Stata2R：针对 Stata 用户的 R 课程

能读懂、能复现、能改进

- **课程：** Stata2R - 针对 Stata 用户的 R 课程
- **嘉宾：** 游万海 (福州大学)
- **时间：** 2023 年 12 月 3/10/17 (三个周日)
- **时段**：上午 9:00-12:00，下午 14:30-17:30，17:30-18:00 答疑
- **课件/软件：** 提供全套 Stata+R 代码和数据 (开课前一周发送)
- **方式：** 在线直播 + 20 天回放（三周课程全部回放至 2024 年 1 月 6 日）
- **预读资料**：<在线查看>，<打包下载 .zip>
- **课程主页**：https://www.lianxh.cn/Stata2R.html
- **报名链接：** https://www.wjx.top/vm/tzLwRUz.aspx#
- **助教招聘：** https://www.wjx.top/vm/OzzTosc.aspx#
- **课程大纲：** PDF 版 | 网页版

目录

1. 为什么要学这门课?

A. 狐狸还是刺猬?

在计量和统计方法的快速发展和学科交叉融合的时代，越来越多的人开始从「刺猬」转变为「狐狸」(ps，刺猬只会一种非常有效的防身术——蜷缩起来；而狐狸则有很多种防身术。…… 二者过得都不错)。

世界银行的 DIME (Development Impact Evaluation) 项目组聚集了很多出色的经济学家。自 2020 以来，DIME 开始了一个培训项目，名为 R for Advanced Stata Users（github-课件）。其项目介绍为： *This material was developed by the DIME Analytics team as an introduction to R Statistical Package for its staff. It builds upon knowledge of Stata to explore features of R with impact evaluation applications in mind* .

我的推测是，多数 DIME 的学者都有 Stata 的使用经验，此时，学习 R 的最佳路径就是「迁移学习」，即将已经习得的概念、方法和经验迁移到新的学习对象上。这显然是一个行之有效的学习方法。

B. Stata2R 模式：why?

其实，**Stata2R** 这一表述并不准确，感觉像是从 Stata 转向了 R。更准确的表述应该是 $Stata^{+}\!\!\rightarrow\!R$：基于 Stata 的经验更快地学习 R，进而充分发挥 Stata 和 R 的比较优势来高效完成研究任务。

(1) $Stata^{+}\!\!\rightarrow\!R$ **模式的必要性。** 从连享会分享的几个前沿专题课程来看，这种工作模式非常必要，也非常奏效。比如，杨海生老师的「政策学习专题」会同时使用 Stata 和 R 来讲解，王菲菲老师的「文本分析」课程全程使用 R。越来越多的 Top 期刊论文 会同时使用 Stata 和 R 甚至多种软件来完成。

(2) **打消顾虑。** 有些同学一直对 R 心存顾虑，认为学习门槛高，开源特征导致用户要花很多时间来应对各种莫名奇妙的错误信息。但这一切在最近五年中发生了实质性的变化：**R 有了新生态**。得益于 Hadley Wickham 等大牛出色地开发和整合工作，R 已经形成了一个多元化、模块化的生态系统。(1) 新手们无需在众 R 包之间穿梭，只需安装 `tidyverse`，`ggplot2`，`tidyr`，`purrr`，`tidymodel` 等几个包，就以应对多数分析任务，因为每个包其实就是针对某一类任务而开发的模块。(2) R 可以将「统计分析+可视化+写作」无缝衔接起来，可以实现「**数据收集和清洗 → 统计建模 → 论文和报告撰写 → 成果展示**」整个流程。比如 RStudio 提供的 R Markdown，可以很方便地撰写实证研究报告、幻灯片，甚至是讲义，并以 HTML，PDF 等格式输出； Shiny 可以方便实现 APP 的制作，网页生成，用于成果展示；新近推出的 Quarto 平台，则可以方便地将 R, Python, Julia 等整合在一起，实现统计分析和可视化报告、幻灯片甚至是电子书的制作。换言之，软件 (工具) 之间的界限会越来越模糊。

Hadley Wickham

(3) **学科之间的交叉融合趋势。**很多 Top 期刊的论文会由多位合作者协作完成，大家可能有着不同专业背景，擅长的领域或使用的统计软件也不同，可以发挥各自的比较优势。因此，一篇文章中同时使用多种软件或语言不足为奇，比如：

- **Stata + R** (Deryugina et al., 2019, AER , Bajari et al., 2015, AER)；
- **Stata + Python** (Butters et al., 2022, AER)

- **Stata + Matlab** (Corbae and D'Erasmo, 2021, Econometica)
- **R + Matlab** (Bernanke, 2020, AER)

C. Stata2R 模式：How?

有趣的是，这种「$Stata \overset{+}{\rightarrow} R$」的学习模式也被很多大牛采用。例如，对多期 DID 和 Bartik IV 估计都有深刻洞见的 Paul Goldsmith-Pinkham 就经常在 Stata 和 R 之间穿梭，还写了一篇有趣的博文 Comparing tidyverse R to Stata。在他给耶鲁大学的博士生们开设 Applied Empirical Methods 课程时，课件 中的代码总是不断地在 Stata 和 R 之间穿梭。事实上，十多年前就已经有这方面的教材了：

- Muenchen, R. A., J. Hilbe. **R for Stata Users**[M]. Springer, **2010**. -Link-, -Website-, PDF, Data-Codes
- Oscar Torres-Reyna, Getting Started in $R \leftrightarrow Stata$: Notes on Exploring Data (v. 1.0), 2010, Princeton University, -PDF-
- Chuck Lanfear, R and Stata Equivalencies, -Link-, -Github-

2. 授课嘉宾

游万海，福州大学经济与管理学院副教授，主要研究领域为空间计量模型、分位数回归模型及应用，成果见诸《统计研究》，World Development, Energy Economics, Economics Letters, Finance Research Letters, Economic Analysis and Policy 等期刊，担任 Energy Economics, Economic Modelling 等期刊匿名审稿人。讲授《统计学》、《统计分析软件（R 语言）》和《计量经济学前沿》等课程，受邀在中国人民银行福州支行讲授「数据挖掘」专题。

3. 课程目标

本课程的目标帮助大家转换为新的工作模式：

$$Task = \beta_1 Stata + \beta_2 R + v$$

对于一个老牌 Stata 用户而言，新的理念是：**用 Stata 搞定 Stata 的事儿；用 R 搞定 R 的事儿。**虽然很多时候 β_2 的取值很小 (e.g., 0.1 或 0.2)，但却能帮你节省很多时间，避免了重复造轮子，把更多的精力放在选题、研究设计、深度思考上。经过一段时间的磨合，大家可以找出让自己工作效率最高的 $\widehat{\beta}_1$ 和 $\widehat{\beta}_2$。

本次课程尝试通过对 `Stata` 和 `R` 在语法风格和使用方法上的对比讲解，辅以多篇 `TOP` 期刊论文复现，以期实现「**能读懂** → **能复现** → **能发表**」的目标。

- **能读懂、能复现。** 公开代码和数据是大势所趋，诸如 AER、QJE、JAE，以及中文期刊 中国工业经济、世界经济、数量经济技术经济研究 和 管理世界 都已要求或建议公布数据和源代码，此外，很多作者也会在其主页公布数据和代码，如 James MacKinnon 和 Josh Angrist 教授等。要充分利用这些资源，就要求我们能读懂多种来源的代码，通过复现加深对前沿方法的理解。
- **能追踪前沿方法。**研读 Journal o Econometrics，Econometrica 等期刊所涉及的前沿方法时，要求我们不仅能读懂作者提供的代码，还要能适当修改和调整，应用于自己的研究中，以此达到跟踪和借力前沿方法目的。
- **掌握实证分析的新流程：Stata + R。**目前有不少发表于 QJE, AER 等期刊上论文都采用了这种「混合代码模式」，即数据清洗和多数实证分析用 Stata 完成；而涉及文本分析、机器学习等方面的内容，则用 R 来完成；最终的结果输出、绘图等可以转回 Stata。基本的原则是，不重复造轮子，尽量站在巨人的肩膀上。

目标人群：

- 立志做学术研究的小白，想掌握两种以上的软件，以便紧跟统计学和计量经济学的前沿方法；
- 有一些 Stata 基础，想学习文本分析和机器学习等在 R 中能够轻松实现的前沿方法；
- 使用 Stata 多年的老手，了解很多新方法的原理，但发现对于很多新方法，作者只提供 R 代码或关键代码是 R 写的。

4. 课程详情

T1. Stata2R：基础知识 (3 小时)

本节尝试通过对比，将大家已经掌握的 Stata 经验迁移到 R 语言的学习中。主要内容包括：(1) 厘清 `Stata` 和 `R` 在语法和使用上的区别，了解 `R` 中对象和函数等概念和用法，掌握函数返回值的获取方式；(2) 基于 `tidyverse` 进行数据基本操作，包括数据导入导出和变量操作等，保证代码语法风格的统一性；(3) 基于 `RStudio` 讲解动态文档和幻灯片等制作，可直接用于报告或者达到 SSCI 发表要求的格式框架，打通实证分析的最后一公里。

- 操作界面初识
- 面向过程和面向对象编程
- 基于 `tidyverse` 的数据基本操作
- 基于 `Rstudio` 的动态文档/幻灯片制作
- 案例：经典最小二乘回归模型实现

T2. R 语言进阶：语法与数据处理 (3 小时)

数据处理是实证研究中至关重要的一环，预计占据了整个实证研究工作的三分之一。本部分将分别就数值型、文本型和日期型数据进行讲解，掌握**dplyr**、**stringr**和**lubridate**常用函数与用法，学习常用数据清洗步骤 (数据的横向合并和纵向追加、长宽数据变换等)，为论文写作奠定基础。

- 数据结构：向量 (`vector`)、矩阵 (`matrix`)、数组 (`array`)、数据框 (`data.frame`) 和列表 (`list`)：创建/引用/增加/删除
- 基于 **dplyr** 数值型数据整理
- 基于 **stringr** 文本型数据整理
- 基于 **lubridate** 日期型数据整理
- 实例

T3. Stata2R：实践运用 (3 小时)

本部分针对实证论文中常用的模型，针对同一研究问题，利用 `Stata` 和 `R` 语言进行同步处理，从而完成对比学习，以期达到事半功倍的效果。

- OLS 回归模型
- 经典面板数据模型
- DID 模型
- 空间面板数据模型

T4. 非 `R` 莫属 (3 小时)

与 `Stata` 相比， `R` 在网页爬取、机器学习和自然语言处理等方面有着独特的优势，本讲就当前较为热门且被使用较多的静态/动态网页爬取和自然语言处理 (如：文本挖掘、文本分词) 为示例，对 `R` 的优势进行介绍；最后，基于 `ggplot2` ，就实证论文中关键步骤数据/结果可视化进行讲解。

- 静态和动态网页爬取

- 自然语言处理
- 基于 `ggplot2` 包可视化
    - 数据可视化
    - 实证结果可视化

T5. Stata2R：论文复现 I (3 小时)

从方法角度来看，一篇论文往往涉及多种计量方法的综合应用；从论文构成角度来看，一篇论文由多个部分组成，通常涉及数据搜集、数据清洗、统计分析和结果展示等部分。然而，不同软件有各自的擅长点和专注点，因此，一篇论文通常结合多门软件完成。

本专题以 QJE 和 AER 上所发表的论文为例，拆解其中用到 `Stata` 与 `R`，一方面帮助学员能够读懂论文代码，另一方面能够对其中的代码做一些基本的修改，并将论文方法和代码用于自己研究中，达到「**能读懂，能复现**」的目标。

讲解如下四篇论文的实证分析和代码复现过程：

- Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. American Economic Review, 105(5), 481-485. -Link-, -Replicaton-
- Le Barbanchon, T., Rathelot, R., & Roulet, A. (2021). Gender differences in job search: Trading off commute against wage. The Quarterly Journal of Economics, 136(1), 381-426. -Link-, -Replicaton-
- Harding, M. C., & Lamarche, C. (2021). Small steps with big data: using machine learning in energy and environmental economics. Annual Review of Resource Economics, 13, 469-488. -Link-, -Replicaton-
- Dong, Y., & Kolesár, M. (2023). When can we ignore measurement error in the running variable?. Journal of Applied Econometrics. -Link-, -Replicaton-

T6. Stata2R：论文复现 II (3 小时)

跟踪文献、借力前沿方法是当前实证计量的一个主要套路。其中面临的一个难题是所需模型是否有现成代码。我相信很大一部分人在考虑是否采用某模型之时，首先想到的是能否有现成的代码可以套用。然而，代码可能基于不同语言编写的，因此，我们应该**掌握多种软件和语言，把更多的精力放在选题、深度思考上**，而不是**重复造轮子，把精力放在复杂编程上**。

本专题通过重现 3-4 篇发表于 Top 期刊上的论文 (AER、QJE、JAE)，与大家分享基于 `R` 语言的论文复制经验。课程将从基本的数据读取、程序放置、程序运行和 debug 等进行细致的讲解，课程之后有能力根据作者提供的代码和帮助文件，复制论文结果，并将之用于自己的研究中。

参考文献：

- Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106. -Link-, -Replicaton-
- Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), 2369-2429. -Link-, -Replicaton-
- Kelejian, H. H., & Piras, G. (2016). An extension of the j-test to a spatial panel data framework. Journal of applied econometrics, 31(2), 387-402. -Link-, -Replicaton-
- Canay, I. A. (2011). A simple approach to quantile regression for panel data. The Econometrics Journal, 14(3), 368-386. -Link-, -Replicaton-

5. 开课前预读资料

为提高听课效率，建议大家花点时间预先学习一下 R 的基础知识。学习 R 没你想象的那么困难，你只需要老老实实地对着 **R4DS** 操作一下就可以很快上手了。

> R 入门和基础

- *R4DS* | Hadley W., and G. Grolemund. **2023**. **R for Data Science**. O'Reilly Media.
  - R4D-2E-在线阅读， Solutions
  - Data 社区: TidyTuesday
  - 这本书是各领域学习 R 的必读书目
- *ModernDive* | Ismay C., and A.Y. Kim, **2022**. Statistical Inference via Data Science: A ModernDive into R and the Tidyverse. -在线阅读-.
  - 采用最新的 `tidyverse` 和 `moderndive` 包进行数据处理、回归分析、统计推断、Bootstrap 等。
  - 大量使用了管道 (pipe) 操作符，符合 R 代码的主流规范

> 其他:

- **RStudio 之 R 学习导航**： Finding Your Way To R，介绍了入门、中级和高级学习过程中的主要参考资料
- **Roadmap**： A roadmap for getting started with R, R 打怪升级指南
- **BBR** Baruffa, Oscar. 2023. Big Book of R. -Link-. R 语言资料清单：包括各类书籍、包等资料的清单。
- 连享会： R 资料， Stata v.s R 对照表

6. 报名和缴费信息

- **主办方**： 太原君泉教育咨询有限公司
- **标准费用** (含报名费、材料费)：499 元/人 (全价)
- **联系方式**：
  - 邮箱： wjx004@sina.com
  - 电话 (微信同号)： 王老师 18903405450 ; 李老师 18636102467

报名链接

> **报名链接：** https://www.wjx.top/vm/tzLwRUz.aspx# 或 长按/扫描二维码报名：

缴费方式

> **方式 1：对公转账**

- 户名：太原君泉教育咨询有限公司
- 账号：35117530000023891 (晋商银行股份有限公司太原南中环支行)
- **温馨提示：** 对公转账时，请务必提供「**汇款人姓名-单位**」信息，以便确认。

> **方式 2：微信扫码支付**

> **温馨提示：**

- 支持公务卡转账：请使用已经绑定公务卡的「微信/支付宝/云闪付」等扫码付款
- 微信转账时，请务必在「**添加备注**」栏填写「**汇款人姓名-单位**」信息。

7. 听课指南

7.1 软件和课件

听课软件：**支持**手机，ipad ，平板以及 windows/Mac 系统的笔记本，**但不支持台式机**

> **特别提示：**

- 为保护讲师的知识产权和您的账户安全，系统会自动在您观看的视频中嵌入您的「用户名」信息。
- 一个账号绑定一个设备，且听课电脑不能外接显示屏，请大家提前准备好自己的听课设备。
- 本课程为虚拟产品，**一经报名，不得退换。**
- 为保护知识产权，课程不允许以任何形式录屏及传播。

7.2 实名制报名

本次课程实行实名参与，具体要求如下：

- 高校老师/同学报名时需要向连享会课程负责人 **提供真实姓名，并附教师证/学生证图片**；
- 研究所及其他单位报名需提供 **能够证明姓名以及工作单位的证明**；
- 报名即默认同意「连享会版权保护协议条款」。

8. 诚聘助教

说明和要求

- **名额：** 10 名
- **任务：**
  - **A. 课前准备**：协助完成 1 篇推文，风格类似于 lianxh.cn ；
  - **B. 开课前答疑**：协助学员安装课件和软件，在微信群中回答一些常见问题；
  - **C. 上课期间答疑**：针对前一天学习的内容，在微信群中答疑 (8:00-9:00，19:00-22:00)；参见
  - Note: 下午 5:30-6:00 的课后答疑由主讲教师负责。
- **要求：** 热心、尽职，熟悉 Stata 或 R 语言的基本语法和常用命令，能对常见问题进行解答和记录。
- **特别说明：** 往期按期完成任务的助教自动获得本期助教资格，不必填写申请资料，直接联系连老师即可。
- **截止时间：** 2023 年 11 月 15 日 (将于 11 月 18 日公布遴选结果于连享会主页 lianxh.cn)。

申请链接： https://www.wjx.top/vm/OzzTosc.aspx#

或扫码填写助教申请资料：

课程主页： https://www.lianxh.cn/Stata2R.html

附：相关课程

> ⇛ 计量和因果推断 · 强基班

- **嘉宾：** 司继春 (上海对外经济贸易大学)
- **时间：** 2023 年 11 月 5/12/19 (三个周日，线上)
- **详情：** PDF 大纲 | <资料下载>
- **报名/助教：** -点我报名- (3300 元/人) | 助教招聘

> ⇛ 文本分析：从文本到论文

- **嘉宾**：王菲菲 (中国人民大学)
- **时间**：2023 年 11 月 11, 18, 25 日 (三个周六，线上)
- **详情**：PDF 版 | 网页版 | 预读资料
- **报名/助教**：-点我报名- (3300 元/人) | 助教招聘

> 主页 || 视频 || 推文 || 知乎 || b 站