R语言在经济学中的应用

南开大学周恩来政府管理学院 吕小康

2017-07-12

R简介

R是一个免费自由且跨平台通用的统计计算与绘图软件。

- 它有Windows、Mac、Linux等版本，均可免费下载使用。

从R主页中选择download R链接可下载到对应操作系统的R安装程序。

- 打开链接后的网页会提示选择相应的CRAN镜像站。目前全球有超过一百个CRAN镜像站，用户可选择就近下载。

R与STATA等统计软件的区别

R为开源免费的软件，其他基本为商业付费软件。

- 如果你有钱，可以只选贵的、不选对的；但如果你没钱……

R是一种脚本语言，强调英文命令操作。

- R的学习比较费时、对汉字编码不友好，但掌握之后的自由性更强

R在数据可视化上的表现更佳，选择更丰富。

- R的统计绘图是它最有标志性的功能，可以制作达到出版的各种图形

R在经济学中的综合应用

R及与之相关的配套开源软件（如RStudio）已构成一个丰富的数据分析网络生态，具有同类软件很难同时满足的多种可能性。

用于课程教学

用于数据获取与预处理

用于数据建模

用于数据可视化

用于撰写学术报告

作为课堂教学的辅助软件

可以作为两门经济学基础课程的教学辅助软件

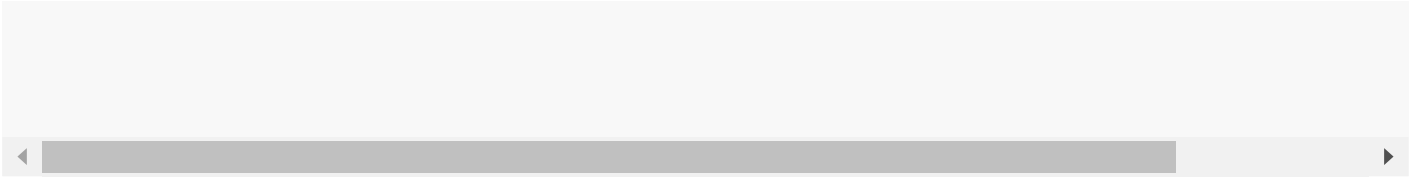
- 《概率论与数理统计》
- 《计量经济学》

我本人在清华大学出版社2017年出版的《R语言统计学基础》，内容差不多覆盖经济学类入门概率论与数理统计的教学要求，全程使用R作为分析和绘画软件。

作为课堂教学的辅助软件

这里使用一个经常在计量经济学中使用到的数据（　　）进行示例。这是美国 *Psychology Today* 杂志于1969年采集的关于婚外情的数据。该数据经常用于广义线性模型的示例。

OLS回归



OLS 回归

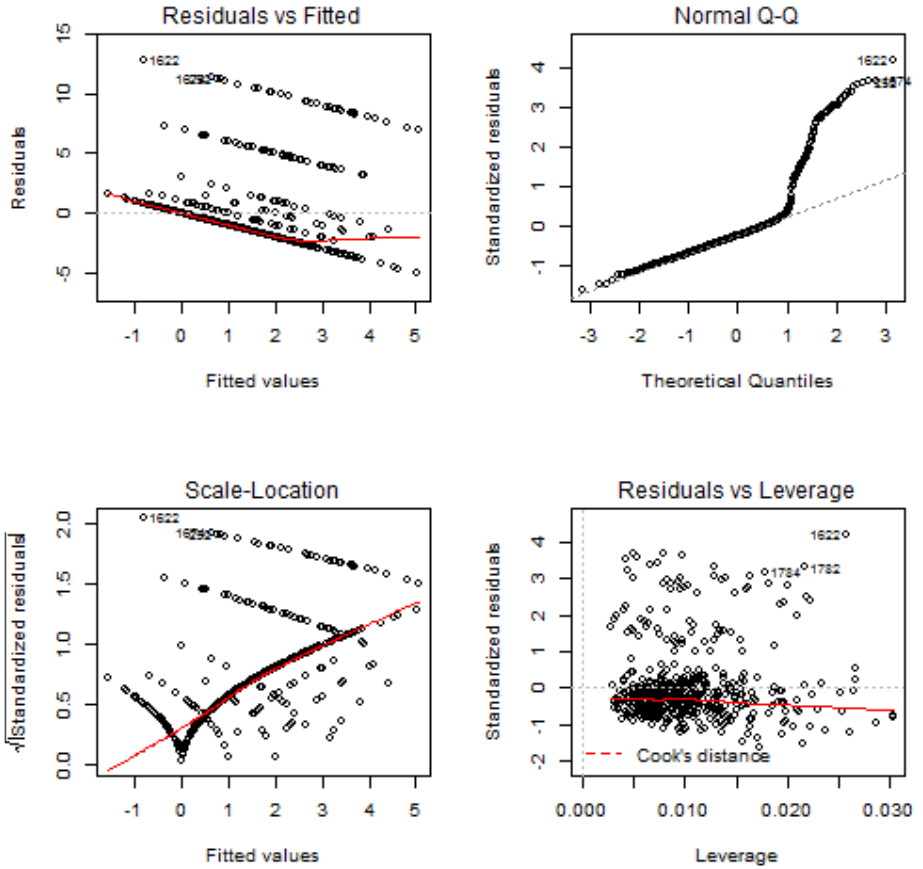
查看模型拟合值

```

```

查看模型残差

```

```

查看用于模型诊断的相关图示



广义线性模型

广义线性模型（Generalized Linear Models）的一般形式：

$$f(\mu_Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$$

其中

- $f(\mu_Y)$表示响应变量的条件均值的某种函数（称为连接函数，link function）。
- 此时对 Y 不再有服从正态分布的要求，而可以服从任何指数分布族中的某一分布。
- 设定好连接函数与分布类型后，就可以利用极大似然法通过多次迭代推导出各参数值。

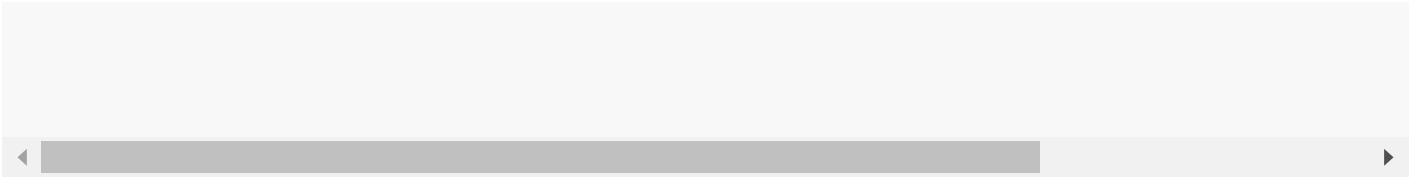
常用的广义线性模型

- Probit/Logistic 回归模型
- Poisson 回归模型
- Negative Binomial 回归模型
- Zero Inflation 回归模型
- Tobit 回归模型
- ……

这些都可通过R的相关函数方便求得。

广义线性模型

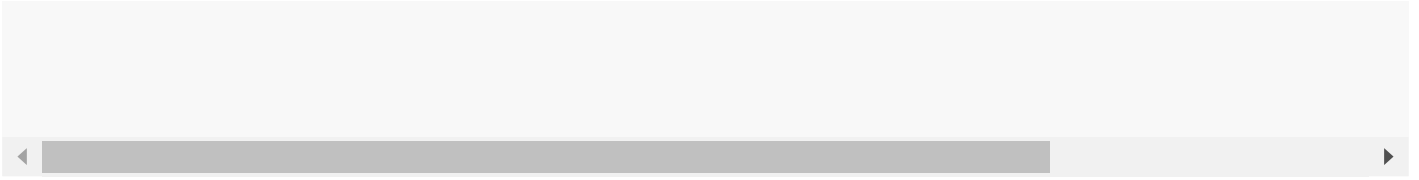
Probit 回归



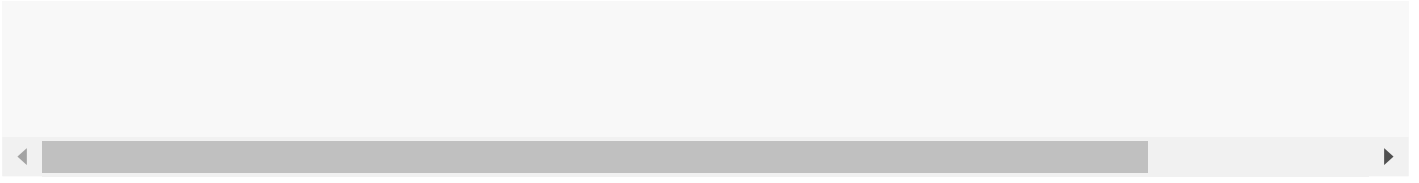
Probit 回归

查看模型拟合值

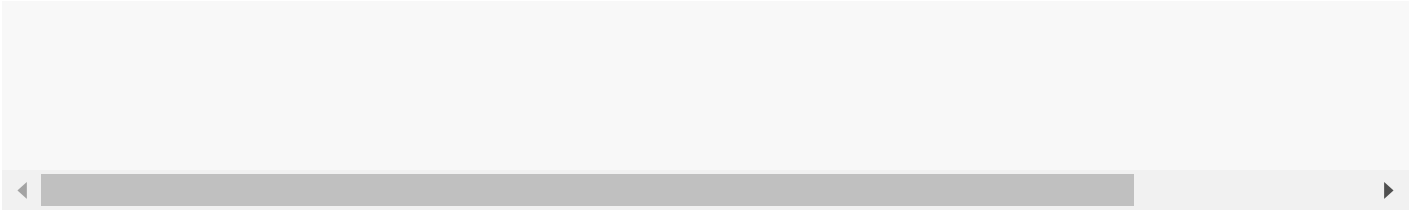
Logistic/Logit 回归



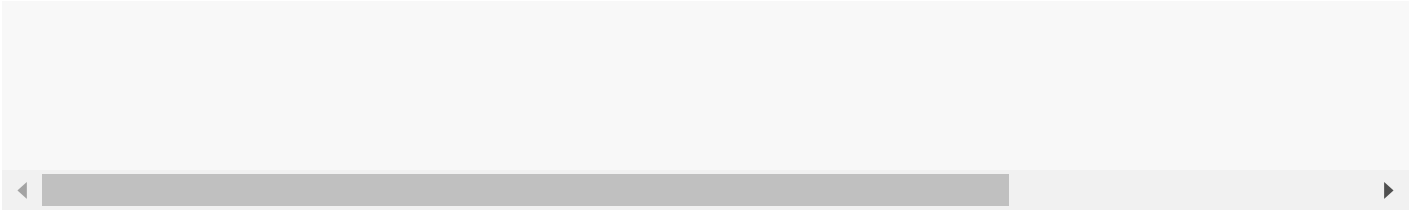
Poisson 回归



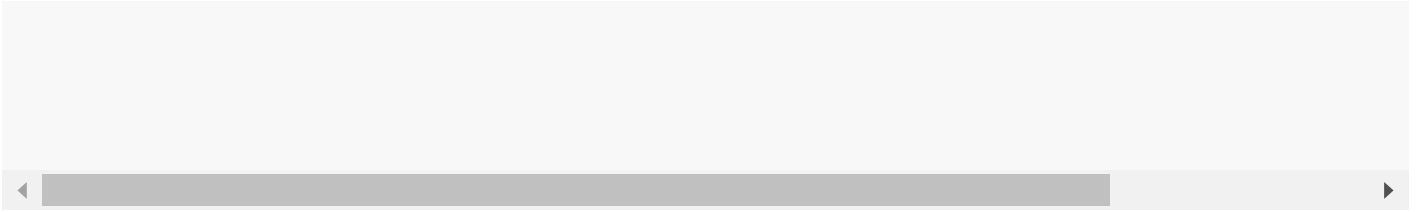
Negative Binomial 回归



Zero Inflation 回归



Tobit 回归



在数据获取、预处理和可视化中的应用

- tidyverse系列数据处理包
  - dplyr: 数据操纵
  - tidyr: 数据操纵
  - stringr: 文本数据操纵
  - rvest: 在线抓取文本
  - ……
- 可视化系列数据处理包
  - ggplot2
  - ggtheme
  - ggvis
  - shiny
  - wordcloud2
  - ……

数据处理示例1：一手问卷调查数据

我们项目组目前正在编制《中国医患社会心态调查问卷》，问卷已经基本完成编制并已进行预测试。对初测数据的统计分析工作正在进行。初测问卷使用问卷星填答，要求被调查者使用自身手机或在访问员的手机上完成填答。数据示例见Excel文件。

以下命令可简单地统计被试的地理位置分布。

地理位置信息分布结果

数据获取与处理示例2

政府工作报告抓取与分析

传统社会科学的量化分析以对数字数据（numeric data）的量化分析为主，对文本数据（text data）的分析较少。这主要是受研究工具的局限所致。

R及Python等开源软件的出现，很大程度上改变这种现状，使得文本分析成为当下社会科学研究的一大潮流。

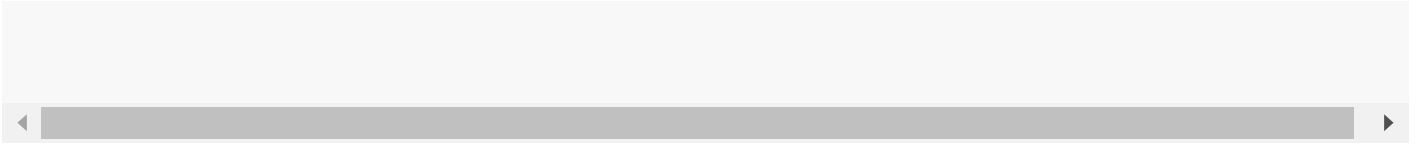
中国政府网提供了自1954年以来所有的政府工作报告全文。这里以中国政府工作报告（2017）为例做一简单的R语言示例（该示例得益于雪晴数据网陈堰平老师的讲座）。

政府工作报告的抓取与简单分析

2017政府工作报告

政府工作报告的抓取与简单分析

政府工作报告的抓取与简单分析



如何通过循环来遍历所有年份政府工作报告的链接，留待大家作为思考题。

提示如下：

数据获取、处理与可视化

经济学研究的常用数据、世界银行数据可使用两个R包获取：

- WDI
- wbstats

一个复制Hans Rosling的Gapminder软件的动态交互式气泡图

- Hans Rosling的TED演讲，中文翻译版
- R中的复制

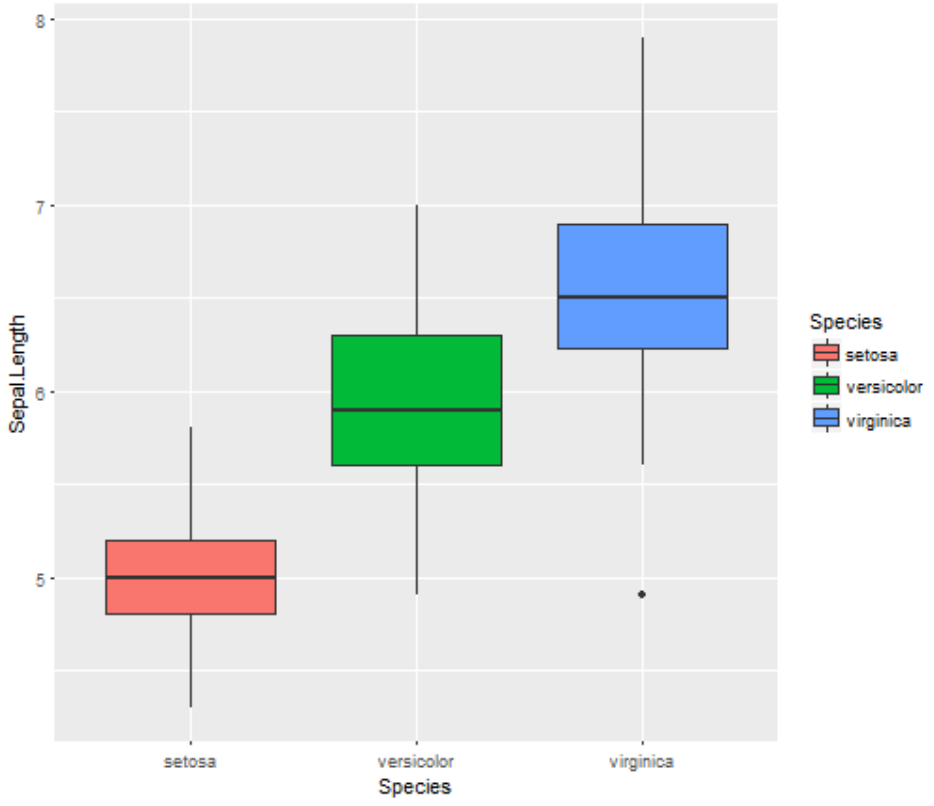
ggplot系列图形

利用ggplot2及ggthemes、ggsci等包，可便捷产生符合特定杂志风格的图形。

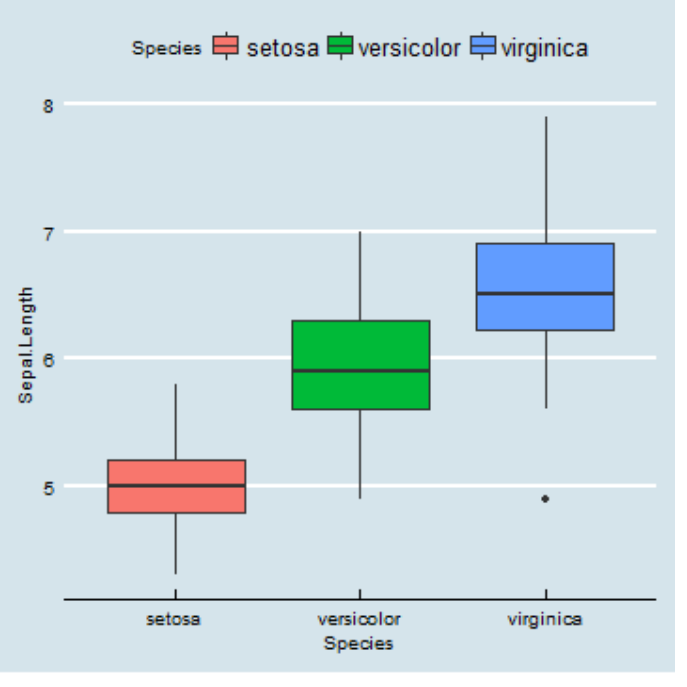
常用ggplot系列可视化包

- ggplot2
- ggthemes
- ggsci
- ggcorrplot
- ……

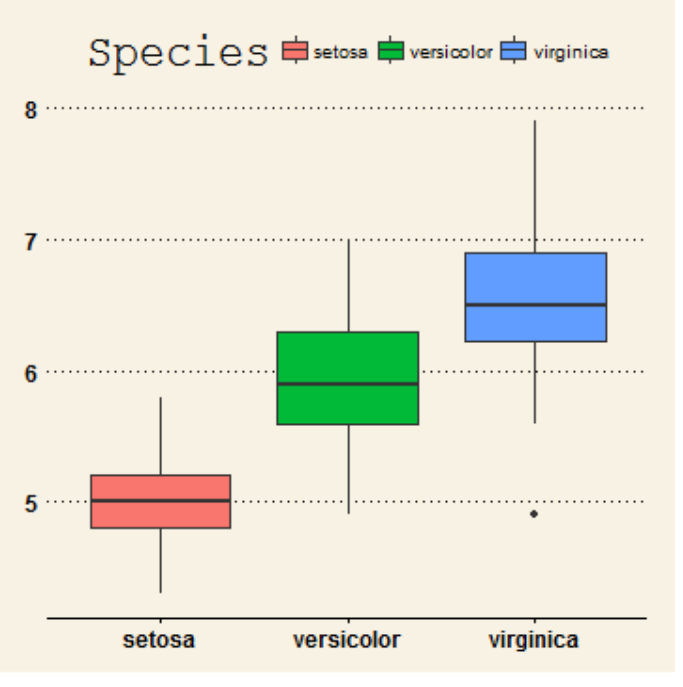
ggplot2 原始风格



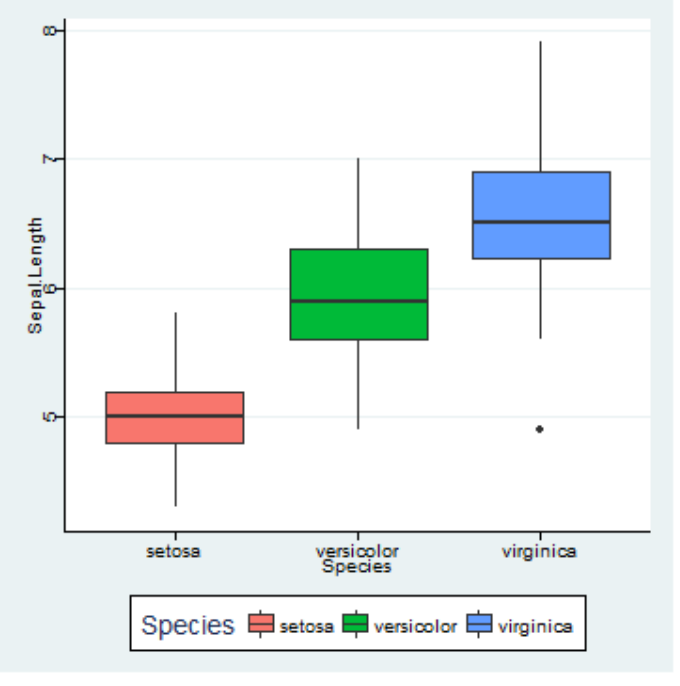
The Economist 风格图形



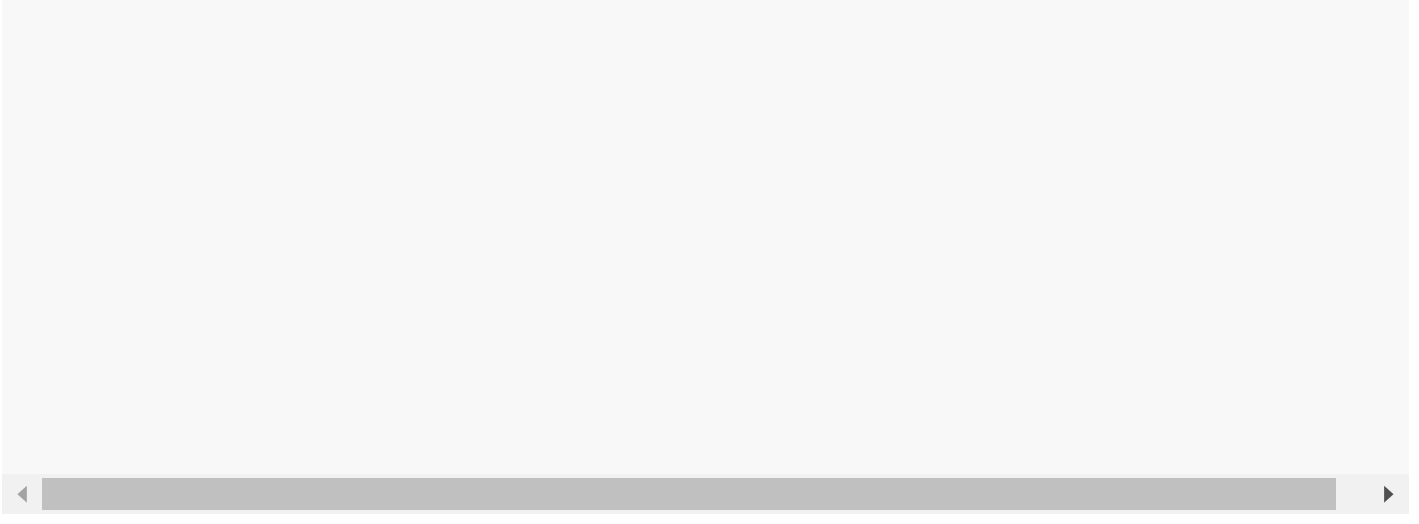
The Wallstreet Journal 风格图形



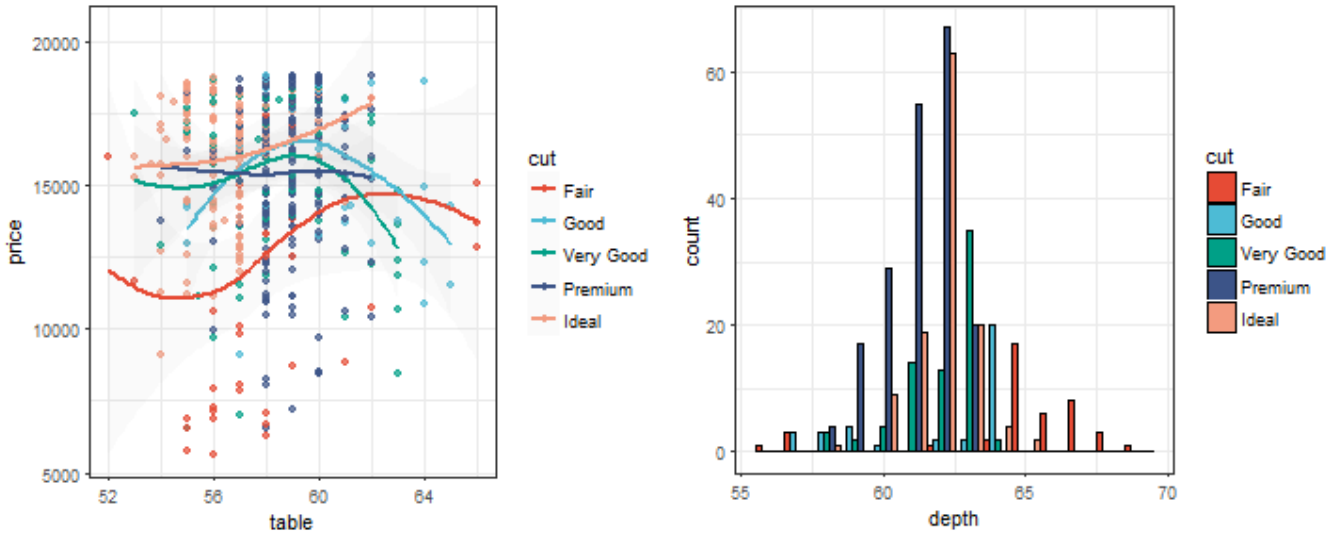
Stata风格图形


Sepal.Length
8
7
6
5
setosa
versicolor
virginica
Species
Species
setosa
versicolor
virginica

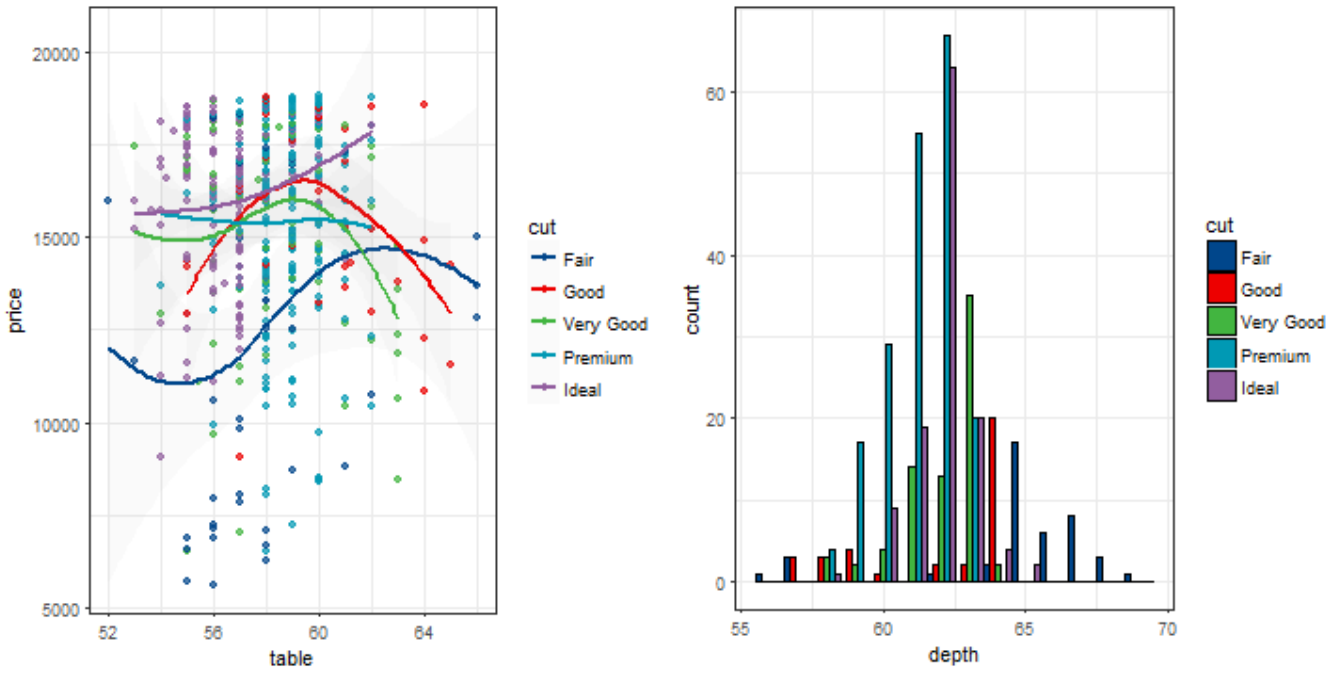
Nature 风格



Nature 风格


price
table
20000
15000
10000
5000
52
56
60
64
cut
Fair
Good
Very Good
Premium
Ideal
count
depth
60
40
20
0
55
60
65
70
cut
Fair
Good
Very Good
Premium
Ideal

Lancet 风格


price
20000
15000
10000
5000
52
56
60
64
table
cut
Fair
Good
Very Good
Premium
Ideal
count
60
40
20
0
55
60
65
70
depth
cut
Fair
Good
Very Good
Premium
Ideal

更多的R可视化图例

- RStudio图库
- ggplot2图库
- ggthemes示例
- ggsci示例

用于撰写学术报告

- rmarkdown: html 格式报告
- xaringan: html 格式幻灯片
- rticles: AER 等经济学类顶级刊物LaTeX模板
- stargazer: 生成LaTeX表格

常用资源

- 计量经济学中的常用 R 包索引：https://cran.r-project.org/web/views/Econometrics.html
- 用R做计量分析网站：https://econometricswithr.wordpress.com/
- Using R for Introductory Econometrics(Wooldridge 计量经济学导论配套R语言网站): http://www.urfie.net/
- bookdown官方网站：https://bookdown.org/home/
- *R for Data Science* 在线版本：http://r4ds.had.co.nz/

谢谢观看！

本幻灯片由谢益辉的 R 包 **xaringan** 生成

吕小康 副教授

南开大学周恩来政府管理学院

xkdog@126.com

本报告原始文档可从以下链接下载：

https://github.com/xkdog/StatsUsingR

简略版可从以下网址在线观看（图片未能正确显示）：

https://github.com/xkdog/StatsUsingR/blob/master/R4Eco201707.Rmd