连享会：R 语言初级

> 多开一扇门：我要「Stata+R」了！

1. 课程概览

- **时间：** 2022 年 10 月 4-6 日
  - **上午**：9:00-12:00，**下午**：14:30-17:30，**答疑：** 17:30-18:00
- **方式：** 网络直播 + 30天回放
- **授课方式：** 幻灯片+R 实操演示，全程电子板书+R 演示截图，课后以 PDF 形式分享给学员。
- **授课嘉宾：** 游万海 (福州大学)
- **全程答疑：** 15 位经验丰富的助教，答疑文档公布于 课程主页
- **课程主页：** https://gitee.com/lianxh/Rcourse ↬ PDF课纲
- **报名链接：** http://junquan18903405450.mikecrm.com/qWXY25e
- **助教招聘：** https://www.wjx.top/vm/hngNjLi.aspx#

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、分位数回归模型及相关实际问题的应用，已在《World Development》、《Energy Economics》、《Economics Letters》、《Finance Research Letters》、《Journal of Cleaner Product》、《Energy Sources, Part B: Economics, Planning, and Policy 》、《统计研究》等期刊发表学术论文 30 余篇，担任《Energy Economics》、《Journal of Cleaner Product》、《Economic Modelling》、《International Review of Economics & Finance》等期刊匿名审稿人。

3. 为什么要 Stata+R？

我从 2003 年开始接触 Stata，几乎每天都用，快 20 年了。过去的一年中，我追踪因果推断、机器学习、政策学习 (Policy Learning)、模型平均化 (Model Averaging) 等领域的最新进展时，发现很多新方法的实现多以 R 语言为主。我逼自己花了 1 个多月去了解 R 语言，发现她有很多可爱之处，完全打消了我多年来对开源软件的误解。

与几位好友交流的结论是：一个软件包打天下的时代已经过去了，我们需要同时掌握多种工具，原因如下：

其一，学科之间的交叉融合。很多 Top 期刊的论文会由多位合作者协作完成，大家来自不同学校、不同专业，既有思想碰撞又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)

其二，统计和计量方法的快速发展。同样一个政策评价问题，往往有多种各有优劣的识别方法，为了确保结果的稳健性，往往需要同时使用多种方法和模型进行估计和检验。这也为实证分析工作提出了新的挑战：如何有效掌握多种方法，合理应用之？此时，同时掌握多个工具便是很好的选择。比如，就面板数据模型而言，Stata 具有绝对优势，然而，在机器学习、文本分析和可视化等方面，R 语言和 Python 明显占优。尤其是最近十年中，很多基于大数据、机器学习发展出来的因果推断方法都是由统计学家主导的，而 R 是他们的主要工具。相比之下，Stata 在这些领域的更新速度就显得过于缓慢了。

目前，Stata 官方已经提供了与 Python 的交互功能，应该很快就会提供与 R 语言的交互 (目前，外部命令 `rcall`，`rsource` 等命令已经可以在 Stata 中执行 R 命令)。从 R 和 Python 的角度来看，二者都可以很轻松地读入 Stata 格式的数据，甚至执行 Stata 命令。

在这种大趋势下，我们的问题不再是「我该学习 **哪种** 软件？」，而是「我该学习 **哪几种** 软件？如何搭配？」

大体建议如下：

- **Stata**：常规数据的处理、经典计量模型估计、结果输出等。
- **R 语言**：非结构化文本数据的获取和清洗、机器学习和深度学习算法、因果推断前沿方法等。
- **Python**：数据分析、机器学习，……。

当然，这里只是列举各个工具的典型优势，至于最终选择 **Stata+R** 还是 **Stata+Python**，抑或其它组合完全取决于我们的研究领域和个人的偏好。

简言之，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。**

对于多数已经具有 Stata 基础的同学而言，R 或 Python 很容易上手，大道至简，很多基本思想和语法都相通。即使仅仅掌握如何在 R 软件中调用各类包，学会运行命令也足为我们开启一扇新的大门。

作为一个 Stata 老用户，我为自己的「叛变行为」感到自豪，这种「反叛」意味着更开放、更包容的学习态度。

基于上述考虑，连享会在后续的课程中会尝试促进各个工具软件的融合使用。本次课程，我们邀请了有多年 R 使用和编程经验的游万海老师，帮助大家搭建起 R 语言的学习框架，以便为后续的进阶学习打好基础。

> 连玉君 (连享会课程负责人)
> `2022/9/9`

4. 课程导引

作为一门免费、开源的语言，R 被广泛应用于 **数据挖掘、机器学习、数据可视化、计量经济学和空间统计** 等领域。正是因为其拥有众多使用者，大量的外部包被开发应用于各个领域 (18549 个，截止 2022.8.28)。这也是 **为什么 R 体积小，功能却如此多** 的原因所在。

R 用户群体非常庞大，且呈现逐年递增趋势，资源丰富，遇到的问题大都能找到答案，如 统计之都论坛、RWeekly、stackoverflow 等。这就带来很多益处，包括：

- **第一**，众多新提出的计量和统计模型在 R 中可以找到相应的工具包实现。如在因果推断中，各类的 DID 衍生模型可以在 R 中方便的实现，如 多期DID、双重稳健DID、合成 DID。又如，近年来发展迅猛的模型平均化 (model averaging)，最前沿的方法基本上都能在 R 中实现，参见 MA-R，以及 R pakcages。相比之下，Stata 中只有 2009 年发布的 `bma` 命令。
- **第二**，与 Stata 的友好衔接。让 Stata 用户可以临时做客 R，使用 R 中独有的新命令和功能。当前，Stata 提供了 `rcall` 和 `rsource` 两个命令，使得 Stata 用户可以很方便的调用 R 代码。如输入 `rcall: library(ggplot2)` 即可调用 R 中的 `ggplot2` 包绘制精美图形。随着时间的推移，Stata 与 R 间的互动将越来越紧密，对于 Stata 用户来说，是时候掌握一些 R 的基础操作，以达到事半功倍的效果。
- **第三**，可重复研究。也许大家都碰到这样的一个问题，好不容易等到审稿意见回来，审稿人让更新实证数据。看似简单的一条审稿意见，其实不然，包括了：**运行实证结果** → **整理表格和图形** → **将图表插入到论文指定位置** ...，那么有没有一种模式能将文字和代码融为一体，使得只需要更新原始数据，相应的结果将自动更新呢？R 中的可重复研究 Reproducible Research为这问题的解决提供了可能，包含了诸如 `knitr`，`rmarkdown` 等包。
- **第四**，与 Stata 相对，R 在机器学习和自然语言处理方面的优势也十分明显，包括 随机森林-Random Forests、支持向量机-Support Vector Machines、递归分类算法-Recursive Partitioning and Regression Trees、神经网络与深度学习-Neural Networks and Deep Learning 等机器学习算法，以及 quanteda、tidytext、tm、topicmodels 等自然语言处理技术。
- **第五**，相比于 Stata，在图形绘制方面，R 主要有以下几点优势：(1) 能绘制的图形丰富多样，包括 雷达图、相关矩阵图、多层网络图等，还包括了 patchwork、ggthemes 等图形修饰包；(2) 语法相对统一，比如可以通过 `geom_point`，`geom_line`，`geom_bar` 等分别绘制点图、线图和柱状图，同时图形可以通过 `+` 叠加多个图层，更加符合人们的画图习惯。

在此次课程中，力求通过三天课程的系统学习，实现如下目标：

> **其一**，建立起 R 的基本架构，熟知 R 能做什么、如何做，以期为后续学习打下宽厚扎实的基础；
>
> **其二**，掌握使用 R 进行实证分析的流程，从数据导入与整理、模型估计、表格图形输出，以及文档的制作。这样可以避免繁琐的图表插入，大大提高论文写作效率。

在**内容安排**上，基本上遵循了由浅入深，循序渐进的原则。

第 1-3 讲 依序介绍 R 的基本用法、常规数据处理、程序编写，学习这些内容对于提高实证分析能力和分析效率大有裨益。同时，也会同时列出一些常用命令的 Stata 和 R 代码，通过平行的比较，以便大家将 Stata 的经验迁移到 R 语言学习中，做到融会贯通。其实，不管是 Stata，R，或是 Python，都可以相互调用。比如，借助 PyStata，用户可以直接在 Stata 17 中执行 Python 指令；R 中最受欢迎的编辑器 RStudio 可以直接运行 Python 指令，稍加设置即可调用 Stata 指令。

第 4-5 讲 介绍文本分析处理和文本数据分析方法，包括各类文本数据的读取、匹配、抽取等，以及文本相似度、文本复杂度计算和主题模型。

第 6 讲 介绍实证分析可视化中常见图形的绘制，包括散点图、线形图、相关矩阵图、雷达图等，通过讲解，希望各位同行不仅能够自行绘制这些图形，更重要的是要了解这些图形的应用场景。

具体说明如下：

- **第 1 讲** 介绍 R 的基本语法结构，并对数据处理过程中的关键问题进行介绍，如离群值的处理等。通过帮助文件和外部工具包的使用等讲解 R 语言的学习路径。数据处理能力的高低直接决定实证分析的效率，而对于离群值等问题的处理是否妥善会直接影响全文结果的稳健性，是多数人不够重视但却至关重要的问题。
- **第 2 讲** 介绍 R 编程的基础知识。谈到程序撰写，很多人可能会产生恐惧心理，其实一旦掌握了最基本的原理和语法格式，程序的撰写并没有想象的那么困难。在掌握基本的命令使用和四则运算后，可以将内容分解为基本数据类型、数据结构和控制语句，然后通过函数编写将程序进行封装，使之成为可以被方便使用的命令。一旦掌握了基本的编程知识和理念，你的实证分析便开始进入「快车道」了。
- **第 3 讲** 介绍常用的数据处理工具包，讲解常规数据处理方法，包括数据的横向合并和纵向追加、长宽数据变换、多文件操作。本部分将同时展示 R 和 Stata 用于处理数据的方法，通过对照学习，加深理解。
- **第 4 讲** 通过具体实例介绍文本数据的清洗方法，包括各类文本数据的导入、文本匹配、文本截取等，讲解基础的正则表达式，R 在文本型数据方面的优势更为明显。
- **第 5 讲** 介绍文本数据中涉及的常用分析方法，包括文本分词，词频统计，文本复杂度和相似度计算，主题模型以及文本数据可视化等。
- **第 6 讲** 介绍实证分析中常用的图形绘制方法，特别是外部包 ggplot2 及其扩展包的功能实现。为什么学习 `ggplot2`？归纳起来有如下几点理由：第一，绘制的图形非常美观，这点也是最为重要的；第二，延伸功能非常强大，大量基于 `ggplot2` 的扩展包被开发，例如：扩展包；第三，绘图语法非常直观，实现了数据和图层的分离，真正做到图形的「**可加**」。利用 `ggplot2` 所画的图形已经出现在了 Nature，Nature Communications，PANS 和 Cell 等期刊上。

5. 专题介绍

A1. R 语言基础

- 为什么用 `R`？ R vs. Stata
- 数据的导入和导出
- 执行命令和基本统计分析
- 行列操作
- 重复值、缺漏值的处理
- 帮助文件使用和外部包安装
- 一篇范例文档

A2. R 语言程序

- 基本数据类型
- 数据结构：向量、矩阵、数据框、数组、列表
- 控制语句（条件语句、循环语句）
- 函数编写与调用
- 如何使用他人分享的程序?
- 利用 RStudio 制作文档及幻灯片

A3. 数据处理

- 常用的数据处理程序包 `dplyr` 、 `data.table` 介绍
- 数据的横向合并和纵向追加
- 长宽数据变换
- 高级函数的使用： `Reduce()` , `do.call()` 等
- 导入 Stata, Python 格式的数据
- 案例： `WDI` 数据整理为面板数据格式

A4. 文本数据清洗

- 文本数据的读取：**txt，word，pdf**
- 常用函数讲解: `grep()、grepl()` 等
- 基于 `stringr` 的文本清洗
- 一篇范例文档：上市公司年报分析

A5. 文本数据分析

- 文本数据分析的基本步骤
- 文本分词
- 词频统计
- 文本相似度和复杂度计算
- 主题模型
- 文本数据可视化
- 一篇范例文档

A6. 实证分析可视化

- 为什要可视化?
- 基础绘图工具: `plot()`
- 基于 `ggplot2` 绘图工具：散点图、线图、雷达图等
- 相关系数矩阵图 `ggcorrplot`

- 图形排版与图注 `patchwork`、`geomtextpath`
- 线性回归分析: OLS，虚拟变量，交叉项
- 系数及系数差异的可视化呈现
- 调节效应和边际效应的可视化

参考文献

- 张成思, 孙宇辰, 阮睿. 宏观经济感知、货币政策与微观企业投融资行为[J]. **经济研究**, 2021, 56(10): 39-55. -Link-
- 张叶青, 陆瑶, 李乐芸. 大数据应用对中国企业市场价值的影响——来自中国上市公司年报文本分析的证据[J]. **经济研究**, 2021, 56(12): 42-59. -Link-
- 李晓溪, 杨国超, 饶品贵. 交易所问询函有监管作用吗?——基于并购重组报告书的文本分析[J]. **经济研究**, 2019, 54(05): 181-198. -Link-
- Deryugina, Tatyana, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif. 2019. The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction. American Economic Review**, 109 (12): 4178-4219. -Link-
- Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2015. Machine Learning Methods for Demand Estimation. American Economic Review**, 105 (5): 481-85. -Link-
- Butters, R. Andrew, Daniel W. Sacks, and Boyoung Seo. 2022. How Do National Firms Respond to Local Cost Shocks? American Economic Review**, 112 (5): 1737-72. -Link-
- Corbae D, D'Erasmo P. 2021. Capital Buffers in a Quantitative Model of Banking Industry Dynamics. **Econometrica**, 89: 2975–3023. -Link-
- Bernanke, Ben S. 2020. The New Tools of Monetary Policy. American Economic Review**, 110 (4): 943-83. -Link-

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**：2200 元/班/人
- **优惠方案**：(以下各项优惠不能叠加使用)
  - **专题课/现场班老学员报名：** 9 折，1980 元/人
  - **学生 (需提供学生证/卡照片)**：9 折，1980 元/人
  - **会员报名：** 85 折，1870 元/人
- **联系方式：**
  - 邮箱：wjx004@sina.com
  - 王老师：18903405450 (微信同号)；李老师：18636102467 (微信同号)

> **报名链接：** http://junquan18903405450.mikecrm.com/qWXY25e
>
> 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 (山西省太原市晋商银行南中环支行)
- **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：扫码支付**

> **温馨提示：** 扫码支付后，请将「**付款记录**」截屏发给王老师：18903405450 (微信同号)

7. 听课指南

7.1 软件和课件

听课软件：本次课程可以在手机，ipad ，平板以及 windows/Mac 系统的电脑上听课 (**台式机除外**)。

> **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。**

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. 助教招聘

说明和要求

- **名额：** 15 名
- **任务：**
  - A. **课前准备**：协助完成 2 篇介绍 R/Stata/Python 和计量经济学基础知识的文档；
  - 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)；
  - 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 R 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可以直接联系连老师直录。
- **截止时间：** 2022 年 9 月 25 日 (将于 9 月 27 日公布遴选结果于 课程主页，及 连享会主页 lianxh.cn)

> **申请链接：** https://www.wjx.top/vm/hngNjLi.aspx#

> 扫码填写助教申请资料：

> **课程主页：** https://gitee.com/lianxh/Rcourse

关于我们

- **Stata 连享会** 由中山大学连玉君老师团队创办，定期分享实证分析经验。
- 连享会-主页 和 知乎专栏，700+ 推文，实证分析不再抓狂。直播间 有很多视频课程，可以随时观看。
- ⛳ 课程主页: https://gitee.com/arlionn/Course