# 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（`Affairs`）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。

```
if(!require(AER)) install.packages("AER")
data("Affairs")
head(Affairs)
```

```
##    affairs gender age yearsmarried children religiousness education
## 4        0   male  37        10.00       no             3        18
## 5        0 female  27         4.00       no             4        14
## 11       0 female  32        15.00      yes             1        12
## 16       0   male  57        15.00      yes             5        18
## 23       0   male  22         0.75       no             2        17
## 29       0 female  32         1.50       no             2        17
##    occupation rating
## 4           7      4
## 5           6      4
## 11          1      4
## 16          6      5
## 23          6      3
## 29          5      5
```

# OLS回归

```
fm_ols <- lm(affairs ~ age + yearsmarried + religiousness + occupati
data = Affairs)
summary(fm_ols)
```

```
##
## Call:
## lm(formula = affairs ~ age + yearsmarried + religiousness + occupation +
##     rating, data = Affairs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0382 -1.7076 -0.7780  0.2086 12.8134
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.60816    0.79660   7.040 5.31e-12 ***
## age           -0.05035    0.02211  -2.278   0.0231 *
## yearsmarried   0.16185    0.03690   4.387 1.36e-05 ***
## religiousness -0.47632    0.11131  -4.279 2.18e-05 ***
## occupation     0.10601    0.07110   1.491   0.1365
## rating        -0.71224    0.11829  -6.021 3.03e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.087 on 595 degrees of freedom
## Multiple R-squared:  0.1314,    Adjusted R-squared:  0.1241
## F-statistic:    18 on 5 and 595 DF,  p-value: < 2.2e-16
```

# OLS 回归

## 查看模型拟合值

```
fit <- fitted(fm_ols)
head(fit)
```

```
##           4           5          11          16          23          29
##  1.8279300  0.7779612  3.2055393 -0.1406526  2.1685688  0.2559938
```

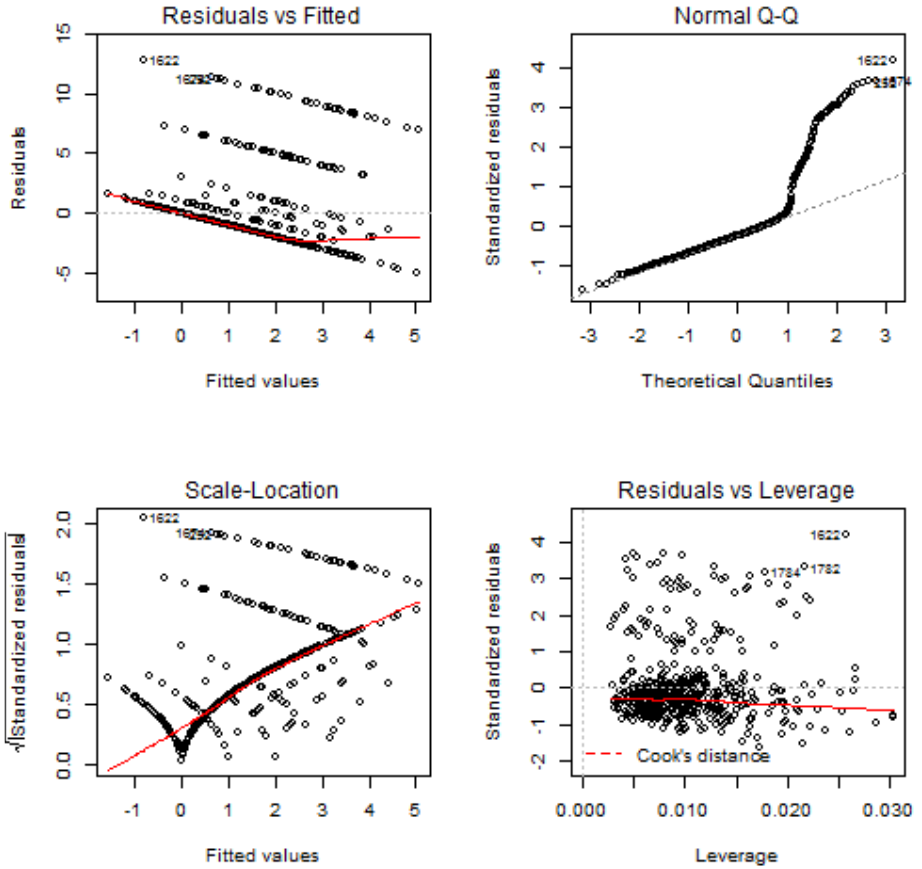
## 查看模型残差

```
re <- residuals(fm_ols)
head(re)
```

```
##           4           5          11          16          23          29
## -1.8279300 -0.7779612 -3.2055393  0.1406526 -2.1685688 -0.2559938
```

查看用于模型诊断的相关图示

```
opar <- par(no.readonly = T)
par(mfrow = c(2, 2))
plot(fm_ols)
```



# 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

```
fm_probit <- glm(I(affairs > 0) ~ age + yearsmarried + religiousness
data = Affairs, family = binomial(link = "probit"))
summary(fm_probit)
```

```
##
## Call:
## glm(formula = I(affairs > 0) ~ age + yearsmarried + religiousness +
##     occupation + rating, family = binomial(link = "probit"),
##     data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6143  -0.7678  -0.5841  -0.2368   2.4615
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.97667    0.36537   2.673 0.007516 **
## age            -0.02202    0.01032  -2.134 0.032821 *
## yearsmarried    0.05990    0.01712   3.499 0.000468 ***
## religiousness -0.18365    0.05172  -3.551 0.000384 ***
## occupation      0.03751    0.03285   1.142 0.253399
```

# Probit 回归

## 查看模型拟合值

```
fit <- fitted(fm_probit)
head(fit)
```

```
##           4           5          11          16          23          29
## 0.26779939 0.16359028 0.47303362 0.07518241 0.33574947 0.11842761
```

# Logistic/Logit 回归

```
fm_logit <- glm(I(affairs > 0) ~ age + yearsmarried + religiousness +
data = Affairs, family = binomial(link = "logit"))
summary(fm_logit)
```

```
##
## Call:
## glm(formula = I(affairs > 0) ~ age + yearsmarried + religiousness +
##     occupation + rating, family = binomial(link = "logit"), data = Affairs
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6633  -0.7500  -0.5750  -0.2691   2.4189
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.74904    0.62802   2.785 0.005352 **
## age            -0.04009    0.01785  -2.245 0.024748 *
## yearsmarried    0.10543    0.02952   3.572 0.000354 ***
## religiousness -0.32332    0.08950  -3.613 0.000303 ***
## occupation      0.07250    0.05677   1.277 0.201565
## rating         -0.46842    0.08928  -5.247 1.55e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
```

# Poisson 回归

```
fm_pois <- glm(affairs ~ age + yearsmarried + religiousness + occupa
data = Affairs, family = poisson)
summary(fm_pois)
```

```
##
## Call:
## glm(formula = affairs ~ age + yearsmarried + religiousness +
##     occupation + rating, family = poisson, data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5968  -1.5728  -1.1627  -0.7067   8.3473
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.533905   0.196924  12.867  < 2e-16 ***
## age           -0.032255   0.005851  -5.512 3.54e-08 ***
## yearsmarried   0.115698   0.009908  11.677  < 2e-16 ***
## religiousness -0.354037   0.030892 -11.460  < 2e-16 ***
## occupation     0.079828   0.019449   4.105 4.05e-05 ***
## rating        -0.409443   0.027381 -14.953  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2925.5  on 600  degrees of freedom
```

# Negative Binomial 回归

```
if(!require(MASS)) install.packages("MASS")
fm_nb <- glm.nb(affairs ~ age + yearsmarried + religiousness + occup
data = Affairs)
summary(fm_nb)
```

```
##
## Call:
## glm.nb(formula = affairs ~ age + yearsmarried + religiousness +
##     occupation + rating, data = Affairs, init.theta = 0.142555597,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1139  -0.8067  -0.6943  -0.4533   2.4548
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.189666   0.727326   3.011  0.00261 **
## age           -0.002624   0.020312  -0.129  0.89722
## yearsmarried   0.084819   0.034205   2.480  0.01315 *
## religiousness -0.422227   0.104081  -4.057 4.98e-05 ***
## occupation     0.060443   0.066262   0.912  0.36167
## rating        -0.431331   0.107449  -4.014 5.96e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.1426) family taken to be 1)
```

# Zero Inflation 回归

```
if(!require(pscl)) install.packages("pscl")
fm_zero <- zeroinfl(affairs ~ age + yearsmarried + religiousness + o
yearsmarried + religiousness + occupation + rating, data = Affairs)
summary(fm_zero)
```

```
##
## Call:
## zeroinfl(formula = affairs ~ age + yearsmarried + religiousness +
##     occupation + rating | age + yearsmarried + religiousness + occupation
##     rating, data = Affairs)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.4643 -0.5190 -0.3827 -0.2444 14.3993
##
## Count model coefficients (poisson with log link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.201940   0.210148  10.478  < 2e-16 ***
## age           -0.007238   0.006343  -1.141    0.254
## yearsmarried   0.049224   0.010990   4.479 7.50e-06 ***
## religiousness -0.131668   0.031154  -4.226 2.37e-05 ***
## occupation     0.016029   0.020101   0.797    0.425
## rating        -0.118672   0.028693  -4.136 3.53e-05 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.74075    0.62879  -2.768 0.005633 **
```

# Tobit 回归

```
library(AER)
fm_tobit <- tobit(affairs ~ age + yearsmarried + religiousness + occ
data = Affairs)
summary(fm_tobit)
```

```
##
## Call:
## tobit(formula = affairs ~ age + yearsmarried + religiousness +
##     occupation + rating, data = Affairs)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##            601            451            150              0
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    8.17420    2.74145   2.982  0.00287 **
## age           -0.17933    0.07909  -2.267  0.02337 *
## yearsmarried   0.55414    0.13452   4.119 3.80e-05 ***
## religiousness -1.68622    0.40375  -4.176 2.96e-05 ***
## occupation     0.32605    0.25442   1.282  0.20001
## rating        -2.28497    0.40783  -5.603 2.11e-08 ***
## Log(scale)     2.10986    0.06710  31.444  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 8.247
```

# 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。

```
library(readxl)
library(stringr)
library(tidyverse)
PDSurveyBasic <- read_excel("PDSurveyBasic.xlsx")
ip.location <- str_extract(PDSurveyBasic$ip, "(?<=\\().*(?=\\))") %>%
  str_split("-", n = 2, simplify = TRUE) %>%
  as_tibble %>%
  transmute(province = .[[1]], city = .[[2]]) %>%
  group_by(province) %>%
  summarise(n=n()) %>%
  arrange(desc(n))
```

# 地理位置信息分布结果

```
## # A tibble: 27 x 2
##    province     n
##       <chr> <int>
##  1     天津    97
##  2     云南    71
##  3     辽宁    63
##  4     新疆    57
##  5     河南    47
##  6     山东    44
##  7     北京    36
##  8     四川    34
##  9     重庆    34
## 10     山西    29
## # ... with 17 more rows
```

# 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

```
if (!require(rvest)) install.packages('rvest')
if (!require(wordcloud2)) install.packages('wordcloud2')
if (!require(jiebaR)) install.packages('jiebaR')
if (!require(stringr)) install.packages('stringr')
url2017 <-
  "http://www.gov.cn/premier/2017-03/16/content_5177940.htm"
report2017 <- read_html(url2017)
text2017 <- report2017 %>%
  html_nodes("p") %>%
  html_text() %>%
  paste(collapse = "")
writeLines(text2017, "report2017.txt")
library(jiebaR)
cutter <- worker(
  bylines = T,
  user = "./UsrWords.txt",
  stop_word = "./stopWords.txt",
  output = "report2017output.txt"
)
report_seg_file <- cutter["./report2017.txt"]
report_segged <-
  readLines("./report2017output.txt", encoding = "UTF-8")
report <- as.list(report_segged)
doc.list <- strsplit(as.character(report), split = " ")
term.table <- table(unlist(doc.list))
term.table <- sort(term.table, decreasing = TRUE)
```

# 政府工作报告的抓取与简单分析

```
head(vocabDF, 10)
```

```
##    Var1 Freq
## 1  发展 3375
## 2  改革 1836
## 3  推进 1755
## 4  建设 1458
## 5  经济 1404
## 6  推动 1134
## 7  加快 1080
## 8  政府 1080
## 9  创新  972
## 10 企业  945
```

# 政府工作报告的抓取与简单分析

```
library(wordcloud2)
wordcloud2(vocabDF, color = "random-light", backgroundColor = "grey")
```

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：

```
url <- "http://www.gov.cn/guowuyuan/baogao.htm"
reports <- read_html(url)
links <- reports %>%
  html_nodes(".history_report a") %>%
  html_attr("href") %>%
  str_trim()
head(links)
```

```
## [1] "http://www.gov.cn/premier/2017-03/16/content_5177940.htm"
## [2] "http://www.gov.cn/premier/2016-03/17/content_5054901.htm"
## [3] "http://www.gov.cn/guowuyuan/2015-03/16/content_2835101.htm"
## [4] "http://www.gov.cn/guowuyuan/2014-03/14/content_2638989.htm"
## [5] "http://www.gov.cn/premier/2013-03/19/content_2357136.htm"
## [6] "http://www.gov.cn/premier/2012-03/15/content_2067314.htm"
```

# 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

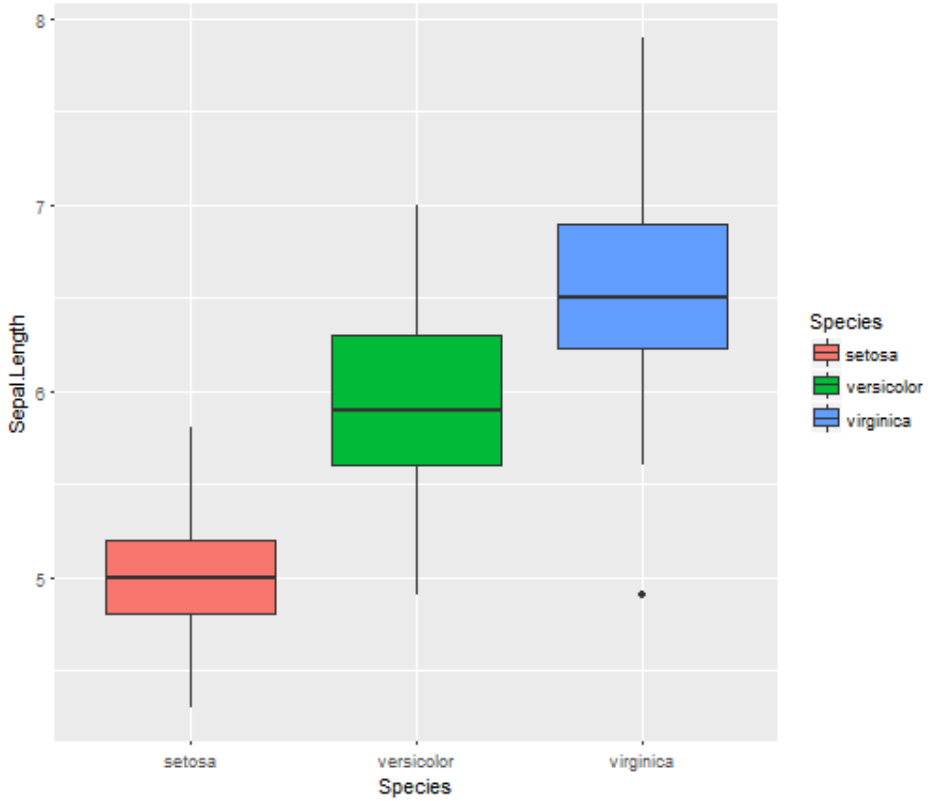
利用ggplot2及ggthemes、ggsci等包，可便捷产生符合特定杂志风格的图形。

# 常用ggplot系列可视化包

- ggplot2
- ggthemes
- ggsci
- ggcorrplot
- ……

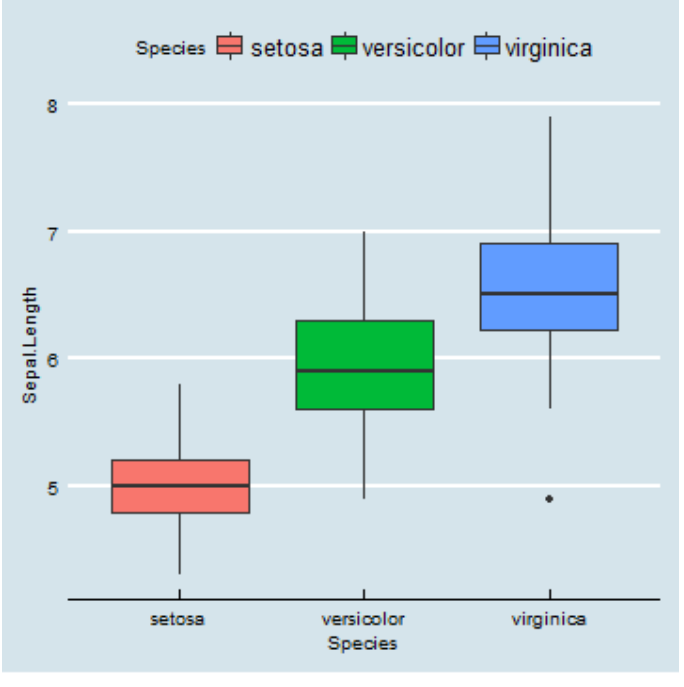
# ggplot2 原始风格

```
library(ggplot2)
ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Length, fill = Species))
```



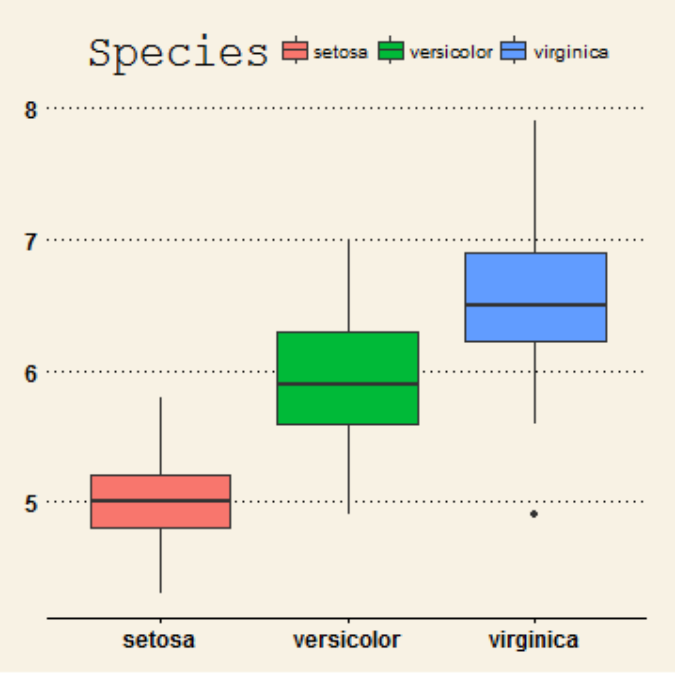
# *The Economist* 风格图形

```
ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Length, fill = Species)) +
  ggthemes::theme_economist()
```



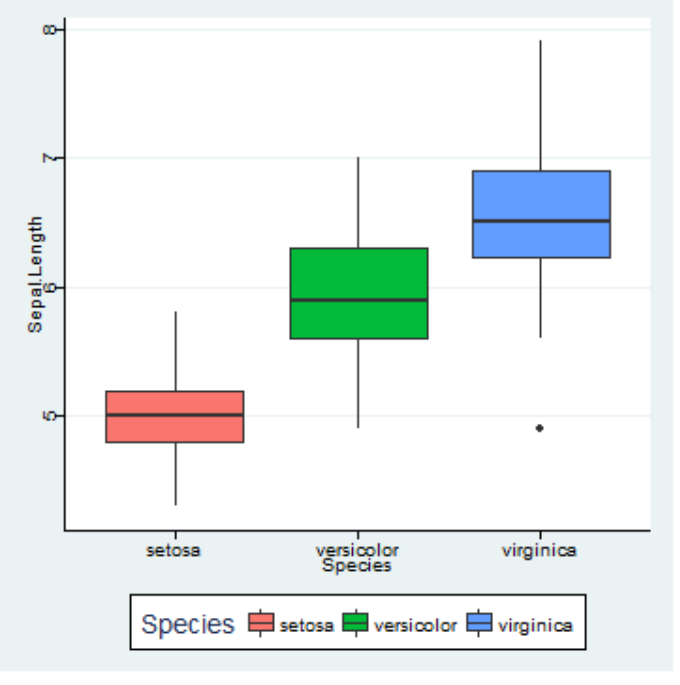
# *The Wallstreet Journal* 风格图形

```
ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Length, fill = Species)) +
  ggthemes::theme_wsj()
```



# Stata风格图形

```
ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Length, fill = Species)) +
  ggthemes::theme_stata()
```

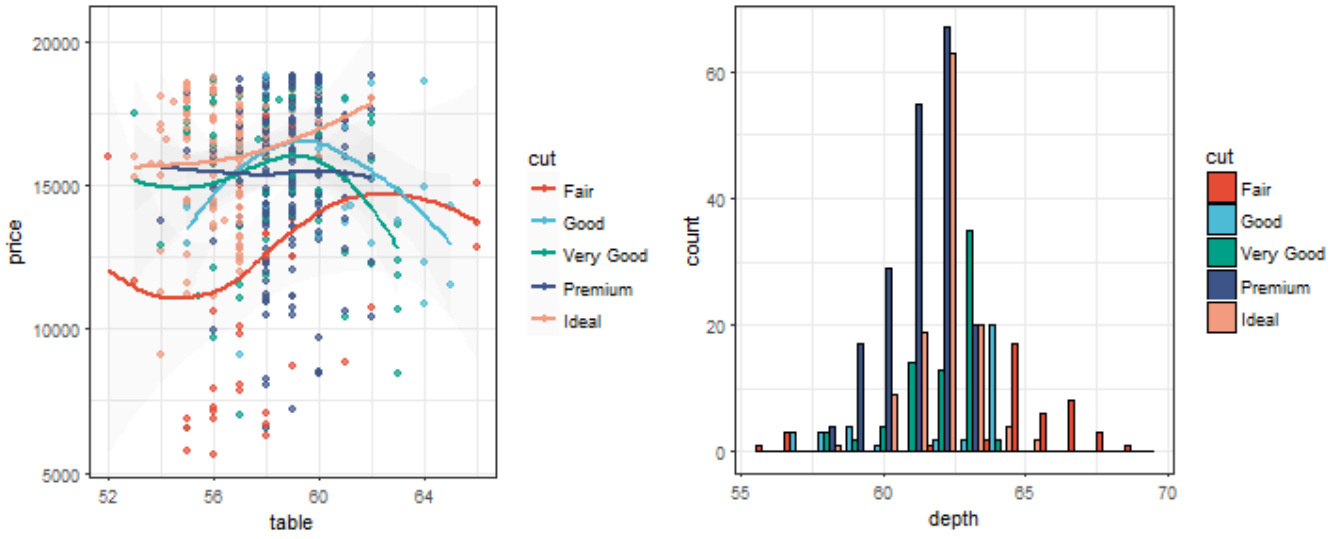


# *Nature* 风格

```
library("ggsci")
library("ggplot2")
library("gridExtra")
data("diamonds")
p1 = ggplot(subset(diamonds, carat >= 2.2),
       aes(x = table, y = price, colour = cut)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "loess", alpha = 0.05, size = 1, span = 1) +
  theme_bw()
p2 = ggplot(subset(diamonds, carat > 2.2 & depth > 55 & depth < 70),
       aes(x = depth, fill = cut)) +
  geom_histogram(colour = "black", binwidth = 1, position = "dodge") +
  theme_bw()
```

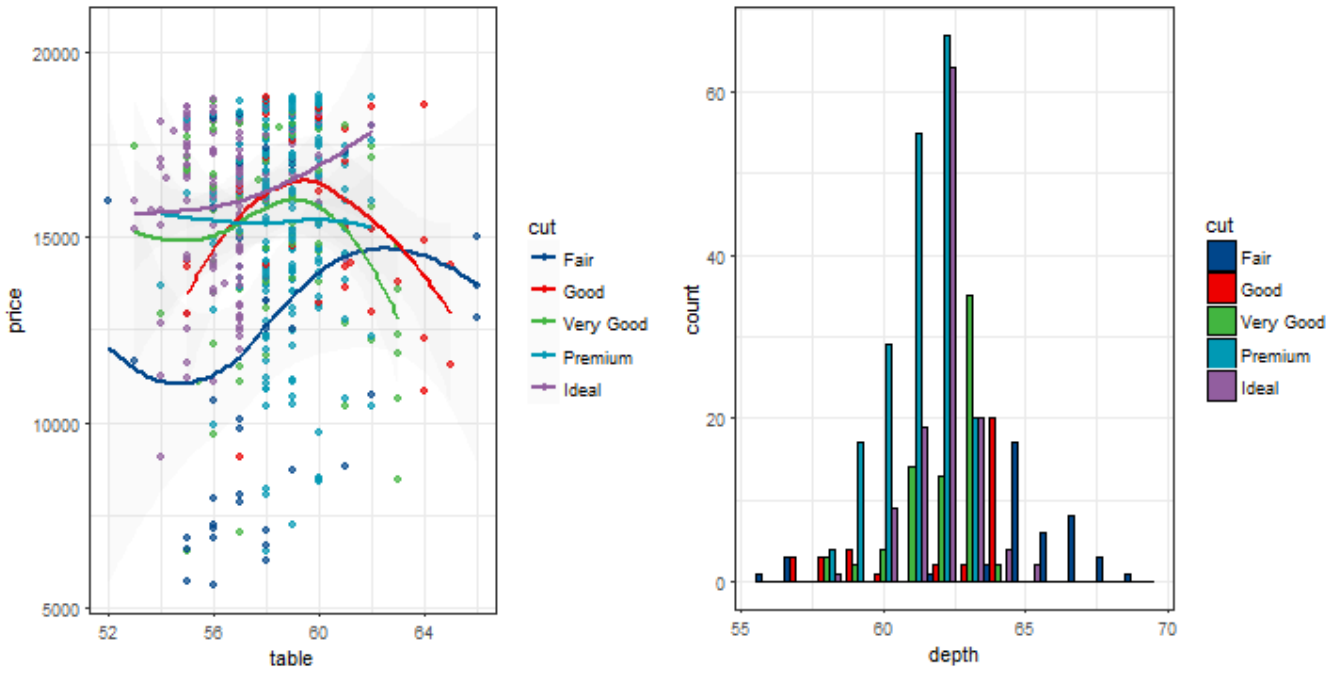
```
p1_npg = p1 + scale_color_npg()
p2_npg = p2 + scale_fill_npg()
grid.arrange(p1_npg, p2_npg, ncol = 2)
```

# Nature 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

# *Lancet* 风格

```
p1_lancet = p1 + scale_color_lancet()
p2_lancet = p2 + scale_fill_lancet()
grid.arrange(p1_lancet, p2_lancet, ncol = 2)
```



# 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 xkdog@126.com

南开大学周恩来政府管理学院

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

http://rpubs.com/xkdog/r4eco2017