法。为帮助你做到这一点，对某一指定任务，我通常会讨论多种解决办法。与此同时，我还给出了一些低效率做事方法的案例，因为对读者来说，亲眼看到错误方法所造成的后果远比耳闻正确方法之优越性更令人印象深刻。这些案例都是真实的，来源于我曾经犯过的诸多错误和在帮助他人做数据分析时遇到的问题。读者需要做的就是选择一种能够与自己的项目特点、拥有的资源和自己的习惯爱好相匹配的工作流程。可以说，有多少做数据分析的人，就有多少种工作流程，没有哪一种工作流程是适合所有人或适用于所有项目的。关键是读者需要全面考虑问题，选择自己的程序并且严格执行，如果没有好的理由，就绝不随意修改。

在本章的剩下部分，我提供了一个用来理解和评估工作流程的框架。首先要讲的就是可复制性这一基本原则，该原则应该指导工作流程的各个方面。无论你如何做数据分析，都要确保结果的可证性和可复制性。接下来要介绍的是在所有类型的数据分析中都不可缺少的四个步骤：清理数据、进行分析、报告结果和保存文件。在每个步骤中都有四项主要任务：规划工作、组织管理资料、记录所做的工作、执行这些工作。如果工作中任何一个特定方面都有若干种解决办法，那么该如何判定工作流程的高下优劣呢？对此，书中提供了评估工作流程的若干标准。这些标准可以帮助读者决定使用哪种程序，也正是这些标准促成了书中有关最佳实践方法的建议。

1.1 可复制性：工作流程的指导原则

能够把已发表或出版的研究结果完整地复制出来是所有工作流程的基础。科学要求可复制性，一个好的工作流程能够提高你重复得到同样研究结果的能力。制订项目计划、保存工作、写程序以及保存分析结果都应该考虑到可复制性的需要。在多数情况下，研究者直到自己的工作遇到巨大挑战时才开始担心可复制性的问题。这并不是说他们都在走捷径、在分析中做假，或者做出了错误决定，而是说他们需要完成几个必要的步骤，以便将来可以毫不费力地重复之前已经完成的工作。举例来说，设想一个同事想要扩展你之前的某项研究，于是向你索要已发表的论文中使用的数据和命令。这时，你肯定不想手忙脚乱地复制出研究结果。虽然要找出以前的分析结果可能要花好几个小时（例如，我自己的很多资料保存在自己的笔记本中，这些笔记本都堆放在储物间），但这应该是一个检索工作记录的过程，而不是去回忆以前做了哪些工作，更不是发现自己的记录和报告的研究结果根本不匹配。

在整个工作流程中都应该始终牢记可复制性原则。在完成每一阶段的工作之后，如有必要，应该拿出一小时或一天的时间来回顾一下已经完成的工作，检查工作程序是否已经被记录，确认所用资料已经存档。当一篇要发表的文章的初稿写完时，需要回顾所有的文件记录，检查用过的文件是否均已保存，确认 do 文件是否仍然可以运行，并再次确认论文中的数字与结果中的数字是否一致。最终，确认整个过程都已记录在自己的研究日志中（详见第 37 页）。

如果你想在分析工作完成了几个月之后再来重复这项工作，或者试图只用一篇文章及其使用的原始数据就来重复某个作者的分析结果，就会发现复制有多么困难。要知道实现工作的可重复性有哪些要求的一个很好的方法是考虑一下有哪些使复制变得不可能的因素。这些因素中的大部分会在后文中得到详细的讲解。最开始，要找到原始文件，但随着时间的推移，要找到这些原始文件会变得越来越难。在找到这些文件

之后，看一下文件格式是不是当前统计软件能分析的格式。如果能打开这个数据文件，那么，你是否确切地知道数据中的变量结构或者数据中的案例是如何被选出来的？你知不知道每个回归模型中包含哪些变量？即使这些信息你都有，你也有可能发生现在使用的分析软件和当时用来分析的原始软件在计算方式上有不同之处。一个有效的工作流程可以使复制变得更容易。

一个最近出现的例子说明，即使是很简单的分析结果也可能很难复制。我从一个同事发表的论文中选取了一些数据，希望能够复制其结果并拓展该研究。但由于驱动故障，该同事的部分资料遗失了。我与该同事都无法再准确地计算出论文中的分析结果。计算结果很接近，但始终不能达到完全一致。这是为什么呢？假设在构建变量和选取分析样本时共做了 10 个决定。很多时候，这些决定是在两个正确的选项中做选择。举例来说，究竟是保留原研究结果的平方根呢，还是在原研究结果加上 0.5 之后取对数？10 个这样的决定就会产生 $2^{10}=1\,024$ 种不同的结果。这些结果都能导向相似的研究发现，但不完全相同。如果在构建数据的过程中偏离了某些决定，就会发现很难重复执行之前做过的事情。顺便提一下，另一位使用该数据的研究者成功地复制出了论文中的分析结果。

即使已经有了原始数据和分析文件，复制研究结果还是很困难。对于已发表的论文来说，通常很难获得原始数据和数据分析过程的详细记录。弗里兹（Freese，2007）提出了一个很有说服力的论点，针对的是为什么不同学科都要制定一些规则来管理信息的可获取性，因为这些信息是复制结果的必要条件。我完全赞同他的论点。

1.2 工作流程的步骤

数据分析包括四个主要步骤：清理数据、进行分析、报告结果以及保存文件。虽然这些步骤之间存在逻辑顺序，但是一个有效的工作流程的机制是灵活的，且主要取决于特定的项目本身。在理想情况下，一次前进一步，从开始一直做到项目完成。但我从来都没有按这个逻辑顺序做过。实际上我通常是根据工作的进展情况，在这几个步骤中来回穿梭：可能在分析时发现变量存在问题，就需要返回到清理数据这一步；或者分析结果提供了没有预计到的发现，于是就要修改研究计划。但是，我认为让这四个步骤互相独立是有益的。

1.2.1 清理数据

在开始实际分析之前，必须确认你的数据是准确无误的，数据中的变量已被加上了变量名并且加上了合适的标签。也就是说，你已经清理过数据了。一开始，需要把

数据导入Stata。如果你收到的数据格式是Stata格式，就很简单，只需要用use命令就能把数据直接导入Stata。如果你收到的数据是其他格式，则需要确保数据被准确地导入Stata。同时，需确认变量名和标签无误。不恰当的变量名会让数据分析起来更加困难，并会导致错误。同理，不完整或者不恰当的标签也会让分析结果阅读起来更加困难，而且同样会产生错误。变量的取值是否正确无误？有没有给缺失数据做合理的编码？数据是否具有内部一致性？样本量是否正确？变量的分布类型是否与预期的分布类型一致？在确认这些问题后，就可以根据分析需要选取样本并构建所需变量。

1.2.2 进行分析

在数据清理完成后，为论文或著作构建模型、绘制图表其实是工作流程中最简单的一个环节。实际上，书中和这部分有关的章节也相对简短。虽然这里不探讨具体的分析类型，但我在后文中会讲述确保结果精确性的几种方法，以便将来无论使用何种统计方法都能够复制分析过程，并有助于妥善保存好do文件、数据文件和log文件。

1.2.3 报告结果

数据分析一旦完成，就要把结果呈现出来。这里会探讨一些与报告流程有关的几个问题。首先，需要把Stata的输出结果导到论文或报告中。一个有效的工作流程可以自动完成大部分的导出工作。其次，记录下报告中所有分析结果的出处。如果报告中没有保存结果出处，日后（例如，其他研究者想复制出你的研究结果，或者你必须回复评审人）将很难追溯这些分析结果的来源。最后，还可以做很多简单的事情让报告更有效。

1.2.4 保存文件

在清理数据、分析数据和写作时，需要妥善保存各种文件，以免因为硬盘损坏、文件崩溃或者误删之类的原因而造成文件遗失。没有人喜欢在数据遗失之后重做分析或者重写论文。有很多简单的方法让自动保存文件变得更容易。随着备份软件越来越容易找到和硬盘存储成本越来越低，文件备份工作中最困难的部分是记录所做的一切。存档与备份相互独立，而且存档比备份更加困难，因为它需要在很长的一段时间内保存文件，以便在若干年之后还能找到这些文件。而且必须考虑到使用的操作系统（现在已经很难读取CP/M操作系统保存的数据了）、存储介质（你能读取20世纪80年代的5 $\frac{1}{4}$″容量的软盘或者几年前的ZIP压缩磁盘中存储的数据吗？）、自然灾害和黑客。

1.3 每个步骤中的任务

在四个主要步骤的每一步中，都有四个主要任务：规划工作、组织管理资料、记录所做的工作以及执行这些工作。尽管在每个步骤中，这四个任务的重要性是不一样的（例如，在做规划时，组织管理更重要），但在工作流程的四步骤中每一项任务都是重要的。

1.3.1 规划

大多数人在规划上花的时间太少，在工作上花的时间太多。在将数据导入 Stata 之前，应该对想做的事情草拟一个计划，并评估一下这些工作的优先排序。要思考：需要哪些类型的分析？如何处理缺失数据？需要创建哪些新变量？随着工作的推进，应在已完成的工作的基础上修正工作目标和分析策略，阶段性地调整工作计划。一个小小的计划，就能起到事半功倍的效果，而且我几乎总是发现制定规划能节省时间。

1.3.2 组织管理

精心的组织管理有助于加快工作速度。查找文件和避免重复工作的需要推动了组织管理工作。好的组织管理有助于避免寻找已遗失的文件、重建丢失的文件等工作。如果很好地记录下了已做的工作，却找不到工作中用过的文件，那就等于做了无用功。组织管理需要系统地思考如何给文件和变量命名，如何管理硬盘中的文件目录，如何记录哪台电脑里存储了哪些资料（如果有不止一台电脑的话），以及研究资料保存在什么地方。当你离开某个项目一段时间后或急需什么东西时，组织管理的问题就会出现。本书给出了一些如何组织管理资料的建议，而且讨论了一些让查找和工作更容易的工具。

1.3.3 记录

没有充分的记录，就不可能实现工作的可复制性，且更容易出错，通常工作时间也会更长。工作记录包括研究日志和编码表，研究日志记录已做的工作，编码表记录已创建的数据集和数据集中所包含的变量。完整的工作记录还应包括 do 文件中的注释、数据文件中的标签和注释。虽说写工作记录是一项很麻烦的工作，而且是整个数据分析过程中最无聊的一部分，但我发现做好工作记录在后期会相应地节省几周的工作时间，减少失望沮丧。虽然在记录上花时间是不可避免的，但本书给出了一些能更

快且更有效地记录工作的建议。

1.3.4 执行

执行包括完成每步中的具体任务。有效的执行需要合适的工作工具。一个简单的例子是写程序时用到的编辑器。掌握一个好的文本编辑器在写程序时能节省好几个小时的时间，而且写出来的程序质量更高。另一个例子是学习 Stata 中最有效的命令。花几分钟的时间掌握 recode 命令的用法会节省用在写命令 replace 上的几个小时的时间。本书中的很多章节包含了如何为工作选择正确工具的内容。在探讨这些工具的过程中，需要强调的是对工作进行标准化和自动化。之所以要标准化，是因为通常用以往工作中用过的方法来做事情会比用新的方法更快。如果给日常工作建立了一套模板，工作就会变得更加统一，这样就更容易进行查找工作，也更容易避免出错。有效的执行需要在花时间学习新工具、新工具所能带来的工作精确程度以及节省的时间三者之间进行权衡取舍，以便更有效地工作。

1.4 选择工作流程的标准

在执行工作流程每个步骤中的不同任务时，有各种不同的做事方法可供选择。如何确定使用哪个程序呢？本章介绍了几条评估正在用的工作流程以及从可供选择的程序中进行选择的标准。

1.4.1 准确性

一个好的工作流程的必备要素是能够得出正确的答案。对于这一点，奥利韦拉和斯图尔特（Oliveira，Stewart，2006，30）说得非常好：“如果程序不对，那么其他的一切都没有任何意义。”在工作的每个步骤中，都必须确认结果是正确的。结果是否回答了你提出的问题呢？结果是你想要的吗？一个好的工作流程也会出错。不变的是，错误永远会出现，而且有时会很多。一个有效的工作流程可以减少错误，也应该能够帮助你快速地找到错误和改正错误。

1.4.2 高效性

兼顾精确性和可复制性之后，你想尽快完成数据分析工作。完成工作和认真工作二者之间总是矛盾的。一方面，如果花费了太多时间在确认和记录所做的工作上，就会永远无法完成项目，那么这样的工作流程就是不可行的。另一方面，项目完成，但