关于作者

David S. Brown 是科罗拉多大学博尔德分校政治学教授和社会科学系主任。他获得了加利福尼亚大学洛杉矶分校的政治学博士学位，并且是科罗拉多大学博尔德分校行为科学研究所肯尼斯 · 博尔丁[1]的首届博士后研究员。在加入科罗拉多大学政治学系之前，他曾在莱斯大学担任助理教授。他研究的是比较政治学，侧重于体制及其对经济发展的影响。他的研究成果发表在《美国政治科学评论》（*American Political Science Review*）、《美国政治学期刊》（*American Journal of Political Science*）、《英国政治学期刊》（*British Journal of Political Science*）和《美国地理学家协会年鉴》（*Annals of the Association of American Geographers*）上。

1　译注：此人开辟了多个经济学和社会科学研究领域。

致谢

首先，我要感谢科罗拉多大学博尔德分校 PSCI 2075 班的学生，他们没让我幽默和过时的文化典故干扰到学习。他们的幽默、精神和毅力让课堂教学充满了乐趣。我还要感谢科罗拉多大学优秀的同事，特别是安迪 · 飞利浦（Andy Philips），他在本书的早期阶段提供了重要的反馈。最后，我要感谢 SAGE 的团队。Leah Fargotstein 给了我写作本书的机会，并耐心地帮助我将一些漫不经心的课程笔记转化为这本最终产物。切尔西 · 尼夫（Chelsea Neve）也参与了这个过程的重要部分，她就如何将散文转化为教学法提供了帮助。责任编辑克里斯蒂娜 · 韦斯特（Christina West）也做了大量的工作，她的耐心、技巧和友善让写作之旅的最后变得非常愉快。艾维 · 梅勒姆（Ivey Mellem）和丽贝卡 · 李（Rebecca Lee）在后端处理了非常复杂的流程，让前端显得简单又容易。尽管有其他人的这些启发、支持和帮助，但余下的失误都在我。

SAGE 和作者还要感谢以下审稿人对撰写本书提出的意见：

Zachary Albert, Brandeis University

Shavonne Arthurs, Seton Hall University

Hunter Bacot, University of North Carolina at Greensboro

Nathanial Bastian, Northwestern University

Salem Boumediene, University of Illinois Springfield

Scott Comparato, Southern Illinois University

Renato Corbetta, University of Alabama at Birmingham

Sarah Crocco, University of Maryland–College Park

Todd Daniel, Missouri State University

Eric Dunford, Georgetown University

Catherine Garcia, University of Nebraska–Lincoln

Jonathan Hack, Social Science Research Council

Troy Hooper, Texas Tech University Health Sciences Center

Ahmed Ibrahim, Johns Hopkins University

Whitt Kilburn, Grand Valley State University

David Lamb, University of Southern Florida

Alice Long, Penn State Shenango

Matt Miles, Brigham Young University–Idaho

Joseph Nedelec, University of Cincinnati

Carl Palmer, Illinois State University

Galen Papkov, Florida Gulf Coast University

Esther Pearson, Lasell University

Peter Peregrine, Lawrence University

Catherine Persall, St. Joseph's College

Matthew Phillips, University of North Carolina at Charlotte

Chris Prener, Saint Louis University

Alessandro Quartiroli, University of Wisconsin–La Cross

Jason Renn, Utah State University

Matthew Risler, Loras College

George Robinson, North Carolina A&T University

Andrew Rosenberg, University of Florida

Shayna Rusticus, Kwantlen Polytechnic University

Josh Ryan, Utah State University

Jennifer Samson, Arkansas Tech University

Jeffery Stone, University of California, Los Angeles

Joseph Szmania, Moravian College

Bradly Theissen, St. Ambrose University

Ches Thurber, Northern Illinois University

James Walke, Alabama A&M University

Lili Wang, Arizona State University

Kyle Woosnam, University of Georgia

Jingshun Zhang, Florida Gulf Coast University

习题答案

1 入门指南

参考答案

知识检验

1. a
2. b
3. b, c, d
4. b, c
5. a, d
6. b
7. a, b, d
8. b, c, d
9. a, d
10. b, d
11. b, c, d
12. b, d
13. b, d
14. b, c, d
15. b, d

数据分析与可视化练习

1. a, d
2. c
3. a, b, c
4. c, d
5. a, c
6. b, c
7. a, b, c
8. a, c
9. a, b
10. d

2 数据分析导论

参考答案

知识检验

1. a
2. a, b, c
3. a, b, c
4. a, b, c
5. b
6. c
7. d
8. b, d
9. a, c
10. b
11. a
12. b, c, d
13. c, d
14. d
15. a, c, d
16. a, b, c, d
17. c, d
18. a, c
19. a, b, c, d
20. a, b, c, d

数据分析与可视化练习

1. c
2. a
3. a, b, c, d
4. b
5. a
6. d
7. c
8. b
9. d
10. a, c

3 描述数据

参考答案

知识检验

1. b
2. d
3. b, c
4. 变量（a），变量（b），数据集（c），变量（d）
5. a, b
6. b, c
7. b
8. b, c, d
9. a, d
10. 问题（a），困惑（b），问题（c），困惑（d）
11. 错误（a），错误（b），正确（c），正确（d）
12. b
13. b, c, d
14. a
15. a, b
16. 都不是（a），可靠性（b），可靠性（c），可靠性（d）

数据分析与可视化练习

1. 数据集（a），变量（b），数据集（c），数据集（d）
2. `head(world)`

3. `head(dplyr::select(states, state, st, murderrate, hsdiploma))` 或者 `df <- states %>% head(dplyr::select(state, st, murderrate, hsdiploma))`
4. a, b, c, d
5. a
6. a, b
7. 困惑（a），问题（b），问题（c），问题（d）
8. 可靠性（a），可靠性（b），有效性（c），有效性（d）
9. a, d

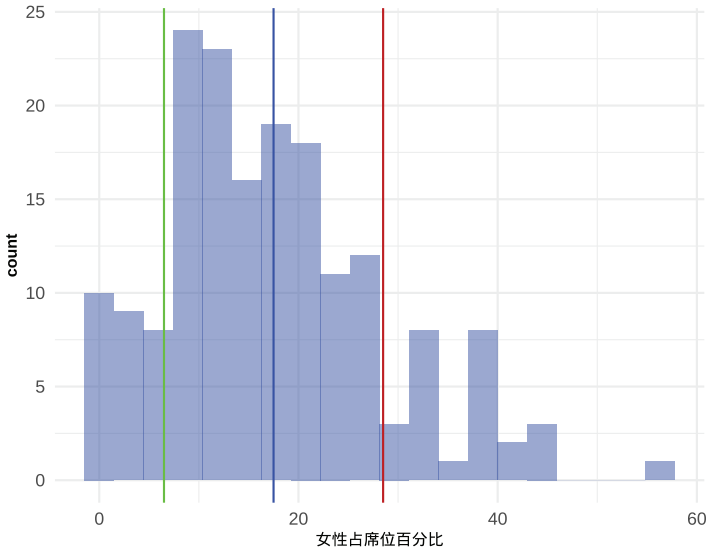
4 集中趋势和离散程度

参考答案

知识检验

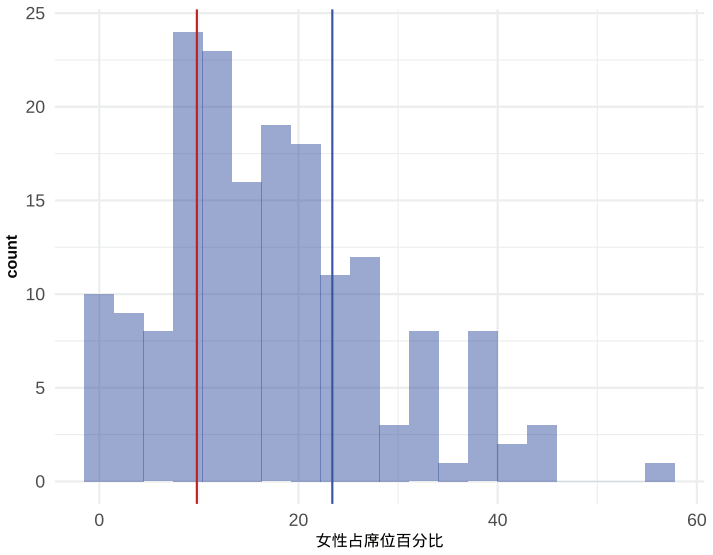
1. 众数（a），中位数（b），众数（c），平均数（d）
2. 60.6，67.5，两者（a）;16,326.8，9,863.2，中位数（b）;85.5，86.1，两者（c）;40,725，39,833，两者（d）
3. 坚定的民主党人（a），已婚（b），全职（c），白人（d）
4. 中位数（a），中位数（b），平均数（c），中位数（d）
5. 平均数增大，中位数移动（a）；平均数减小，中位数移动（b）；平均数增大，中位数不变（c）；平均数减小，中位数不变（d）
6. 122,910，17,934.3，18,456.39（a）;14.7，5.64，3.959（b）;100，40，27.1425（c）;100，34，38.05（d）

7.
```
ggplot(world, aes(womleg)) +
  geom_histogram(bins=20, fill = "#0000ff", alpha=0.5) +
  labs(title="world$womleg 变量直方图 ") +
  xlab(" 女性占席位百分比 ") +
  geom_vline(xintercept=(17.5 + 11), col="#bf0000") +
  geom_vline(xintercept=17.5, col="#0000ff") +
  geom_vline(xintercept=(17.5 - 11), col="#00ff00") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))
```



world$womleg 变量直方图

8.
```
ggplot(world, aes(womleg)) +
  geom_histogram(bins=20, fill = "#0000ff", alpha=0.5) +
  labs(title="world$womleg 变量四分位距 ") +
  xlab(" 女性占席位百分比 ") +
  geom_vline(xintercept=9.8, col="#bf0000") +
  geom_vline(xintercept=23.4, col="#0000ff") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold")) +
  theme(axis.title = element_text(size = 8, face = "bold"))
```



world$womleg 变量四分位距

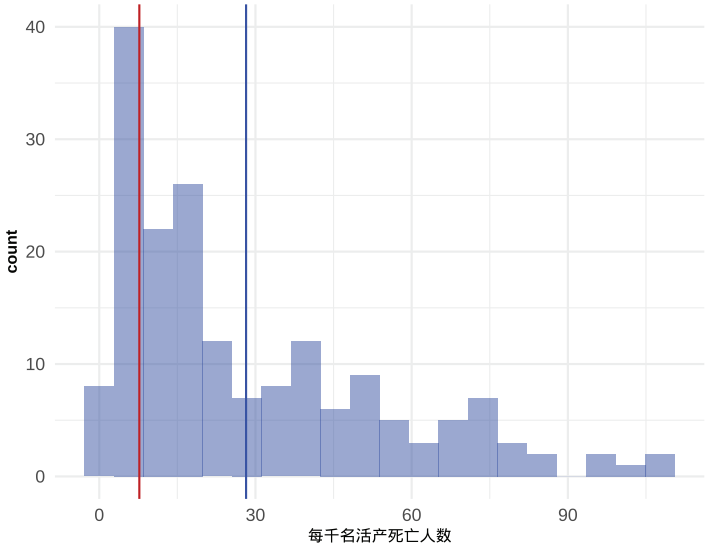
9. 四分位距（a），四分位距（b），标准差（c），四分位距（d）

10. 增大，增大（a）；增大，增大（b）；不变，增大（c）；不变，增大（d）

11. 四分位距

12. 7.7，28.2

```
ggplot(world, aes(inf)) +
  geom_histogram(bins=20, fill = "#0000ff", alpha=0.5) +
  labs(title="婴儿死亡率直方图") +
  xlab("每千名活产死亡人数") +
  geom_vline(xintercept=7.7, col="#bf0000") +
  geom_vline(xintercept=28.2, col="#0000ff") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold")) +
  theme(axis.title = element_text(size = 8, face = "bold"))
```



婴儿死亡率直方图

13. 四分位距

数据分析与可视化练习

1. 均可（a），均可（b），中位数（c），中位数（d）
2. extremely important（非常重要）
3. 均可（a），均可（b），四分位距（c），四分位距（d）
4. 1.49（a），23.96（b），36.49（c），2.55（d）
5. b
6. c
7. 改变直方图颜色的深浅程度
8. 红色 = 平均数，蓝色 = 中位数
9. c
10. b, a, c

5 数据的单变量和双变量描述

参考答案

知识检验

1. a, b, c, d
2. 离散程度（a），集中趋势（b），离散程度（c），离散程度（d）
3. 分类变量（a），连续变量（b），分类变量（c），连续变量（d）
4. 议会制民主（Parliamentary Democracy）（a），皇室独裁制（Royal Dictatorship）（b），35（c），议会制民主（Parliamentary Democracy）（d）
5. 8（a），是（b），否（c），0（d）
6. MN（a）；HI（b）；NE、MS、ID、VT 或 LA（c）；WI 和 WV（d）
7. c
8. a, d
9. 散点图（a），马赛克图（b），箱线图（c）
10. 减小（a）；LA、MD、NY（b）；低支出和低凶杀率（c）；*y*= 凶杀率，*x*= 州学生平均支出（d）
11. 皇室独裁（a）;皇室独裁（b）;SVK、PRT、FRA、ISL、FIN、AUT、IRL（c）;SGP（d）
12. a, b
13. 是（a），是（b），高收入和高凶杀率（c），MS（d）

数据分析与可视化练习

1. a, b, c
2. 白人（a），中东人（b），8（c），863（d）
3. 17（a），60（b），24（c），否（右偏态）（d）
4. 是（a），15（b），100（c），60（d）

5. 经济增长（a），幸福感（b），婴儿死亡率（c），教育改革（d）
6. 四分卫战术（a），周几 / 时间（b），时间（c），投入（d）
7. c
8. c
9. c

6 数据变换

参考答案

知识检验

1. a, b, c
2. a, b, c
3. b, d
4. a, b, d
5. a, c
6. a, b, c
7. a, b, c
8. a, d
9. a, b, c
10. c, d
11. a, d
12. b, d
13. b, d
14. 3（a），4（b），2（c），1（d）

数据分析与可视化练习

1. `nes$agevar <- cut(nes$age, c(30,40))`
   `nes$agevar` 变量是通过 `nes$age` 变量创建的新变量，用来给年龄分组。
2. 首先给 `nes$race` 变量重新分类。

   ```
   nes$whitevar <- ifelse(nes$race=="White", 1, 0)
   ```

   然后给新变量 `nes$whitevar` 加上标签。

   ```
   nes$whitevar <- factor(nes$whitevar, levels = c(0,1),
   labels = c("非白人",  "白人"))
   ```

3. 分类的（a），连续的（b），分类的（c），分类的（d）

4. 分类的（a），连续的（b），分类的（c），连续的（d）
5. 对数版
6. 1（a），0（b），6（c），3（d）
7. 1,000（a），10,000（b），10（c），1（d）
8. a, d
9. a, b, c
10. a

7 数据展示的一些原则

参考答案

知识检验

1. a, b
2. c, d
3. a, b, c, d
4. d
5. a, b, c
6. c
7. b
8. a
9. a, b, c, d
10. a, b, c
11. a, b, c, d
12. b, c, d
13. a, b, d

数据分析与可视化练习

1. `col=ifelse(states=="KS" | states=="NE", "blue", "grey")`
2. `col=ifelse(iso3c=="BRA" | iso3c=="RUS", "red", "grey")`
3. c
4. b, d

5. a
6. a, c
7. d
8. c
9. b
10. a, b, c, d

8 概率论精要

参考答案

知识检验

1. 总体（a），总体（b），样本（c），样本（d）
2. a, d
3. a, b, c, d
4. a, b, c
5. a, b
6. b, c
7. b
8. a
9. c, d
10. b, c, d
11. a, b, c, d
12. a, b, c, d
13. a, b, d
14. b
15. a
16. c
17. c
18. 0.87（a），0.82（b），0.14（c），0.07（d）

数据分析与可视化练习

1. b, c
2. 总体（a），样本（b），总体（c），总体（d）
3. a, d
4. a, c
5. a, b, c
6. b, d
7. a, b
8. c
9. a, b
10. a, b

9 置信区间与假设检验

参考答案

知识检验

1. (0.4015, 0.4575)，0.4295
2. (0.208, 0.232)，0 .22
3. (1.86, 2.058)，是的
4. b
5. (4.71, 7.29)，是的
6. c
7. b, c, d
8. b, c
9. d
10. 4.36（a），1.90（b），(3.41, 5.17)（c），是的（d）
11. 39,954（a），6,418（b），(3,988, 4,223)（c），是的（d）
12. 0.5992（a），0.076（b），(0.599, 0.621)（c），(0.579, 0.641)（d）
13. 1,784，拒绝

14. 1.786，拒绝

15. 3,045，未拒绝

16. a

17. b, c, d

18. a, b, c, d

数据分析与可视化练习

1. (0.489, 0.547), 0.5178
2. (0.82, 0.354), 0.318
3. 17.9（a），9.58（b），(14.64, 21.16)（c），(12.54, 23.26)（d）
4. 7.34（a），1.87（b），(6.7, 7.98)（c），(6.3, 8.3)（d）
5. 17.3（a），9.7（b），(14.08, 20.56)（c），(12.18, 22.46)（d）
6. 15.3，拒绝
7. 1.26，未拒绝
8. 1.6，未拒绝
9. b, c, d

10 进行比较

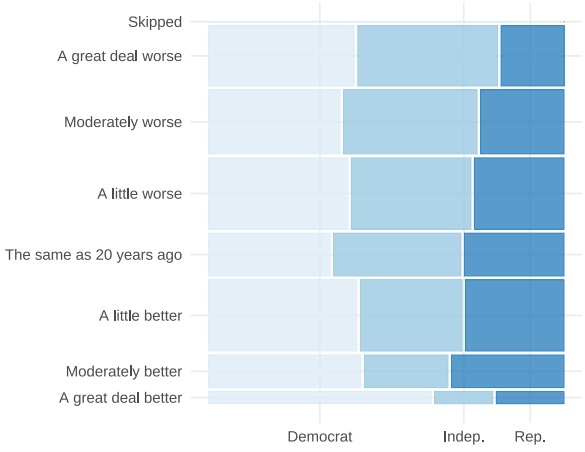
参考答案

知识检验

1. a, b, c, d
2. a, c
3. b, c
4. c, d

5. 马赛克图，是

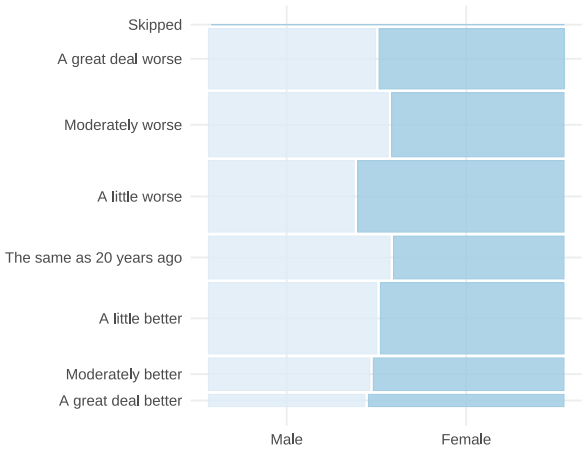
```
nanes$finwell <- droplevels(nanes$finwell)
ggplot(nanes) +
  geom_mosaic(aes(x = product(pid3.new, finwell),
                  fill=pid3.new,na.rm=TRUE)) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  xlab("") +
  ylab("") +
  ggtitle("民主党人表现不错") +
  scale_fill_brewer(palette="Blues") +
  theme(legend.position="none") +
  coord_flip()
```



民主党人表现不错

6. 马赛克图，否

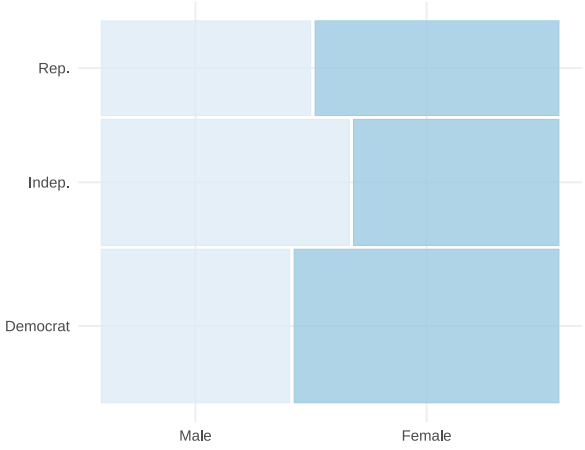
```
nanes$finwell <- droplevels(nanes$finwell)
ggplot(nanes) +
  geom_mosaic(aes(x = product(gender, finwell),
                  fill=gender,na.rm=TRUE)) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  xlab("") +
  ylab("") +
  ggtitle("性别与财务状况") +
  scale_fill_brewer(palette="Blues") +
  theme(legend.position="none") +
  coord_flip()
```



性别与财务状况

7. 马赛克图，是

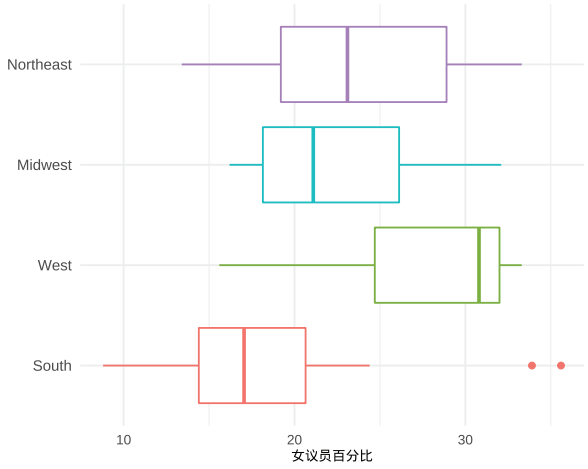
```
nanes$finwell <- droplevels(nanes$finwell)
ggplot(nanes) +
  geom_mosaic(aes(x = product(gender, pid3.new),
                  fill=gender,na.rm=TRUE)) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  xlab("") +
  ylab("") +
  ggtitle(" 无党派人士男性居多 ") +
  scale_fill_brewer(palette="Blues") +
  theme(legend.position="none") +
  coord_flip()
```



无党派人士男性居多

8. 箱线图，最高：西部（West），最低：南部（South）

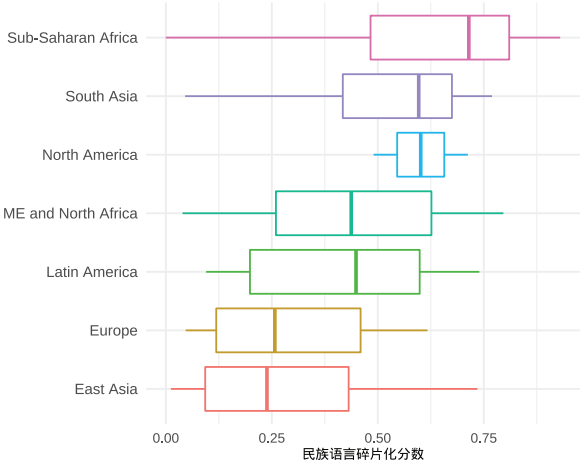
```
ggplot(states, aes(region, femleg, col=region)) +
  geom_boxplot() +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(axis.text.x = element_text(size=8, vjust=0.7),
        legend.position="none") +
  ggtitle("各地区立法机构中的女性") +
  ylab("女议员百分比")+
  xlab("")  +
  coord_flip()
```



各地区立法机构中的女性

9. 箱线图，最高：非洲（Africa），最低：东亚（East Asia）

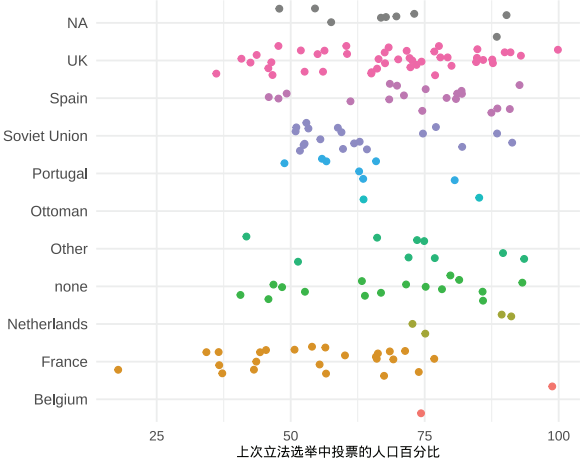
```
ggplot(world, aes(region, ethfrac, col=region)) +
  geom_boxplot() +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(axis.text.x = element_text(size=8, vjust=0.7),
        legend.position="none") +
  ggtitle("各地区的民族语言碎片化情况") +
  ylab("民族语言碎片化分数") +
  xlab("")  +
  coord_flip()
```



各地区的民族语言碎片化情况

10. 抖动图，法国的

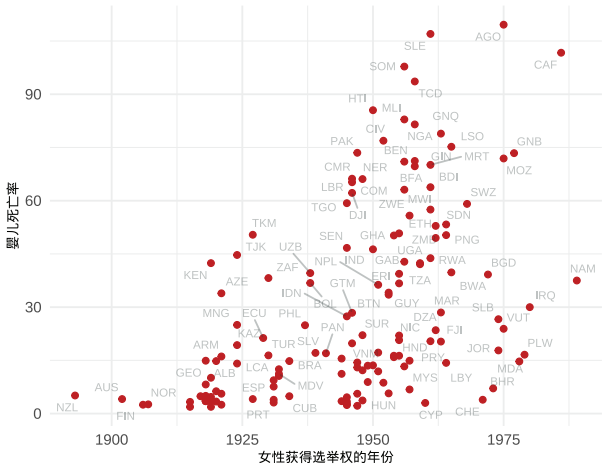
```
ggplot(world, aes(colony, turnout, col=colony)) +
  geom_jitter() +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(axis.text.x = element_text(size=8, vjust=0.7),
        legend.position="none") +
  ggtitle("投票率与殖民地历史") +
  ylab("上次立法选举中投票的人口百分比") +
  xlab("")  +
  coord_flip()
```



投票率与殖民地历史

11. 是；否；是；CAF, AGO, NAM

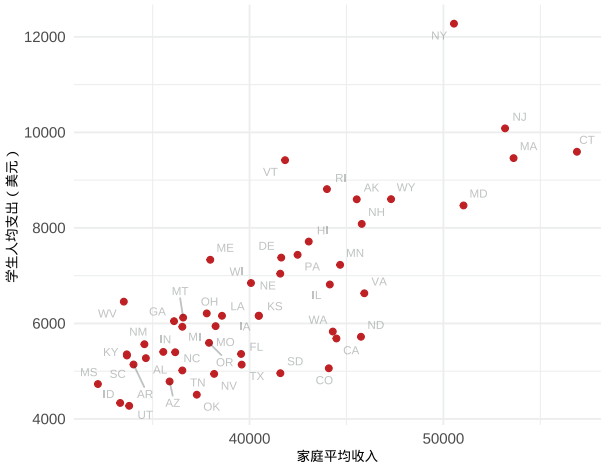
```
ggplot(world, aes(womyear, inf)) +
  geom_point(col="#bf0000") +
  geom_text_repel(size=2.4, vjust=0.5, col="#263333", alpha = 0.3,
                  aes(label= iso3c)) +
  ggtitle("女性选举权和婴儿死亡率") +
  ylab("婴儿死亡率")+
  xlab("女性获得选举权的年份") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))
```



女性选举权和婴儿死亡率

12. 是；是；是；NY

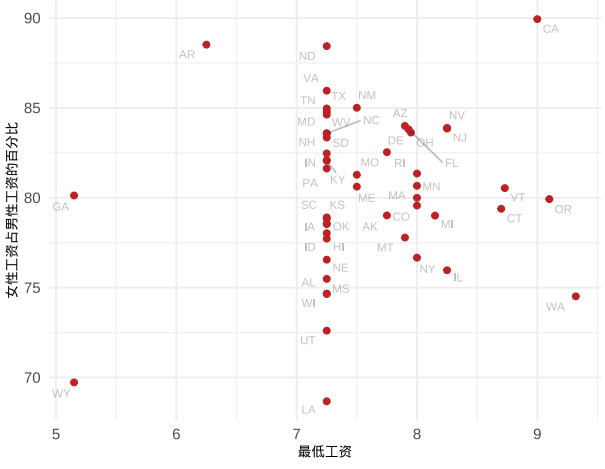
```
ggplot(states, aes(inc, stuspend)) +
  geom_point(col="#bf0000") +
  geom_text_repel(size=2.4, vjust=0.5, col="#263333", alpha = 0.3,
                  aes(label= st)) +
  ggtitle("收入与学生教育支出") +
  ylab("学生人均支出（美元）") +
  xlab("家庭平均收入") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))
```



收入与学生教育支出

13. 否；否；是；CA, WA, WY, AR, GA

```
ggplot(states, aes(minwage, percwom)) +
  geom_point(col="#bf0000") +
  geom_text_repel(size=2.4, vjust=0.4, col="#263333", alpha = 0.3,
                  aes(label= st)) +
  ggtitle("女性与男性的工资之比和最低工资") +
  ylab("女性工资占男性工资的百分比") +
  xlab("最低工资") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))
```



女性与男性的工资之比和最低工资

14. b, c, d

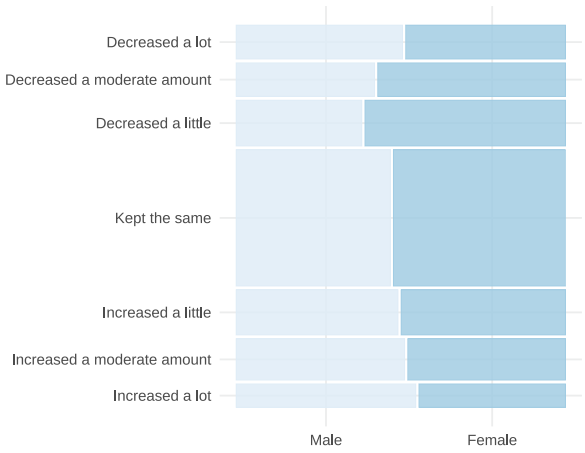
15. a, c, d

16. 图 10-6（a），图 10-8（b），图 10-9（c），图 10-9（d）

数据分析与可视化练习

1. 马赛克图，否

```
ggplot(data = na.omit(nes)) +
  geom_mosaic(aes(x = product(gender, immig_numb),
                  fill=gender,na.rm=TRUE)) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  xlab("") +
  ylab("") +
  coord_flip() +
  ggtitle("无性别差异") +
  scale_fill_brewer(palette="Blues") +
  theme(legend.position="none")
```

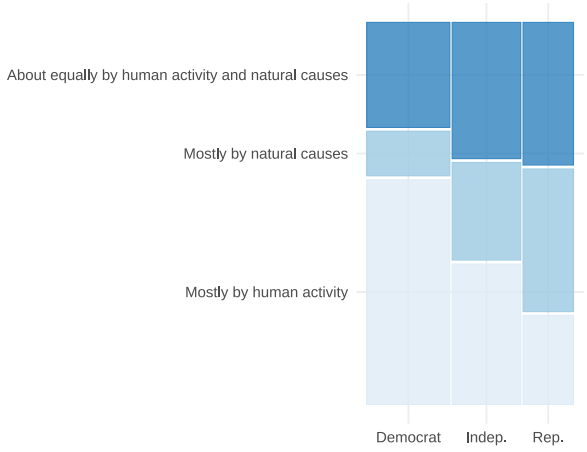


无性别差异

2. 马赛克图，是

```
ggplot(data = na.omit(nes)) +
  geom_mosaic(aes(x = product(warmcause, pid3.new),
                  fill=warmcause,na.rm=TRUE)) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
```

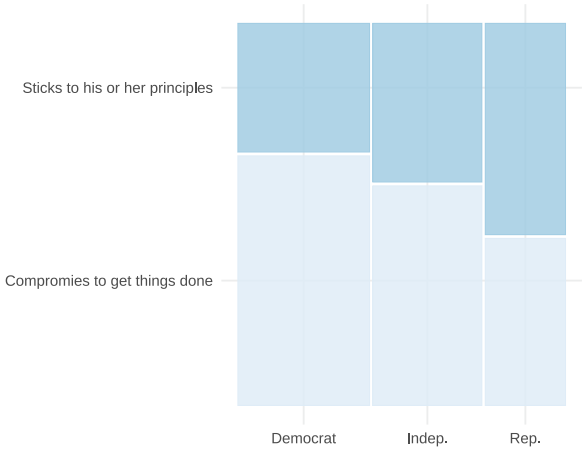
```
        axis.title = element_text(size = 8, face = "bold")) +
  xlab("") +
  ylab("") +
  ggtitle("民主党将气候变化归咎于人类活动") +
  scale_fill_brewer(palette="Blues") +
  theme(legend.position="none")
```



民主党将气候变化归咎于人类活动

3. 马赛克图，是

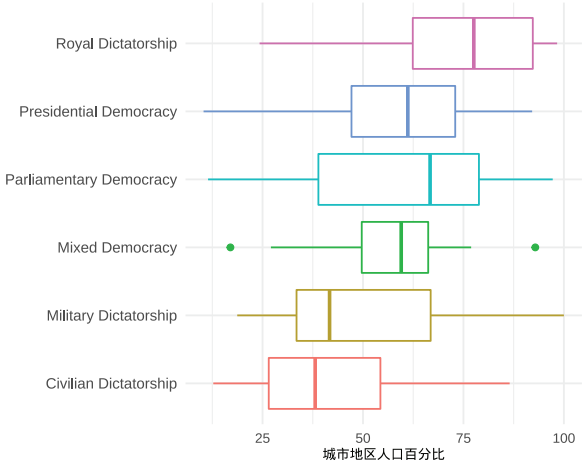
```
ggplot(data = na.omit(nes)) +
  geom_mosaic(aes(x = product(compromise, pid3.new),
                  fill=compromise,na.rm=TRUE)) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  xlab("") +
  ylab("") +
  ggtitle("民主党人赞同妥协") +
  scale_fill_brewer(palette="Blues") +
  theme(legend.position="none")
```



民主党人赞同妥协

4. 箱线图，最多：皇室独裁（Royal Dictatorship），最少：文官独裁（Civilian Dictatorship）

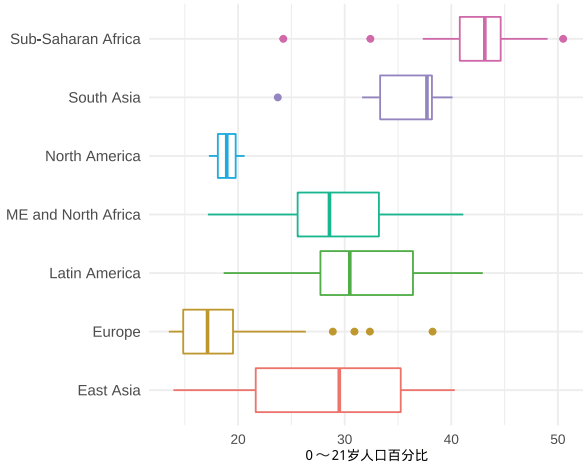
```
ggplot(world, aes(regime, urban, col=regime)) +
  geom_boxplot() +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(axis.text.x = element_text(size=8, vjust=.7),
        legend.position="none") +
  ggtitle("皇室独裁城市人口占比高") +
  ylab("城市地区人口百分比") +
  xlab("") +
  coord_flip()
```



皇室独裁城市人口占比高

5. 箱线图，最年轻：非洲（Africa），最年老：欧洲（Europe）

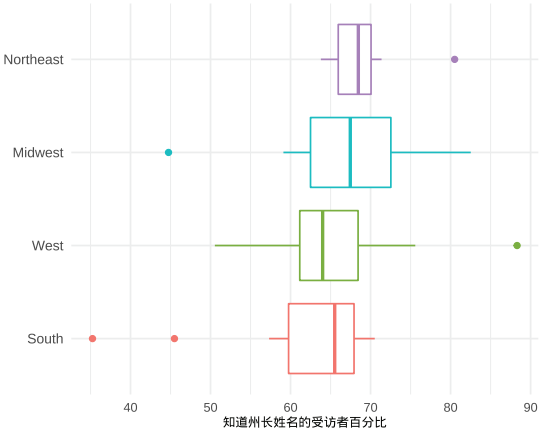
```
ggplot(world, aes(region, young, col=region)) +
  geom_boxplot() +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(axis.text.x = element_text(size=8, vjust=0.7),
        legend.position="none") +
  ggtitle("世界上人口最年轻和最年老的地区") +
  ylab("0~21岁人口百分比") +
  xlab("")  +
  coord_flip()
```



世界上人口最年轻和最年老的地区

6. 箱线图，最多：东北部（Northeast），最少：西部（West）

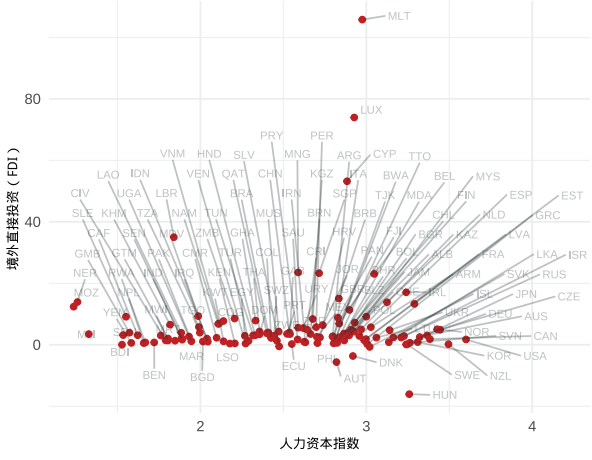
```
ggplot(states, aes(region, knowgov, col=region)) +
  geom_boxplot() +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(axis.text.x = element_text(size=8, vjust=0.7),
        legend.position="none") +
  ggtitle("东北部政治知识最丰富") +
  ylab("知道州长姓名的受访者百分比") +
  xlab("")  +
  coord_flip()
```



东北部政治知识最丰富

7. 否（a），否（b），是（c），马耳他（Malta）、卢森堡（Luxembourg）、塞浦路斯（Cypress）（d）

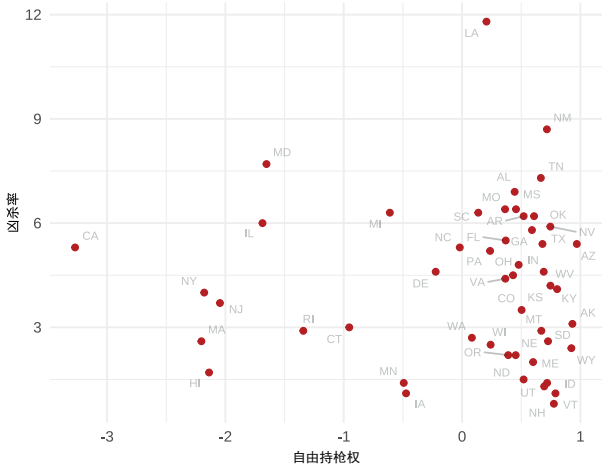
```
ggplot(world, aes(pwthc, fdi)) +
  geom_point(col="#bf0000") +
  geom_text_repel(size=2.4, vjust=1, max.overlaps = Inf,
                  col="#263333", alpha = 0.3,
                  aes(label= iso3c)) +
  ggtitle("问题 7") +
  ylab("境外直接投资（FDI）") +
  xlab("人力资本指数") +
  expand_limits(x=4.2) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))
```



问题 7

8. 否（a），否（b），是（c），洛杉矶（LA）、加利福尼亚（CA）（d）

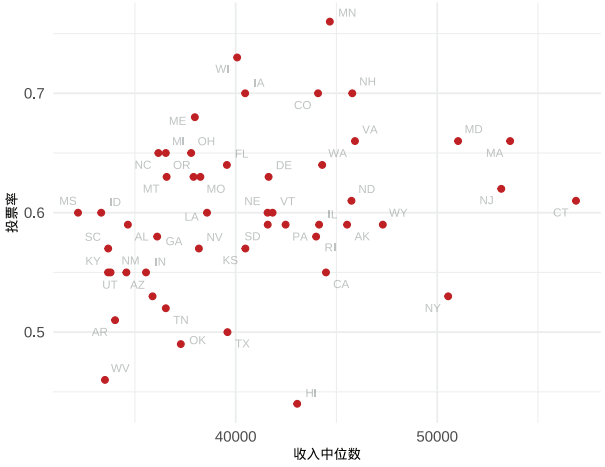
```
ggplot(states, aes(gunfree, murderrate)) +
  geom_point(col="#bf0000") +
  geom_text_repel(size=2.4, vjust=0.5, col="#263333", alpha = 0.3,
                  aes(label= st)) +
  ggtitle(" 问题 8") +
  ylab(" 凶杀率 ") +
  xlab(" 自由持枪权 ") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))
```



问题 8

9. 是（a），是（b），是（c），纽约州（NY）、夏威夷州（HI）（d）

```
ggplot(states, aes(inc, turnout)) +
  geom_point(col="#bf0000") +
  geom_text_repel(size=2.4, vjust=0.5, col="#263333", alpha = 0.3,
                  aes(label= st)) +
  ggtitle(" 问题 9") +
  ylab(" 投票率 ") +
  xlab(" 收入中位数 ") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))
```



问题 9

10. a, b, c, d

11 受控比较

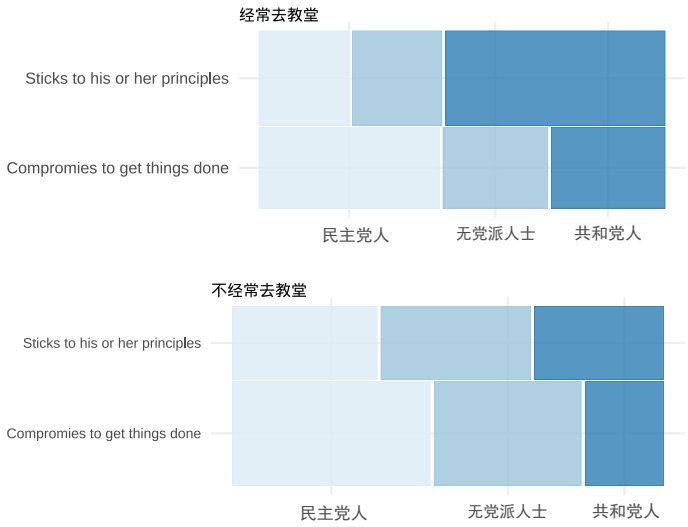
参考答案

知识检验

1. b, c, d

```
p1 <- ggplot(data = subset(nes, nes$pid3.new!="NA" &
                           nes$pew_churatd=="More than once a
                           week")) +
  geom_mosaic(aes(x = product(pid3.new, compromise),
                  fill=pid3.new,na.rm=TRUE)) +
  guides(fill=guide_legend(title=NULL)) +
  ggtitle(" 经常去教堂 ") +
  xlab("") +
  ylab("") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face="bold"),
        axis.title = element_text(size = 8, face="bold")) +
  scale_fill_brewer(palette="Blues") +
  theme(axis.text.x = element_text(size=8, vjust=0.5),
        axis.text.y = element_text(size=8, vjust=0.5),
        legend.position = "none") +
```

```
  coord_flip()

p2 <- ggplot(data = subset(nes, nes$pid3.new!="NA" &
                           nes$pew_churatd!="More than once a
                           week")) +
  geom_mosaic(aes(x = product(pid3.new, compromise),
                  fill=pid3.new,na.rm=TRUE)) +
  guides(fill=guide_legend(title=NULL)) +
  ggtitle("不经常去教堂") +
  xlab("") +
  ylab("") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face="bold"),
        axis.title = element_text(size = 8, face="bold")) +
  scale_fill_brewer(palette="Blues") +
  theme(axis.text.x = element_text(size=7, vjust=0.5),
        axis.text.y = element_text(size=7, vjust=0.5),
        legend.position="none") +
  coord_flip()
grid.arrange(p1, p2, nrow=2,top=textGrob("教会影响了共和党人",
             gp=gpar(fontsize=8)))
```

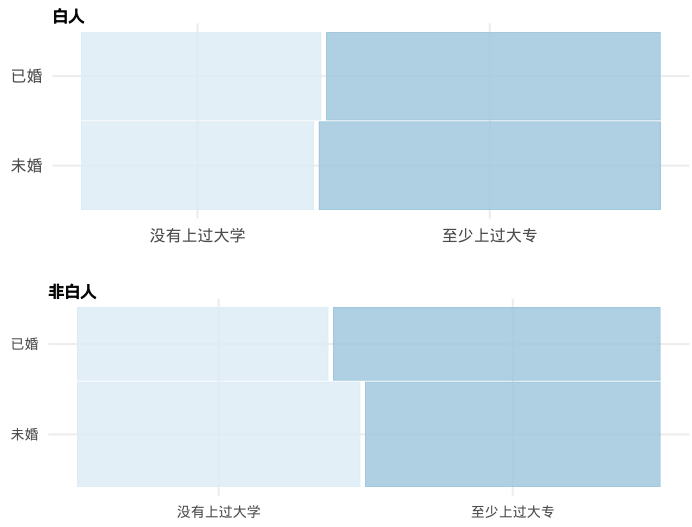


教会影响了共和党人

2. 地区（a），特朗普在 2016 年赢得的州（trumpwin）（b），民主党人（政党认同）（c），人口（d）

```
nes$hsgrad <- ifelse(nes$educ=="No HS", 0,
                     ifelse(nes$educ=="High school graduate", 0, 1))
nes$hsgrad.f <- as.factor(nes$hsgrad)

levels(nes$hsgrad.f)=c("没有上过大学", "至少上过大专")
```

```
nes$White <- ifelse(nes$race=="White", 1, 0)
nes$White.f <- as.factor(nes$White)

levels(nes$White.f)=c("非白人", "白人")

nes$married <- ifelse(nes$marstat=="Married", 1, 0)
nes$married.f <- as.factor(nes$married)
levels(nes$married.f)=c("未婚", "已婚")

p1 <- ggplot(data = subset(nes, nes$White.f=="白人")) +
  geom_mosaic(aes(x = product(hsgrad.f, married.f),
                  fill=hsgrad.f,na.rm=TRUE)) +
  guides(fill=guide_legend(title=NULL)) +
  ggtitle("白人") +
  xlab("") +
  ylab("") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face="bold"),
        axis.title = element_text(size = 8, face="bold")) +
  scale_fill_brewer(palette="Blues") +
  theme(axis.text.x = element_text(size=8, vjust=0.5),
        axis.text.y = element_text(size=8, vjust=0.5),
        legend.position = "none") +
  coord_flip()

p2 <- ggplot(data = subset(nes, nes$White.f!="白人")) +
  geom_mosaic(aes(x = product(hsgrad.f, married.f),
                  fill=hsgrad.f,na.rm=TRUE)) +
  guides(fill=guide_legend(title=NULL)) +
  ggtitle("非白人") +
  xlab("") +
  ylab("") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face="bold"),
        axis.title = element_text(size = 8, face="bold")) +
  scale_fill_brewer(palette="Blues") +
  theme(axis.text.x = element_text(size=7, vjust=0.5),
        axis.text.y = element_text(size=7, vjust=0.5),
        legend.position="none") +
  coord_flip()
grid.arrange(p1, p2, nrow=2, top=textGrob("种族不会影响婚姻和大学之间
的关系", gp=gpar(fontsize=8)))
```



种族不会影响婚姻和大学之间的关系

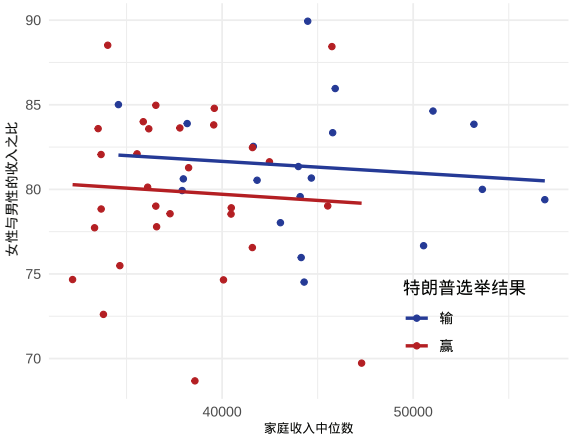
3. a, b, c, d

4. 是（a），是（b），对共和党人有影响（c），自变量（d）

5. 否（a），否（b），都没有影响（c），自变量（d）

6. 否（a），否（b），否（c），贫穷（d）

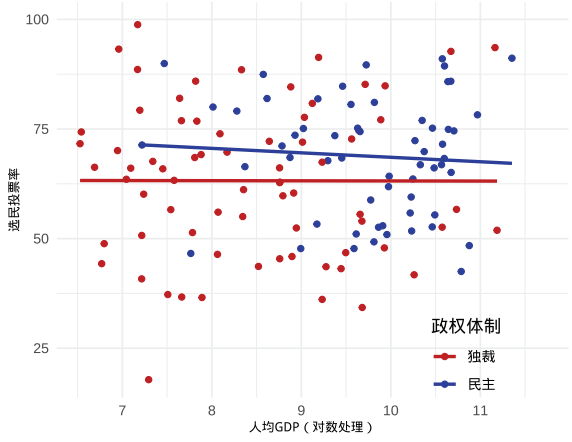
```
states$trumpwin.f <- as.factor(states$trumpwin)
levels(states$trumpwin.f)=c("输", "赢")
ggplot(states, aes(inc, percwom, col=trumpwin.f)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))  +
  theme(legend.position = c(0.8, 0.2)) +
  scale_color_manual(breaks = c("输", "赢"),
                     values = c("#0000bf", "#bf0000")) +
  guides(col=guide_legend("特朗普选举结果")) +
  xlab("家庭收入中位数") +
  ylab("女性与男性的收入之比") +
  ggtitle("这是经济问题")
```



这是经济问题

7. 否（a），否（b），是（c），独裁（dictatorship）（d）

```
ggplot(subset(world, democ!="NA"),
       aes(log(gdppc), turnout,col=democ)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))  +
  theme(legend.position = c(0.8, 0.1)) +
  scale_color_manual(values=c("#bf0000", "#0000bf")) +
  guides(col=guide_legend("政权体制")) +
  xlab("人均GDP（对数处理）") +
  ylab("选民投票率") +
  ggtitle("这是经济问题")
```



这是经济问题

8. b, c

9. a, b, c

数据分析与可视化练习

1. a, b, c, d

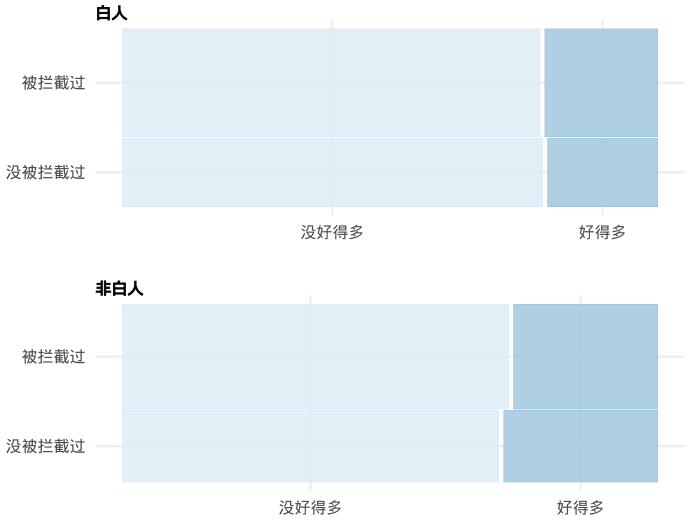
2. a, b, c

3. 赢得比赛（a），赢得比赛（b），赚到很多钱（c），赢得比赛（d）

4. 否（a），种族（b），看法（c），否（d）

```
nes$newpolice <-ifelse(nes$disc_police=="Treats whites much better", 1,
                       ifelse(nes$disc_police=="Skipped", NA, 0))
nes$newpolice <- factor(nes$newpolice, labels=c("没好得多", "好得多"))

nes$newstop <- ifelse(nes$stop_ever=="Has happened", 1,
                      ifelse(nes$stop_ever=="Not asked", NA,
                             ifelse(nes$stop_ever=="Skipped", NA, 0)))
nes$newstop <-factor(nes$newstop,
                     labels=c("没被拦截过", "被拦截过"))

p1 <- ggplot(data = subset(nes, nes$newstop!="NA" &
                             nes$White.f=="白人")) +
  geom_mosaic(aes(x = product(newpolice, newstop),
                  fill=newpolice,na.rm=TRUE)) +
  guides(fill=guide_legend(title=NULL)) +
  ggtitle("白人") +
  xlab(label=NULL) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face="bold"),
        axis.title = element_text(size = 8, face="bold")) +
  scale_fill_brewer(palette="Blues") +
  theme(axis.text.x = element_text(size=8, vjust=0.5),
        axis.text.y = element_text(size=8, vjust=0.5),
        legend.position = "none") +
  xlab("") +
  ylab("") +
  coord_flip()

p2 <- ggplot(data = subset(nes, nes$newstop!="NA" & nes$newpolice!="NA"
                           & nes$White!="白人")) +
  geom_mosaic(aes(x = product(newpolice, newstop),
                  fill=newpolice, na.rm=TRUE)) +
  guides(fill=guide_legend(title=NULL)) +
  ggtitle("非白人") +
```

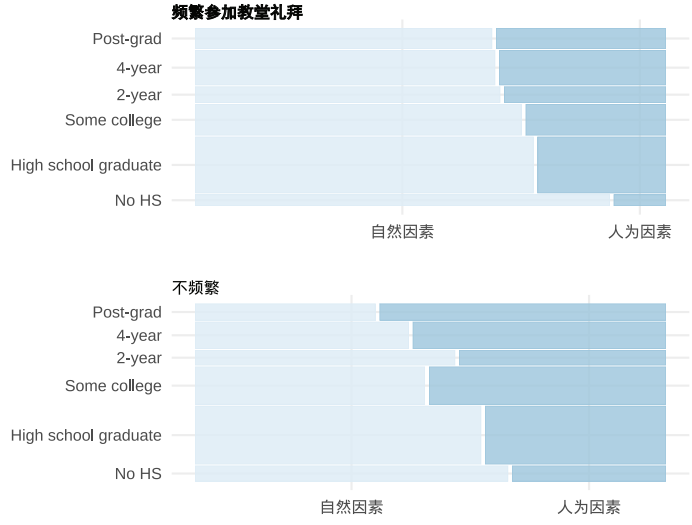
```
  xlab(label=NULL) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face="bold"),
        axis.title = element_text(size = 8, face="bold")) +
  scale_fill_brewer(palette="Blues") +
  theme(axis.text.x = element_text(size=8, vjust=0.5),
        axis.text.y = element_text(size=8, vjust=0.5),
        legend.position = "none") +
  xlab("") +
  ylab("") +
  coord_flip()

grid.arrange(p1, p2, nrow=2, top=textGrob("", gp=gpar(fontsize=8)))
```



5. 是（a），教堂礼拜参加（b），对气候变化的看法（c），是（d）

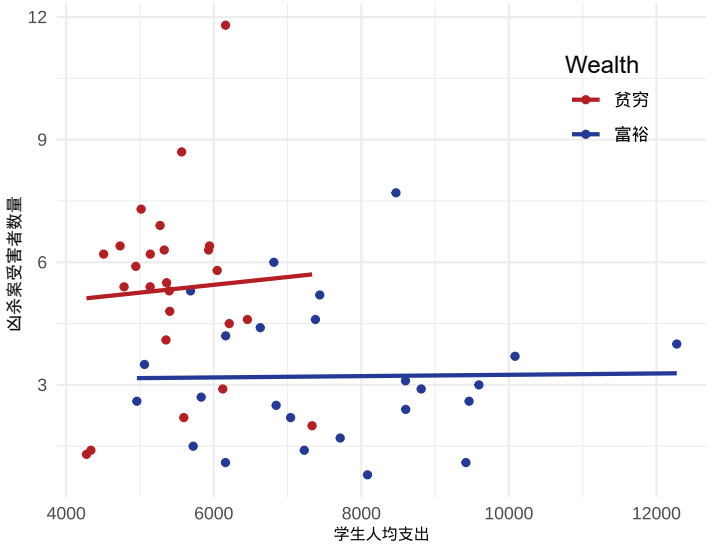
```
nes$newcause <- ifelse(nes$warmcause=="Mostly by human activity", 1, 0)
nes$newcause <- factor(nes$newcause, labels=c("自然因素", "人为因素"))
nes$newchurch <- ifelse(nes$pew_churatd=="More than once a week", 1,
                        ifelse(nes$pew_churatd=="Once a week", 1,
                               ifelse(nes$pew_churatd=="Skipped", NA, 0)))
nes$newchurch <- factor(nes$newchurch,
                        labels=c("每周不超过一次",
                                 "每周一次以上"))

p1 <-ggplot(data=subset(nes, nes$newchurch=="每周一次以上")) +
  geom_mosaic(aes(x = product(newcause, educ),
                  fill=newcause,na.rm=TRUE)) +
  guides(fill=guide_legend(title=NULL)) +
  ggtitle("频繁参加教堂礼拜") +
  xlab(label=NULL) +
```

```
    theme_minimal() +
    theme(plot.title = element_text(size = 8, face="bold"),
          axis.title = element_text(size = 8, face="bold")) +
    scale_fill_brewer(palette="Blues") +
    theme(axis.text.x = element_text(size=8, vjust=0.5),
          axis.text.y = element_text(size=8, vjust=0.5),
          legend.position = "none") +
    xlab("") +
    ylab("") +
    coord_flip()

p2 <-ggplot(data = subset(nes, nes$newcause!="NA" &
                          nes$newchurch!=" 每周一次以上 ")) +
    geom_mosaic(aes(x = product(newcause, educ),
                    fill=newcause, na.rm=TRUE)) +
    guides(fill=guide_legend(title=NULL)) +
    ggtitle(" 不频繁 ") +
    xlab(label=NULL) +
    theme_minimal() +
    theme(plot.title = element_text(size = 8, face="bold"),
          axis.title = element_text(size = 8, face="bold")) +
    scale_fill_brewer(palette="Blues") +
    theme(axis.text.x = element_text(size=8, vjust=0.5),
          axis.text.y = element_text(size=8, vjust=0.5),
          legend.position = "none") +
    xlab("") +
    ylab("") +
    coord_flip()
grid.arrange(p1, p2, nrow=2, top=textGrob("", gp=gpar(fontsize=8)))
```



6. 否（a），富裕（b），凶杀率（c），是（d）

```
states$rich <- cut(states$inc, breaks=c(0,40000,60000))

states$rich <- factor(states$rich, labels = c("贫穷", "富裕"))

ggplot(states, aes(stuspend, murderrate, col=rich)) +
  geom_point() +
  ggtitle("凶杀率、收入和教育支出") +
  ylab("凶杀率") +
  xlab("学生人均支出") +
  geom_smooth(method="lm", se=FALSE) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(legend.position = c(0.85, 0.8)) +
  guides(col=guide_legend("Wealth")) +
  scale_color_manual(breaks = c("贫穷", "富裕"),
                     values=c("#bf0000", "#0000bf"))
```

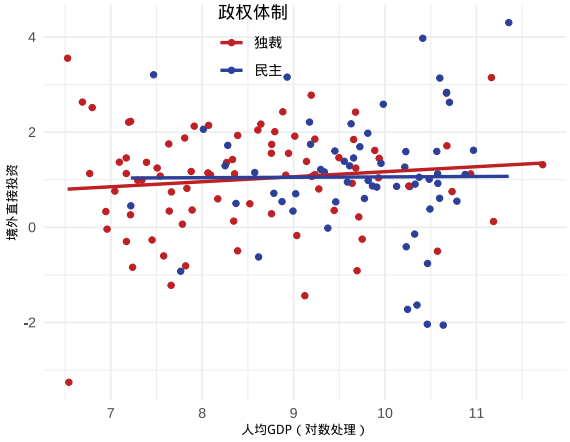


凶杀率、收入和教育支出

7. 否（a），政权体制（b），境外直接投资（FDI）（c），否（d）

```
ggplot(subset(world, democ!="NA"), aes(log(gdppc), log(fdi), col=democ)) +
  geom_point() +
  ggtitle("收入、境外直接投资和民主") +
  ylab("境外直接投资") +
  xlab("人均 GDP（对数处理）") +
  geom_smooth(method="lm", se=FALSE) +
```

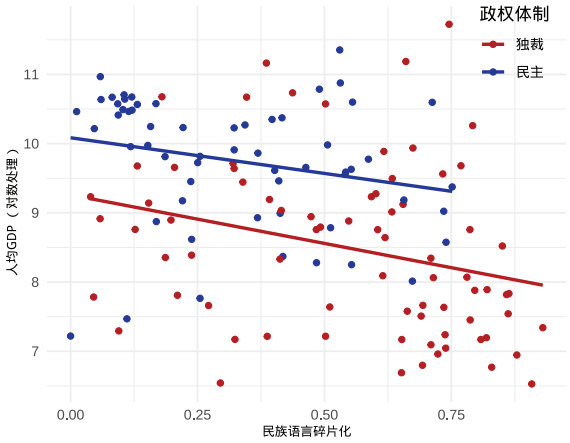
```
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(legend.position = c(0.4, 0.9)) +
  guides(col=guide_legend("政权体制")) +
  scale_color_manual(breaks = c("独裁", "民主"),
                     values=c("#bf0000", "#0000bf"))
```



收入、境外直接投资和民主

8. 否（a），政权体制（b），人均 GDP（c），是（d）

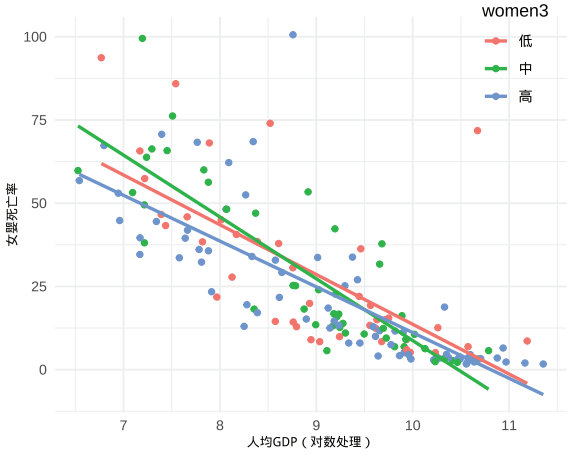
```
ggplot(subset(world, democ!="NA"), aes(ethfrac, log(gdppc), col=democ)) +
  geom_point() +
  ggtitle("种族、收入和民主") +
  ylab("人均 GDP（对数处理）") +
  xlab("民族语言碎片化") +
  geom_smooth(method="lm", se=FALSE) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(legend.position = c(0.9, 0.9)) +
  guides(col=guide_legend("政权体制")) +
  scale_color_manual(breaks = c("独裁", "民主"),
                     values=c("#bf0000", "#0000bf"))
```



种族、收入和民主

9. 否（a），立法机构中的女性百分比（b），女婴死亡率（c），是（d）

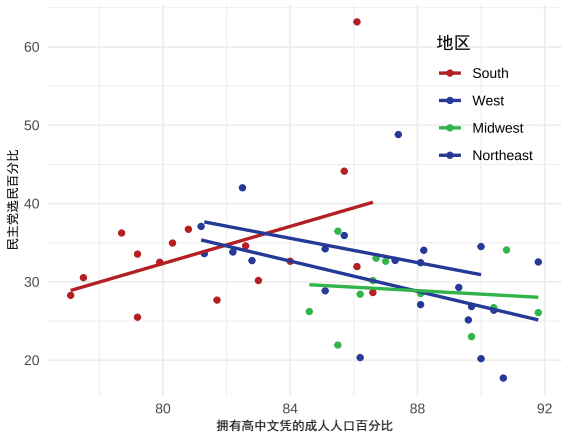
```
world$women3 <- cut(world$womleg, breaks=c(0,10,18,57))

world$women3 <- factor(world$women3, labels = c("低", "中", "高"))

ggplot(subset(world, women3!="NA"), aes(log(gdppc), infemale, col=women3)) +
  geom_point() +
  ggtitle("立法机构中的女性占比与婴儿健康") +
  ylab("女婴死亡率") +
  xlab("人均 GDP（对数处理）") +
  geom_smooth(method="lm", se=FALSE) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(legend.position = c(0.9, 0.9))
```



立法机构中的女性占比与婴儿健康

10. 是（a），地区（b），投票（c），南部（d）

```
ggplot(states, aes(hsdiploma, democrat, col=region)) +
  geom_point() +
  ggtitle("教育与投票") +
  ylab("民主党选民百分比") +
  xlab("拥有高中文凭的成人人口百分比") +
  geom_smooth(method="lm", se=FALSE) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  theme(legend.position = c(0.85, 0.75)) +
  guides(col=guide_legend("地区")) +
  scale_color_manual (values=c("#bf0000", "#0000bf", "#00bf00", "#0000bf"))
```



教育与投票

12 线性回归

参考答案

知识检验

1. a, c, d
2. a, b
3. b, d
4. $\hat{Y}$（a），α（b），β_i（c），x_i（d）

5. d
6. a, b, c
7. 增加（a），不变（b），增加（c），增加（d）
8. a
9. a
10. c, d
11. a, b
12. a
13. b
14. d
15. d

数据分析与可视化练习

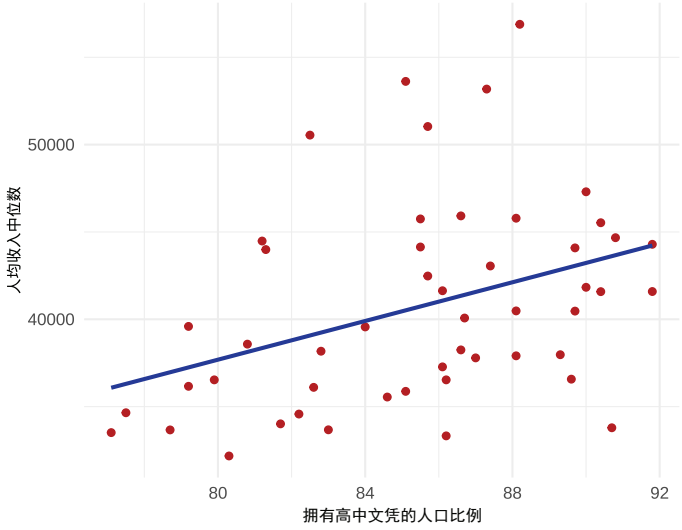
1. b, c
2. a, b, c
3. b, d
4. b
5. b, d
6. d
7. 拥有高中文凭的人口规模每增加 1 个百分点，相应的人均收入就会增加 554 美元（a）；散点图（b）；0.14（c）；2.81（d）

```
income.lm <- lm(inc ~ hsdiploma, data = states)

stargazer(income.lm, type = "text", title =" 教育带来收入增长 ",
          header = FALSE)

ggplot(states, aes(hsdiploma, inc)) +
    geom_point(col = "#bf0000") +
    ggtitle(" 美国的收入和教育 ") +
    geom_smooth(method = "lm", se = F, fullrange = F,
    col = "#0000bf") +
```

```
        theme_minimal() +
        theme(plot.title = element_text(size = 8, face = "bold"),
              axis.title = element_text(size = 8, face = "bold")) +
        ylab("人均收入中位数") +
        xlab("拥有高中文凭的人口比例")
```

教育带来收入增长

```
================================================
                          Dependent variable:
                      ---------------------------
                                  inc
------------------------------------------------
hsdiploma                     553.929***
                               (196.954)

Constant                      -6,626.114
                             (16,853.690)

------------------------------------------------
Observations                      50
R2                               0.141
Adjusted R2                      0.124
Residual Std. Error      5,459.255 (df = 48)
F Statistic            7.910*** (df = 1; 48)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```



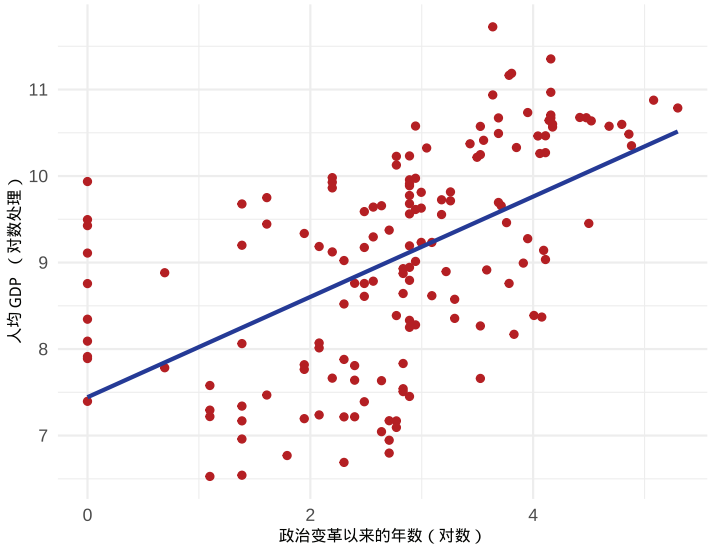
美国的收入和教育

8. 收入每变化 1%，相应的稳定年数变化 0.52%（a）；散点图（b）；0.30（c）；8.125（d）

```
develop.lm <- lm(log(durable + 1) ~ log(gdppc), data = world)

stargazer(develop.lm, type = "text", title = "政治稳定与收入",
          header = FALSE)

ggplot(world, aes(log(durable + 1), log(gdppc))) +
    geom_point(col = "#bf0000") +
    ggtitle("政治稳定与收入") +
    geom_smooth(method = "lm", se = F, fullrange = F,
                col = "#0000bf") +
    theme_minimal() +
    theme(plot.title = element_text(size = 8, face = "bold"),
          axis.title = element_text(size = 8, face = "bold")) +
    ylab("人均 GDP（对数处理）") +
    xlab("政治变革以来的年数（对数）")
```

政治稳定与收入

```
================================================
                        Dependent variable:
                    ---------------------------
                          log(durable + 1)
------------------------------------------------
log(gdppc)                    0.520***
                               (0.064)

Constant                      -1.929***
                               (0.587)

------------------------------------------------
Observations                     154
R2                              0.301
Adjusted R2                     0.297
Residual Std. Error       0.990 (df = 152)
F Statistic           65.527*** (df = 1; 152)
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```



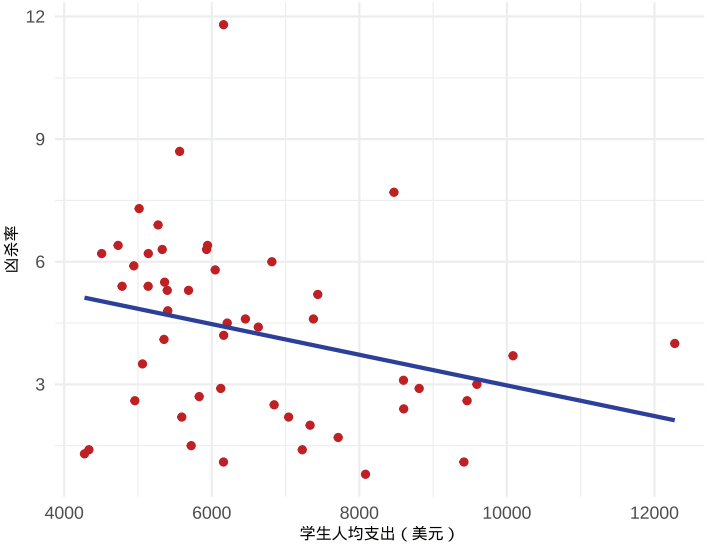
政治稳定与收入

9. 学生人均支出每增加 1 美元，相应的每 10 万人口中凶杀案受害者数量就会减少 0.0004（a）；散点图（b）；0.08（c）；2.04（d）

```
murder.lm <- lm(murderrate ~ stuspend, data = states)

stargazer(murder.lm, type = "text", title = "花在学生上的支出，在警
务支出上省了回来", header = FALSE)

ggplot(states, aes(stuspend, murderrate)) +
    geom_point(col = "#bf0000") +
    ggtitle("美国的凶杀率与教育") +
    geom_smooth(method = "lm", se = F, fullrange = F,
                col = "#0000bf") +
    theme_minimal() +
    theme(plot.title = element_text(size = 8, face = "bold"),
          axis.title = element_text(size = 8, face = "bold")) +
    ylab("凶杀率") +
    xlab("学生人均支出（美元）")
```

花在学生上的支出，在警务支出上省了回来

```
================================================
                          Dependent variable:
                      ---------------------------
                                murderrate
------------------------------------------------
stuspend                        -0.0004**
                                (0.0002)

Constant                        6.724***
                                 (1.238)

------------------------------------------------
Observations                       50
R2                                0.080
Adjusted R2                       0.061
Residual Std. Error          2.200 (df = 48)
F Statistic                4.174** (df = 1; 48)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```



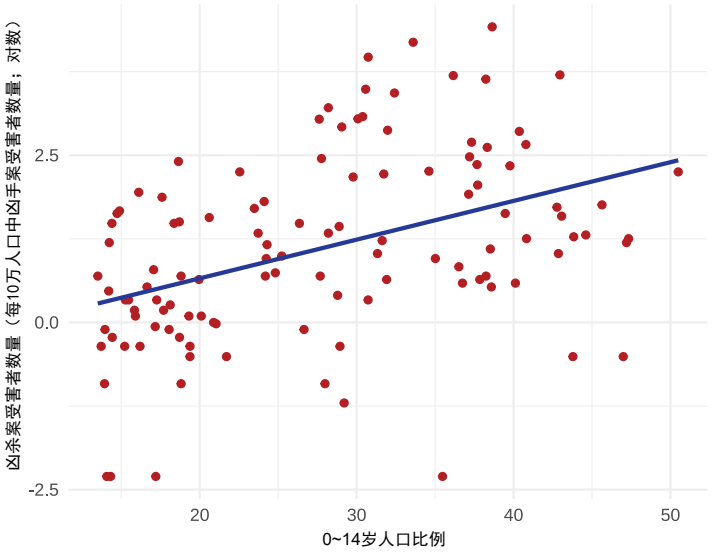
美国的凶杀率与教育

10. 0 ~ 14 岁人口每增加 1 个百分点，相应的凶杀案受害者数量就会增加 6%（a）；散点图（b）；0.18（c）；4.9（d）

```
youth.lm <- lm(log(homicide) ~ young, data = world)

stargazer(youth.lm, type = "text", title = "青年与凶杀案受害者数量",
          header = FALSE)

ggplot(world, aes(young, log(homicide))) +
    geom_point(col = "#bf0000") +
    ggtitle("青年谋杀") +
    geom_smooth(method = "lm", se = F, fullrange = F,
                col = "#0000bf") +
    theme_minimal() +
    theme(plot.title = element_text(size = 8, face = "bold"),
          axis.title = element_text(size = 8, face = "bold")) +
    ylab("凶杀案受害者数量（每 10 万人口中凶杀案受害者数量；对数）") +
    xlab("0 ~ 14 岁人口比例")
```

青年与凶杀案受害者数量

```
================================================
                        Dependent variable:
                    ---------------------------
                           log(homicide)
------------------------------------------------
young                        0.058***
                             (0.012)

Constant                     -0.499
                             (0.349)

------------------------------------------------
Observations                   114
R2                            0.177
Adjusted R2                   0.170
Residual Std. Error      1.279 (df = 112)
F Statistic           24.125*** (df = 1; 112)
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```



青年谋杀

13 多元回归

参考答案

知识检验

1. a, b, c, d
2. a, b, c, d
3. a, b, c, d
4. a, c
5. a, d
6. a, b
7. 教育工会的说客（a），经济学家（b），社会学家（c），政策制定者（d）
8. d

9. c, d
10. a, b, c
11. a, b, d
12. b, d
13. a, c, d
14. b
15. a, c, d

数据分析与可视化练习

1. b
2. b
3. 在人均 GDP 不变的情况下，投票率每增加 1 个百分点，对应的每千名活产婴儿死亡数减少 0.112（a）；在投票率不变的情况下，收入翻倍，对应的每千名活产婴儿死亡数减少 11.8（b）；–1.339（c）；否（d）

```
summary(lm(inf ~ turnout + log2(gdppc), data=world))

Call:
lm(formula = inf ~ turnout + log2(gdppc), data = world)

Residuals:
    Min      1Q  Median      3Q     Max
-26.887  -9.988  -2.745   4.712  80.095

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 189.84427   11.02530  17.219   <2e-16 ***
turnout      -0.10620    0.08423  -1.261    0.209
log2(gdppc) -11.76875    0.76794 -15.325   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.31 on 147 degrees of freedom
  (32 observations deleted due to missingness)
Multiple R-squared:  0.6231,    Adjusted R-squared:  0.618
F-statistic: 121.5 on 2 and 147 DF,  p-value: < 2.2e-16
```

4. 在收入和婴儿死亡率保持不变的情况下，民主党的州人口比例每增加 1 个百分点，相应的拥有高中文凭的人口比例就会下降 0.99 个百分点（a）；在保持

政党认同和婴儿死亡率不变的情况下，人均收入每增加 1 美元，相应的拥有高中文凭的人口比例就会增加 0.0002 个百分点（b）；1.5（c）；否（d）

```
summary(lm(hsdiploma ~ inc + infant + democrat, data=states))

Call:
lm(formula = hsdiploma ~ inc + infant + democrat, data = states)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1201 -2.6018  0.4804  2.5909  5.6360

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.9641683  5.8437041  15.224   <2e-16 ***
inc          0.0001588  0.0001055   1.505   0.1391
infant      -0.9939696  0.4315306  -2.303   0.0258 *
democrat    -0.0949175  0.0743142  -1.277   0.2079
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.385 on 46 degrees of freedom
Multiple R-squared:  0.3142,    Adjusted R-squared:  0.2694
F-statistic: 7.023 on 3 and 46 DF,  p-value: 0.0005504
```

5. 在保持人均 GDP 不变的情况下，女性获得选举权的时间每推迟 1 年，女孩与男孩的受教育程度比例就会增加 0.00026 个百分点（a）;0.641（b）;否（c）;14.5%（d）

```
summary(lm(gtbeduc ~ womyear + log(gdppc), data=world))

Call:
lm(formula = gtbeduc ~ womyear + log(gdppc), data = world)

Residuals:
      Min        1Q    Median        3Q       Max
-0.279973 -0.021841  0.004725  0.034231  0.134698

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2383012  0.8275275   0.288 0.773962
womyear     0.0002609  0.0004055   0.643 0.521361
log(gdppc)  0.0255680  0.0066875   3.823 0.000228 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06162 on 101 degrees of freedom
  (78 observations deleted due to missingness)
```

```
Multiple R-squared:  0.1619,    Adjusted R-squared:  0.1453
F-statistic: 9.758 on 2 and 101 DF,  p-value: 0.0001335
```

6. 在保持人均收入不变的情况下，女性获得选举权的时间每推迟 1 年，每千名活产婴儿的女婴与男婴的死亡比例就会增加 0.06 个百分点（a）；3.18（b）；是（c）；5%（d）

```
summary(lm((infemale/infmale) * 100 ~ womyear + log(gdppc),
           data = world))

Call:
lm(formula = (infemale/infmale) * 100 ~ womyear + log(gdppc),
   data = world)

Residuals:
    Min      1Q  Median      3Q     Max
-14.593  -2.255  -0.034   2.311  20.786

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.20388   39.74096  -1.062  0.29008
womyear       0.06203    0.01965   3.157  0.00195 **
log(gdppc)    0.44781    0.31416   1.425  0.15627
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.903 on 140 degrees of freedom
  (39 observations deleted due to missingness)
Multiple R-squared:  0.06671,   Adjusted R-squared:  0.05337
F-statistic: 5.003 on 2 and 140 DF,  p-value: 0.007967
```

7. 在人口和学生教育支出保持不变的情况下，该州拥有高中文凭的人口比例每增加 1 个百分点，相应的人均收入就会增加 443 美元（a）；在保持人口、学生教育支出和政党认同不变的情况下，该州拥有高中文凭的人口比例每增加 1 个百分点，相应的人均收入就会增加 446 美元（b）；是（c）

```
wout.lm <- lm(inc ~ hsdiploma + poptotal + stuspend,
              data = states)

with.lm <- lm(inc ~ hsdiploma + poptotal + stuspend +
              democrat, data = states)

stargazer(wout.lm, with.lm, header=FALSE, type="text",
          title ="政党认同相对无关紧要")
```

政党认同相对无关紧要

```
================================================================
                                Dependent variable:
                    ---------------------------------------------
                                        inc
                                (1)                  (2)
----------------------------------------------------------------
hsdiploma                    442.701***           446.097***
                             (129.234)            (136.083)

poptotal                     0.0002***            0.0002***
                              (0.0001)             (0.0001)

stuspend                      2.428***             2.417***
                              (0.282)              (0.311)

democrat                                           6.444
                                                  (72.236)

Constant                   -14,200.750          -14,619.970
                           (10,941.430)         (12,018.260)

----------------------------------------------------------------
Observations                    50                   50
R2                             0.701                0.701
Adjusted R2                    0.682                0.675
Residual Std. Error  3,288.938 (df = 46)    3,324.987 (df = 45)
F Statistic         36.015*** (df = 3; 46) 26.431*** (df = 4; 45)
================================================================
Note:                                *p<0.1; **p<0.05; ***p<0.01
```

8. 在保持人均 GDP 不变的情况下，0 ~ 14 岁的人口比例每增加 1 个百分点，CO_2 排放量就会下降 1.29%（a）；否（b）；在青年人口保持不变的情况下，人均 GDP 每变化 1%，相应的碳排放量就会下降 0.08%（c）；否（d）

```
summary(lm(log(co2) ~ young + log(gdppc), data=world ))

Call:
lm(formula = log(co2) ~ young + log(gdppc), data = world)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9582 -0.4761 -0.1397  0.4456  2.1894

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.117069   0.956081   0.122    0.903
young       -0.012545   0.009248  -1.356    0.177
```

```
log(gdppc)   -0.078424    0.078757  -0.996     0.321

Residual standard error: 0.7238 on 141 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared:  0.01301,   Adjusted R-squared:  -0.0009911
F-statistic: 0.9292 on 2 and 141 DF,  p-value: 0.3973
```

9. 在收入和婴儿死亡率保持不变的情况下，民主党的州人口比例每增加 1 个百分点，相应的拥有高中文凭的人口比例就会减少 0.05 个百分点（a）；人均收入每增加 1 美元，相应的拥有高中文凭的人口比例就会增加 0.0002 个百分点（b）；1.83（c）；是（d）。注意：stargazer 对数字进行了四舍五入。更精确的 *t*- 比率和显著性估计值可以使用 `summary()` 函数得到。

```
without.lm <- lm(hsdiploma ~ inc + infant + democrat,
    data = subset(states, st!="NY" & st!="CA" & st!="NJ"))

withall.lm <- lm(hsdiploma ~ inc + infant + democrat,
                 data = states)

stargazer(without.lm, withall.lm, header=FALSE, type="text",
          title = "缺失的个案改变了结果")
```

缺失的个案改变了结果

```
=================================================================
                              Dependent variable:
                    ---------------------------------------------
                                     hsdiploma
                              (1)                    (2)
-----------------------------------------------------------------
inc                         0.0002*                0.0002
                           (0.0001)               (0.0001)

infant                     -1.223***              -0.994**
                            (0.422)                (0.432)

democrat                    -0.051                 -0.095
                            (0.073)                (0.074)

Constant                   88.094***              88.964***
                            (5.715)                (5.844)

-----------------------------------------------------------------
Observations                  47                     50
R2                           0.390                  0.314
Adjusted R2                  0.347                  0.269
Residual Std. Error     3.234 (df = 43)        3.385 (df = 46)
F Statistic         9.156*** (df = 3; 43) 7.023*** (df = 3; 46)
=================================================================
Note:                                *p<0.1; **p<0.05; ***p<0.01
```

10. 在保持人均 GDP 不变的情况下，女性获得选举权的时间每推迟 1 年，女孩与男孩的受教育程度比例就会增加 0.04 个百分点（a）；0.91（b）；否（c）；否（d）

```
woutME.lm <- lm(gtbeduc*100 ~ womyear + log(gdppc),
    data=subset(world, region!="ME and North Africa"))

withME.lm <- lm(gtbeduc*100 ~ womyear + log(gdppc), data=world)

stargazer(woutME.lm, withME.lm, header=FALSE, type = "text",
          title = "是否纳入中东地区的个案差别不大")
```

是否纳入中东地区的个案差别不大

```
=================================================================
                                Dependent variable:
                    ---------------------------------------------
                                    gtbeduc * 100
                              (1)                    (2)
-----------------------------------------------------------------
womyear                     0.039                  0.026
                           (0.042)                (0.041)

log(gdppc)                 2.816***               2.557***
                           (0.686)                (0.669)

Constant                   -3.995                 23.830
                           (86.726)               (82.753)

-----------------------------------------------------------------
Observations                 97                     104
R2                          0.191                  0.162
Adjusted R2                 0.174                  0.145
Residual Std. Error    6.058 (df = 94)        6.162 (df = 101)
F Statistic         11.099*** (df = 2; 94) 9.758*** (df = 2; 101)
=================================================================
Note:                                *p<0.1; **p<0.05; ***p<0.01
```

14 虚拟变量和交互作用

参考答案

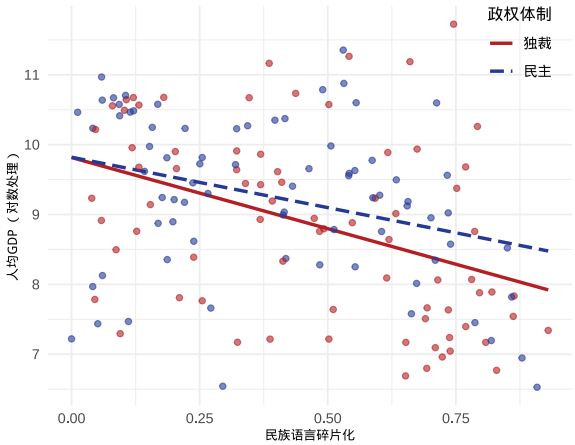
知识检验

1. a, c, d
2. a, c, d
3. a, c, d
4. b
5. c, d
6. a, b
7. a, b, c, d
8. 天主教徒（a）；接受过大学教育的共和党人（b），不是共和党人，或没有接受过大学教育（c）；没有接受过大学教育，也没有工作的公民（d）
9. c
10. d
11. b, d

数据分析与可视化练习

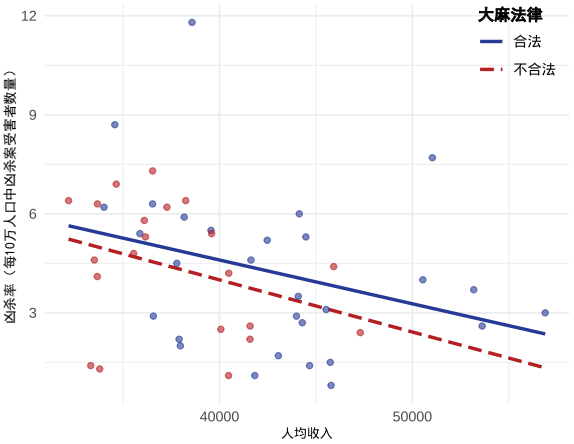
1. b, c
2. 加性模型（a），交互作用模型（b），交互作用模型（c），加性模型（d）
3. c, d
4. a
5. c
6. a
7. 9.818（a），9.819（b），−2.01（c），−1.44（d）
8. 10.31（a），0.6（b），−0.0001579（c），−0.000132（d）

```
9. mod1.lm <- lm(lngdppc ~ ethfrac + dem + ethfrac:dem, data=world)

   interact_plot(mod1.lm, pred = ethfrac, modx = dem, data = world,
                 legend.main = "政权体制",
                 modx.labels = c("独裁", "民主"),
                 plot.points = TRUE,
                 colors = c("#bf0000", "#0000bf")) +
     xlab("民族语言碎片化") +
     ylab("人均GDP（对数处理）") +
     ggtitle("民主与民族语言碎片化之间的交互作用") +
     theme_minimal() +
     theme(plot.title = element_text(size = 8, face = "bold"),
           axis.title = element_text(size = 8, face = "bold"),
           legend.title = element_text(size = 10),
           legend.position = c(0.9,0.9))
```



民主与民族语言碎片化之间的交互作用

```
10. mod2.lm <- lm(murderrate ~ weed + inc + weed:inc, data=states)

    interact_plot(mod2.lm, pred = inc, modx = weed, data =states,
                  legend.main = "大麻法律",
                  modx.labels = c("不合法", "合法"),
                  plot.points = TRUE,
                  colors = c("#bf0000", "#0000bf")) +
      xlab("人均收入") +
      ylab("凶杀率（每10万人口中凶杀案受害者数量）")+
      ggtitle("大麻法律与收入之间的交互作用") +
      theme_minimal() +
      theme(plot.title = element_text(size = 8, face = "bold"),
            axis.title = element_text(size = 8, face = "bold"),
            legend.title = element_text(size = 10),
            legend.position = c(0.9,0.9))
```



大麻法律与收入之间的交互作用

15 诊断 1：普通最小二乘法是否适用

参考答案

知识检验

1. a, b, c, d
2. 无偏（a），有偏（b），有效（c）
3. a, c
4. b, c
5. d
6. b
7. c
8. a, b, c, d

数据分析与可视化练习

1. b

2. b
3. 有偏（a），有偏（b），无偏（c）
4. 无偏（a），有偏（b），有偏（c）
5. 无偏（a），有偏（b），有效（c）
6. a, b
7. d
8. a
9. c
10. d

16 诊断 2：残差、杠杆值与影响力的度量

参考答案

知识检验

1. a, b, c, d
2. c
3. b, c
4. a, c
5. b, c
6. 回归系数（a），在 x 轴上的距离（b），预测值（c），在 y 轴上的距离（d）
7. c
8. a, d
9. 安哥拉（AGO）（a），印度（IND）和中国（CHN）（b），卡塔尔（QAT）（c），否（d）
10. 卢旺达（RWA）（a），卢旺达（RWA）（b），卢旺达（RWA）和美国（USA）（c），是（d）

数据分析与可视化练习

1. b
2. 特拉华州（DE）（a），新泽西州（NJ）（b），特拉华州（DE）（c），杠杆值（d）
3. young（a），斯里兰卡（LKA）和约旦（JOR）（b），马耳他（MLT）和乍得（TCD）（c），乍得（TCD）（d）
4. 马耳他（MLT）（a），乍得（TCD）（b），马耳他（MLT）（c），urban（d）
5. 马耳他（MLT）（a），杠杆值（b），是（c），残差（d）
6. b
7. b
8. c
9. 阿富汗（AFG）和缅甸（MMR）（a），乍得（TCD）和安哥拉（AGO）（b），乍得（TCD）和安哥拉（AGO）（c），是（d）
10. 犹他州（UT）和特拉华州（DE）（a），特拉华州（DE）、夏威夷州（HI）和弗吉尼亚州（VA）（b），特拉华州（DE）（c），evangel（d）

17 逻辑回归

参考答案

知识检验

1. a, d
2. b, d
3. b
4. c
5. a
6. 每 10 万人口中每增加 1 名受害者，对应的一个州实施死刑的对数发生比就会增加 0.26。

7. $\frac{e^{\hat{Y}_i}}{\left(1+e^{\hat{Y}_i}\right)}$

8. d

9. 介于 70% 和 80% 之间

10. b, c, d

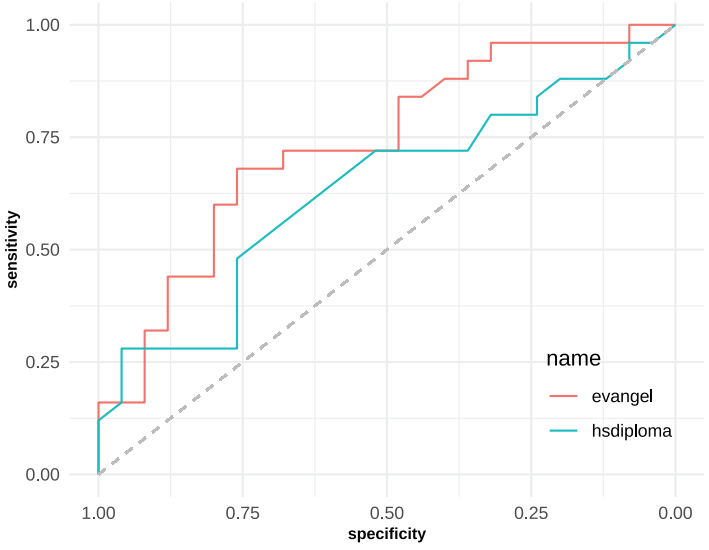
11. a

12. b

13. d

14. evangel（a），没什么帮助（b），是（c）

```
roc.stand <- roc(stand ~ evangel + hsdiploma, data = states)
g.stand <- pROC::ggroc(roc.stand)
g.stand +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"),
        legend.position = c(0.85, 0.2)) +
  geom_segment(aes(x = 1, xend = 0, y = 0, yend = 1),
               color="grey", linetype="dashed") +
  ggtitle("ROC 曲线 ")
```

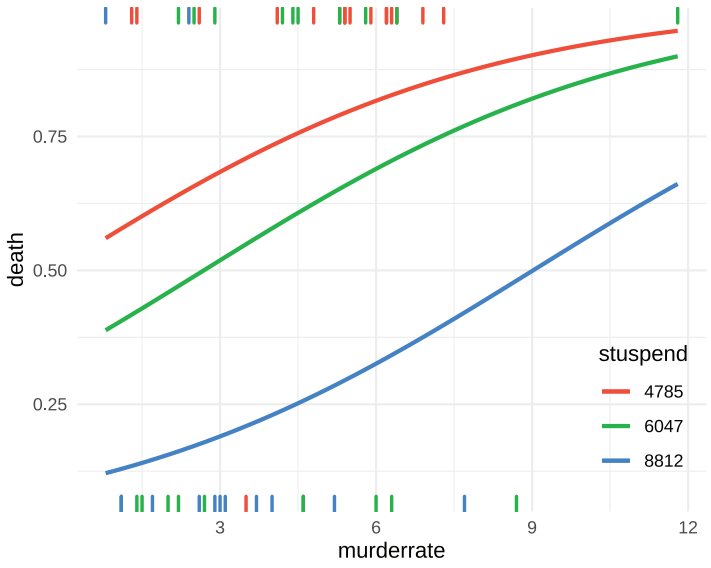


ROC 曲线

数据分析与可视化练习

1. d
2. a, b, c, d
3. c
4. 在每个学生身上每多花 1 美元，一个州有死刑的对数发生比就会降低 0.0005。每 10 万人口中多增加 1 名受害者，一个州有死刑的对数发生比就会增加 0.24 倍。
5. 预测概率

```
death.glm <- glm(death ~ stuspend + murderrate + inc + democrat,
                 data=states, family=binomial)

visreg(death.glm, "murderrate", by="stuspend", overlay=TRUE,
       band=FALSE, scale = "response", gg=TRUE) +
  theme_minimal() +
  theme(legend.position = c(0.9, 0.2))
```



6. 拟 R^2 度量为 Cox-Snell（0.359）、Nagelkerke（0.481）和 McFadden（0.325）。

```
logitR2(death.glm)
```

```
   n      Chi2 df            p       Cox       Nag       RL2
1 50 22.25937  4 0.0001779462 0.3592958 0.4813941 0.3245138
```

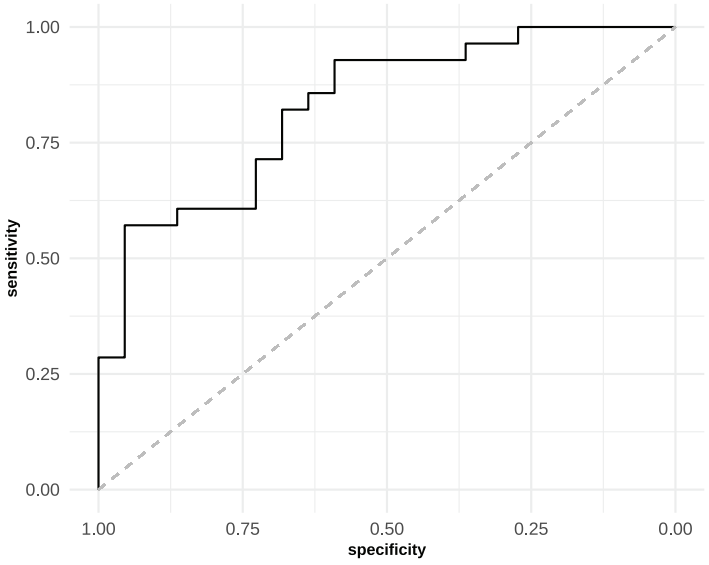
7. 卡方统计量的 p 值为 0.086，表明增加 `democratic` 变量并不会显著改善模型。

```
modelq7.glm  <- glm(death ~ stuspend + murderrate + inc,
                    data=states, family=binomial)
modelq7a.glm <- glm(death ~ stuspend + murderrate + inc + democrat,
                    data=states, family=binomial)

object  <- logitR2(modelq7.glm)
object1 <- logitR2(modelq7a.glm)

pchisqC(modelq7.glm$dev, modelq7a.glm$dev, 1)
```

```
 [1] "0.086357"
```

8. stuspend 变量的 ROC 曲线

```
roc.list <- roc(death ~ stuspend, data = states)
pROC::ggroc(roc.list) +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold")) +
  geom_segment(aes(x = 1, xend = 0, y = 0, yend = 1),
               color="grey", linetype="dashed")
```



stuspend 变量的 ROC 曲线

9. stuspend 变量的 AUC 为 0.8279

```
auc(roc.list)
```

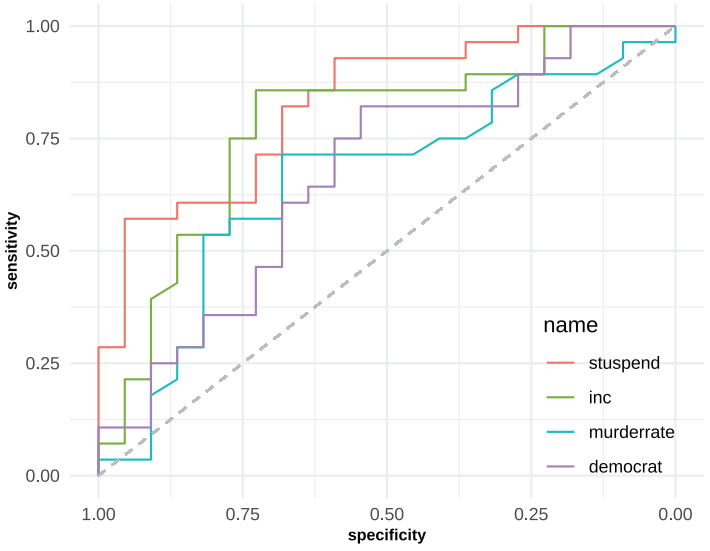
```
Area under the curve: 0.8279
```

10. 模型的 ROC 曲线。stuspend 变量优于其他变量（其 AUC 最大）。

```
aucinc <- roc(death ~ inc, data = states)
auc(aucinc)
```

```
Area under the curve: 0.7752
```

```
roc.full <- roc(death ~ stuspend + inc + murderrate + democrat,
                data = states)
g.full <- pROC::ggroc(roc.full)
g.full +
  theme_minimal() +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"),
        legend.position = c(0.85, 0.2)) +
  geom_segment(aes(x = 1, xend = 0, y = 0, yend = 1),
               color="grey", linetype="dashed") +
  ggtitle("ROC 曲线 ")
```



ROC 曲线

词汇表

第 1 章

R：专为统计计算和图形设计的一种自由（开源）统计编程语言，由 R 统计计算基金会提供支持。

S：20 世纪 60 年代开发的统计程序，R 就是基于此开发的。

包（Package）：可以安装并加载到 R 和 RStudio 中的附加代码，用来帮助我们完成更专业的任务。

R Markdown：一种轻量级标记语言，旨在提供一种能以简便方式呈现专业结果的记录数据分析的方法[1]。做好 Markdown 文件之后，就可以生成 Microsoft Word、HTML 或LaTeX（PDF）格式的文件了。

函数（Function）：在 R 中，对某个对象执行一系列预设的指令。

对象（Object）：在 R 中，可以代表多种不同的东西。对象是函数操作的东西，它可以是数据、图形或一系列函数。

代码块（Code chunk）:在 R Markdown 中，代码块是放置 R 可执行代码的区域（用灰框标记）。在本书中，代码块指的是一系列一起执行的 R 指令。

Knit：在 R 中，Knit 的意思是把 R Markdown 文件转换为 HTML、Word 或 PDF 文档。

YAML（Yet Another Markup Language）：R Markdown 文档开头的代码，用于指定整个文档范围内的特征和特性。

R 脚本（R-script）：我们可以在 RStudio 中创建文件，用来存储在分析过程中使用的指令，以便未来使用或重现分析过程。R 脚本对于用文字描述进行了什么分析用处不大，对于展示目的（创建文档）而言也是如此。

1 译注：感兴趣的读者可以去了解 posit 公司（原 RStudio）新推出的 Quarto。它与 R Markdown 类似，但更强大。

变量（Variable）：记录人、地点或事物集合独特特征的数值列表。

数据集（Data set）：变量的集合。

第 2 章

理论（Theory）：假设（概念）的集合（或系统），用于解释所关注的事物。理论起码要有一些实际依据，即存在某些证据支持。

假设（Hypothesis）：基于有限经验数据的猜测；是任何调查的起点。

模型（Model）：对现实的一种近似；包括一组变量列表，以及它们之间和它们与所关注的现象之间的联系。

诊断（Diagnostic）:一种数据视图或统计摘要，用于了解模型估计是如何产生的，以及模型假设是否成立。诊断也用于发现离群值。

ggplot：一个图形包。“gg”代表“图形语法（grammar of graphics）”概念框架，通过不同的视觉对象（线条、点、颜色、标签）的相互层叠来生成数据的可视化。

散点图（Scatter plot）：一种二维网格，通过在该网格上放置观测来展示两个变量之间的关系。

平滑（Smooth）：用于在散点图中概括总体模式的直线或曲线。

线性（Linear）：任何可以用直线来描述或概括的东西。

总结（Summarize）：用简单的图形表示（例如，直线或曲线）或用描述数据的简单数值（例如，平均数、中位数或标准差）来说明能够描述数据的模式、形状存在或不存在。在统计术语中,这些形状或数字表示被称为对数据的“总结”（或概括）。

正相关（Positive association）：两个或多个变量之间的关系在变化方向上是一致的。当一个变量增长时，其他变量也增长。

二元回归模型（Bivariate regression model）：具有两个变量的回归模型，一个自变量（原因）和一个因变量（结果）。

多元回归模型（Multiple regression model）：具有多个自变量的回归模型。

负相关（Negative association）：两个或多个变量之间的关系在变化方向上是相反的。当一个变量增长时，其他变量在减小。

稳定性（Stability）：根据在估计中纳入的变量或观测的变化，模型估计记录的变化量。它也可以是由于模型函数形式的变化，而在模型估计中记录的变化。

增加变量图（Added variable plot）：在保持模型中其他所有变量不变，同时消除其他自变量影响的情况下，展示模型自变量与因变量之间的关系；也被称为偏回归图（partial regression plot）。增加变量图中的直线斜率与相应的回归系数斜率相同。

残差图（Residual plot）：由回归模型的实际值与预测值之差和因变量绘制所得；y 轴为残差，x 轴为拟合（或预测）值的散点图。这种可视化可以用于发现离群值，以及评估 OLS 是否是合适的估计量。

第 3 章

数据集（Data set）：指明一组人、地点或事物的数量或者特征的数字或类别的集合（一系列行和列）。通常，每列代表一个特征，每行代表一个特定的人、地点或事物。

连续变量（Continuous variable）：计数或计量的变量。连续变量可以取最小值和最大值之间的任意值。

分类变量（Categorical variable）：表示种类差异的变量。分类变量记录着人、地点或事物存在的不同状态或者特征。

有序分类变量（Ordered categorical variable）：一种分类变量，其类别可以根据其背后维度进行排序（例如，从保守到自由的政党光谱；从小学毕业到获得博士学位的受教育程度；根据建议刑期长度不同的犯罪类别）。

形状（Shape）：变量的案例在其范围内的分布（位置）。

分布（Distribution）：此处与形状同义。

密度图（Density plot）：表示变量分布的图。它表明大多数案例位于最大值与最小值之间的何处。

偏态分布（Skewed distribution）：不对称变量的形状。变量在平均数之上和之下的分布形状不呈镜像。

直方图（Histogram）：记录着在变量范围内各区间的案例数量的一种可视化。直方图也提供了实用的数据形状或分布视图。直方图适用于连续变量。

极差（Range）：变量的最小值和最大值之间的差。

dplyr：一个 R 包，用于根据一组特定标准生成数据表。例如，命令可以指定仅列出数据集中的三个变量，且根据其中一个变量列出最高的 10 个值。这个包对于数据管理和数据描述非常有用。

tibble：R 中类似于数据表的数据结构。其目的是在 R 中轻松操纵数据集，让“dplyr”

能够处理数据。

有效性（Validity）：度量（变量）代表概念的程度。

可靠性（Reliability）:度量在跨个体（例如,该度量是否记录了个体间的重要差异）和跨时间（例如，如果对同一个个体重复测量，度量能否记录到相同的数量）比较时的准确度。

第 4 章

众数（Mode）：包含最多观测、个案或个体的变量类别或者值。

平均数（Mean）：变量的平均值。平均数是由所有观测值相加，然后除以观测数量计算得到的。

中位数（Median）：变量的中间个案。中位数是将变量中所有的观测值从最低到最高排列，然后挑出中间个案计算得到的。在有 9 个观测值（奇数个）的变量中，第 5 个案例代表中位数。在有 10 个观测值（偶数个）的变量中，只需计算第 5 个案例和第 6 个案例的平均数即可。

四分位距（Interquartile range）：变量第 25 和第 75 百分位数之间的距离；大致可以认为是“中间一半”的变量。

标准差（Standard deviation）：变量的所有个案与平均数的平均距离，即与变量平均数距离平方的平均数再开平方根。因此，它是以变量的原始单位表示的。

方差：标准差的平方，即与平均数距离平方的平均数。因此，它并不是以原始单位表示的，且在提供变量分散程度信息方面用处不大。

第 5 章

单变量描述（Univariate description）：单个变量的可视化，旨在揭示变量的集中趋势和离散程度。

频率表（Frequency table）：记录分类变量各类别观测数量的表格。频率表适用于分类变量。

条形图（Bar plot）：频率表的可视化。图中的每个条形都代表分类变量中的一个类别。条形长度代表各类别的观测数量。条形图适用于分类变量。

箱线图（Boxplot；Box-and-whisker plot，盒须图）：变量分布的可视化。“方框”代表四分位距，而“须”代表四分位距两侧远离中心的值。箱线图适用于连续变量。

偏度（Skewness）：变量分布偏离正态分布的程度；变量分布不对称的程度。

茎叶图（Stem-and-leaf plot）：代表水平方向直方图的数据可视化。茎代表连续变量的第一数值分区，而叶代表每个分区中存在的单个案例。分区可以代表数字中不同的小数位数。当观测数量相对较少时，茎叶图适用于连续变量。

双变量描述（Bivariate description）：旨在揭示两个变量关系的数据可视化。

负相关（Negative association）：两个变量之间的关系是，其中一个变量增加，则另一个变量减小。

线性关系（Linear relationship）：两个连续变量之间的关系是，其中一个变量的增加或减小，对应着另一个变量恒定量的增加或减小。

相关（Correlation）（系数）：两个连续变量之间关系的数值摘要。介于 –1（强负相关）和 1（高正相关）之间。强相关（接近 1 或 -1）表明一个变量的变化与另一个变量的变化有关。

马赛克图（Mosaic plot）：两个分类变量的可视化，表示一个变量的各类别与另一个变量的各类别共有的观测数量。列的粗细也表明被认定为 X 的变量有多少个案例。马赛克图适合在考察两个分类变量之间的关系时使用。

交叉表（Cross-tab）：马赛克图的数字表示。每个单元格展示的都是一个变量的各类别与另一个变量的各类别共有的观测数量。交叉表还表明了每个单元格中变量的案例占比。交叉表适合在考察两个分类变量之间的关系时使用。

气泡图（Bubble plot）：根据第三个变量调整观测大小的散点图。第三个变量可以是连续的或分类的，而 x 轴和 y 轴上的变量应该是连续的。

第 6 章

数据变换（Transforming data）：以符合假设或能更好地可视化数据的形式重新表示数据。重新表示数据的形式可以是使用数学公式（取对数），将连续变量转换为分类变量，或者改变分类变量中的分类数量。

Box-Cox 变换阶梯（Box-Cox ladder of transformations）：我们可以用一组数学表达式来变换连续变量。这些表达式构成的变换过程，有助于我们从视觉上对数据进行检查。

对数变换（Logarithmic transformations）：对数变换可能是最常见的数据变换了。取对数是根据给定底数计算产生一个数字所需的指数，来重新表示这个数字的。两

个最常见的对数底数为 10 和所谓的自然对数（2.718）。

自然对数（Natural log）：自然对数为 2.718。通过取自然对数重新表示任何数字时，就是在计算将 2.718 转换为该数字所需的指数。例如，如果取 10 的自然对数，我们计算的是能够得到 10 的 2.718 的指数，即 $2.718^{(x)}=10$。本例中 x 为 2.30。换句话说，$2.718^{(2.30)}=10$。

以 10 为底：当以 10 为底取对数重新表示数字时，就是在计算将 10 转换为该数字所需的指数。例如，以 10 为底取 10 的对数，我们计算的是能够得到 10 的 10 的指数，即 $10^{(x)}=10$。本例中 x 为 1。换句话说，$10^{(1)}=10$。

第 7 章

图表杂乱（Chart clutter）：不要被过多的信息搞晕。图表杂乱涵盖了所有多余的图或表的特征，这些特征没有传递任何信息，因此是不必要的。图表杂乱通常与样式有关，例如模糊的坐标轴、标签周围的方框、奇怪的字体，以及灰色的背景。

解释性（Explanatory）：解释性的可视化应该强调一个观点；打造抓住读者眼球的关键洞见。在这类图中，应避免过多的信息。

探索性（Exploratory）：探索性的可视化以中立的方式产生信息，以便对潜在的竞争假设给予同等的重视。

信息性（Informational）：信息性的可视化，每平方英寸要尽可能多地产生信息。人们可以将其视为参考基准，读者可以回头仔细研究，以便更全面地理解某种关系或数据的整体结构（集中趋势、分散程度和案例所处的位置）。

背景 / 语境 / 情境（Context）：在讲故事时，它是开头的部分。引入角色及其面临的状况。在呈现数据时，引入议题和问题。它还建立了受众运用数据和度量的直觉。

因果关系（Causation）：解释为什么两个变量之间相互关联。它讲述了为什么两件事情是相关的。

简化形式（Reduced form）：通常因果解释涉及从 x 引发 y，以及由 y 导致 z 等多种机制。这种因果变化的简化形式是 x 和 z 之间的关系。

第 8 章

总体（Population）：代表关注焦点的一组限定的个体或对象。例如，如果想知道谁可能会赢得美国总统大选，我们关注的总体就是登记选民群体。

样本（Sample）：总体的一个子集。

随机样本（Random sample）：从总体中抽取的样本，其中每个个体或对象被选中的概率相同。

样本偏差（Sample bias）：从总体中抽取样本时，每个个体或对象被选中的概率不同，就会有样本偏差。例如，如果关注的总体是一所大学的学生群体，却只从一个班级中抽取样本，这并不能让这所大学的每个学生都有机会被抽到。

有放回抽样（Sampling with replacement）：每个个体或对象从总体中抽出后，又被放回总体中，以便它们与其他所有个体或对象一起有相同的概率再次被选中。这确保了每个个体或对象都有相同的概率被选中，且每次选择或抽取都是彼此独立的。

大数定律（Law of large numbers）：如果某个偶然性实验在完全相同的条件下重复无限次，而且这些重复是相互独立的，那么某个事件 A 发生的次数将以概率 1 收敛到一个数字，这个数字等于 A 在一次实验重复中发生的概率。

抽样分布（Sampling distribution）:从一系列随机样本中获得的汇总统计量（例如，平均数、中位数、标准差等）的分布。

中心极限定理（Central limit theorem）:从总体中随机抽取样本时确立了以下特征：①随着每个样本观测数量的增加，总体中平均数的抽样分布将趋于正态；②随着每个样本观测数量的增加，总体中和的抽样分布将趋于正态；③从正态分布的总体中抽取样本时，无论观测数量是多少，抽样分布都将趋于正态。

正态分布（Normal distribution）：以样本或总体平均数为中心的对称钟形曲线。

均匀分布（Uniform distribution）：在变量范围内观测数量恒定的分布。均匀分布的形状看起来像矩形。

指数分布（Exponential distribution）：这种分布的观测数量在变量的最小值附近最多，随着最大值的接近而逐渐减少。

双峰分布（Bimodal distribution）：频率最高的观测值出现在变量范围内两个不同的区域，看起来就像有两个“驼峰”在分布中。

标准正态分布（Standard normal distribution）：完全对称的正态分布，以平均数 0 为中心，标准差为 1。在曲线下方全部的面积中，68% 的分布位于平均数上下 1 个标准差之间，95% 的分布位于平均数上下 2 个标准差之间，99.75% 的面积位于平均数上下 3 个标准差之间。这些性质提供了计算置信区间和进行假设检验的机制。

***z*-分数（*z*-score）**：观测值与样本平均数之差，再除以样本标准差。*z*-分数可以

将所度量的任何事物标准化，并将其置于以标准差表示的单位下。这让我们能够确定有多少分布位于观测值的左边或右边。

第 9 章

置信区间（Confidence interval）：有时指的是选举民意调查的误差范围。置信区间代表的是在给定的置信水平下，包含总体参数的样本统计量的取值范围。置信水平越高，区间越大。

总体比例（Population proportion）：属于某一类别的总体百分比。在选举中，总体比例指的是支持某一特定候选人的那部分群体比例。总体比例与总体平均数的置信区间计算公式不同。

临界 *z*- 值（Critical *z*-value）：在标准正态分布中，能够代表给定曲线下方面积的，以标准差度量的与平均数的距离。例如，如果选择 90% 的置信水平，就用 1.65 作为临界 *z*- 值，因为标准正态曲线下方 90% 的面积在平均数上下 ±1.65 个标准差之间。

***t*- 分布（*t*-Distribution）**：与标准正态分布类似，只不过其目的是明确针对小样本。此外，与标准正态分布不同，有许多不同的 *t*- 分布，这都取决于样本中的观测数量有多少。最终，随着观测数量的增加，*t*- 分布的形状会逐渐接近标准正态分布。

自由度（Degrees of freedom）：用于计算统计量的独立数据点的数量。因此，在确定要用哪个 *t*- 分布时，使用的是自由度（$n - 1$），而不是观测数量（n）。这个概念在多种不同的统计学语境下都有出现。

偏差（Bias）：估计的抽样分布的平均数与总体参数之间的差异。如果估计的抽样分布的平均数不以总体平均数为中心，那么这个估计量是有偏差的。

贝塞尔校正（Bessel's correction）：当分母用 n 时，样本标准差是有偏差的（平均而言，是向下的偏差），因此分母中的观测数量减 1 能考虑到偏差，从而得到无偏估计量。

临界 *t*- 值（Critical *t*-value）：*t*- 分布中以标准差度量的与平均数的距离。例如，如果选择 90% 的置信水平，就用 1.697 作为临界 *t*- 值，因为标准正态曲线下方 90% 的面积在平均数 0 上下 ±1.697 个标准差之间。

抽样方差（Sampling variance）：当从总体中抽取样本并计算统计量时，很少会和总体参数完全吻合。样本统计量与总体参数之间的差异被称为抽样方差或误差方差。

零假设（Null hypothesis）：在进行假设检验（比较两个样本的平均数）时，零假设就是平均数之间没有差异。如果 *t*- 统计量表明两个平均数之间存在显著差异，那

么就“拒绝”零假设。

双样本 *t* 检验（Two-sample *t*-test）：该统计检验围绕两个样本的平均数差异构建置信区间。它表示两个样本之间的平均数差异不为 0 的可能性，即零假设。

（双样本 *t* 检验的）*t*- 比率（*t*-Ratio）：通过样本统计量与零假设（通常为零）之间的差异，除以合并样本方差（pooled sample variance）的平方根计算得到的统计量。

第 10 章

探索性数据分析（EDA，exploratory data analysis）：由约翰 · 图基（John Tukey）创造的术语，用于描述一种分析过程，其特征是在产生假设和探讨数据之间来回往复迭代。经典过程是形成假设，然后用数据进行测试，在实验室模拟可能发生的情况。EDA 认为，先查看数据有利于形成更好的假设。EDA 要考虑的一个重要缺点是，来回往复迭代的方法可能会违背统计理论的核心假设（观测的随机选择），导致传统统计检验的用处没有那么大。

情感量表（Feeling thermometer）：用于获取受访者对某个人、地点或事物看法的一种调查工具。通常，受访者会被要求分享他们对某事或某人的感受——有多喜欢，或者有多热烈。量表被设计为生成连续变量，范围通常在 0 和 100 之间。

抖动图（Jitter plot）：找出箱线图的各个案例，将点分开（抖动）以便辨识。箱线图是一种非常有价值的数据可视化方式，而箱线图隐藏的细节则能通过抖动图揭露出来。

第 11 章

受控比较（Controlled comparison）：用于展示两个变量的关系是如何根据第三个变量发生变化的双变量视图。被称为“受控”比较，是因为它们在控制第三个变量的情况下，展示了两个变量之间的关系。等同于在论证两个变量之间的因果关系时，考虑第三个变量。

二分变量（Dichotomous variable）：有两个可能取值的变量。如果用数字表示，取决于某种条件、特征或特质是否存在，变量通常记为 0 或 1。那些被当作二选一的分类变量也是二分变量（例如，是否接受过大学教育、男性或女性、赢或输）。二分变量也被称为“虚拟（dummy）”变量。

第 12 章

线性回归（Linear regression）：在两个变量的情境下（本章的重点），线性回归将一条概括了两个变量之间关系的直线拟合到散点图上。

实质显著性（Substantive significance）：指的是回归模型中系数的大小：x 的单位变化对应的 y 的变化。

统计显著性（Statistical significance）：指的是我们对估计系数不为零的确定程度（或置信水平）。如果我们能拒绝估计是偶然结果的可能性，那么该估计在统计上是显著的。

斜率（Slope）：x 的单位变化对应的 y 的变化。

截距（Intercept）：回归线与 y 轴截距的交点。有时它揭露了重要的信息，有时则用处不大。

系数（Coefficient）：回归线估计的斜率。

预测值（Predicted value）：回归线本身。对于给定的 x 的值，预测的 y。

残差（Residual）：观测的实际值与预测值之间的差异 $(y_i - \hat{y}_i)$。

R^2 统计量（R^2 statistic）：y 的可解释方差 $\left[\sum(\hat{y}_i - \bar{y})\right]$ 除以 y 的总方差 $\left[\sum(y - \bar{y})\right]$。这是一个范围从 0 到 1 的数字，表明 y 的变化有多少是由模型中的变量解释的。

总偏差（Total deviation，在 R^2 的语境下）：Y_i 与 $\bar{Y}$ 之间的差异。

可解释偏差（Explained deviation，在 R^2 的语境下）：预测值 $\hat{Y}_i$ 与平均数 $\bar{Y}$ 之间的差异。

无法解释的偏差（Unexplained deviation，在 R^2 的语境下）：观测值 Y_i 与预测值 $\hat{Y}_i$ 之间的差异。

（回归系数的）t- 比率（t-Ratio）:统计显著性的度量，即回归系数除以其标准误差。它表明对于描述两个变量之间关系的直线，我们有多少信心认为其真实斜率不为零。绝对值大于 2 表示我们可以确信，按照常规的统计显著性标准，真正的斜率不为零。

第 13 章

多元回归分析（Multiple regression analysis）：在控制其他自变量的情况下，拟合一条线来概括两个变量（一个因变量和一个自变量）之间的关系。

加性多元回归模型（Additive multiple regression model）：一种严格的线性模型，假设模型中的自变量是相互独立的。

校正后的 R^2 统计量（Adjusted R^2 statistic）：R^2 统计量的一种版本，考虑了模型中预测变量的数量，只有增加的预测变量改善了模型的拟合度，这个值才会增加。其解读方法与 R^2 统计量相同。

经验蕴涵（Empirical implication）：根据现有的经验关系，在逻辑上得出的假设关系。

第 14 章

虚拟变量（Dummy variable）：由 0 和 1 组成的变量，表示案例是被归类到一个类别还是另一个类别。在 R 中，虚拟变量可以只包含两个不同的字符串，表示案例是属于一个类别还是另一个类别（例如，民主或独裁、男性或女性、共和党或民主党）。

加性模型（Additive model）：假设自变量之间没有关系的多元回归模型。从数学上讲，加性模型的方程含有多个相互相加的项：$y=a+b+c$。

交互作用模型（Interactive model）：这种多元回归模型不假设所有自变量之间是独立的。从数学上讲，交互作用模型的方程含有多个项，其中某些项是相乘的：$y=a+b+c+a\times c$。

第 15 章

效率（Efficiency）：用于比较估计量或统计量的标准。如果估计量的抽样分布具有较小的方差，那么它会被认为是更有效的。如果一个统计量或估计量的抽样分布的方差小于另一个，则意味着一般而言，它对总体参数的估计更准确。

一致性（Consistency）：估计量的准确度随着观测数量的增加而增加。

高斯－马尔可夫定理（Gauss-Markov theorem）：指出如果满足三个重要条件，那么 OLS 就是最佳线性无偏估计量。这三个条件如下：①模型误差之和为零；②误差之间相互独立；③误差的方差恒定。

误差（Error）：观测的实际值与真实模型产生的预测值之间的差异。由于我们永远无法知道真实模型是什么，所以误差只是一个理论上的概念。

第 16 章

离群值（Outlier）：这个术语通常用来描述和其他数据与众不同的观测。本章中

给出了一个更具体的定义：在 y 轴上和其他数据与众不同的观测。

杠杆值（Leverage）：自变量（x 轴上）和其他数据与众不同的观测。

影响力（Influence）：任何给定的观测对于多种不同的统计摘要或回归模型估计产生的影响。

库克距离（Cook's D）：基于残差以及杠杆值的影响力的度量。它衡量的是某个观测对模型估计的影响。由此可见，库克距离衡量的是当从回归中移除第 i 个观测时，所有预测值的变化程度。

dfbeta：一种影响力的度量，记录的是某个观测对每个回归系数的影响。

第 17 章

逻辑回归（Logistic regression）：用逻辑曲线拟合数据的回归模型来解释二分（二元）结果。

对数发生比（Logged odds）：发生比的对数。发生比就是（事件发生的）概率 x 除以（事件不发生的）概率 $1-x$。

预测概率（Predicted probability）：逻辑回归模型预测的某个结果（通常由记为“1”的二元变量表示）的发生概率。在逻辑模型中，预测概率遵循 S 形曲线。类似于 OLS 回归中的预测值 $\hat{y}_i$，自变量的每个值都存在一个预测概率。对于某个具体的 x，我们说编码为“1”的结果有某概率发生。

McFadden's R^2：一个拟 R^2 统计量，通过在仅有常量的简单模型中增加自变量，展示该自变量对模型的改进。这是 `logreg2()` 函数生成的三个拟 R^2 度量中最保守的一个。

Cox-Snell R^2：一个拟 R^2 统计量，通过在仅有常量的简单模型中增加自变量，展示该自变量对模型的改进。它比 Nagelkerke R^2 更保守，但不像 McFadden's R^2 那么保守。

Nagelkerke R^2：一个拟 R^2 统计量，通过在仅有常量的简单模型中增加自变量，展示该自变量对模型的改进。其提供的估计是所有拟 R^2 度量中最不保守的。

卡方检验（Chi-square test）：该检验统计量为空偏差（来自自变量较少的模型）和残差偏差（来自完整模型）之间的差异。然后计算卡方数的 p 值，表明完整模型是否比自变量较少的模型有所改进。p 值小于 0.05，表明完整模型比常规的显著性水平有改进。

接收者操作特征曲线（ROC curve，Receiver operating characteristic curve）：描绘了检验或模型在不同阈值下灵敏度和特异度的关系。其提供的指标表明了模型（或检验）区分两种逻辑回归结果的效果如何。

曲线下面积（AUC，Area under the curve）：提供了 ROC 曲线下面积的数值摘要。数字越大，越接近 1，越远离 0.5，意味着模型很好地解释了为什么你观察到的是某个结果，而不是另一个。

真阳性（True positive）：检出阳性，结果也为阳性。类似于医学检测，患者检出阳性且确实患病。

假阳性（False positive）：检出阳性，而结果实际上为阴性。类似于医学检测，患者检出阳性，却没有患病。

真阴性（True negative）：检出阴性，结果也为阴性。类似于医学检测，患者检出阴性且确实没有患病。

假阴性（False negative）：检出阴性，而结果实际上为阳性。类似于体检，患者检出阴性，实际却患病。

敏感度（Sensitivity）：获得真阳性的概率。类似于患者患病被检出的概率。

特异度（Specificity）：获得真阴性的概率。类似于患者没有患病检出阴性的概率。

参考文献

第1章

Gaubatz, K. T. (2015). *A survivor's guide to R*. SAGE.

第2章

Alesina, A. (2003). The size of countries: Does it matter? *Journal of the European Economic Association*, *1*(2-3), 301-316.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., & Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth*, *8*(2), 155-194.

Cukier, K., & Mayer-Schönberger, V. (2013). *Big data: A revolution that will transform how we live, work, and think*. Eamon Dolan/Houghton Mifflin Harcourt.

Jones, C. (2015). *The facts of economic growth*. National Bureau of Economic Research.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science*, *343*(6176), 1203-1205.

Lloyd, P., & Lee, C. (2018). A review of the recent literature on the institutional economics analysis of the long‐run performance of nations. *Journal of Economic Surveys*, *32*(1), 1-22.

Paulos, J. A. (1995). *A mathematician reads the newspaper*. Basic Books.

Pollock, P. H. (2014). *An R companion to political analysis*. SAGE/CQ Press.

Przeworski, A., Alvarez, R. M., Alvarez, M. E., Cheibub, J. A., Limongi, F., & Neto, F. P. L. (2000). *Democracy and development*. Cambridge University Press.

"Violent Crime Is Down in Chicago." (2018 May 5). *The Economist*. https://www.economist.com/united-states/2018/05/05/violent-crime-is-down-in-chicago.

第 3 章

de Fezensac, M. (1852). *A journal of the Russian campaign of 1812*. Parker, Furnirall & Parker.

Wickham, H., & Grolemund, G. (2017). *R for data science*. O'Reilly.

第 4 章

Cassidy, J. (2009). *How markets fail: The logic of economic calamities*. Farrar, Strauss, and Giroux.

Fisher, M., & Keller, J. (2017 November 7). What explains US mass shootings? International comparisons suggest an answer. *New York Times*.

第 5 章

Lucas, R. E., Jr. (1988). On the mechanics of economic development. *Journal of Monetary Economics*, *22*(1), 3-42.

Ruger, W. P., & Sorents, J. (2009 February). *Freedom in the 50 states: An index of personal and economic freedom*. George Mason University Mercatus Center.

Tufte, E. R. (1997). *Visual explanations*. Graphics Press LLC.

第 6 章

Binswanger, M. (2006). Why does income growth fail to make us happier? Searching for the treadmills behind the paradox of happiness. *Journal of Socio-Economics*, *35*(2), 366-381.

Przeworski, A., Alvarez, R. M., Alvarez, M. E., Cheibub, J. A., Limongi, F., & Neto, F. P. L. (2000). *Democracy and development*. Cambridge University Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

第 7 章

Duarte, N. (2010). *Resonate: Present visual stories that transform audiences*. Wiley.

"Economists Are Rethinking the Numbers on Inequality." (2019 November 28). *The Economist*. https://www.economist.com/briefing/2019/11/28/economists-are-rethinkingthe-numbers-on-inequality.

Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. Wiley.

Piketty, T. (2014). *Capital in the twenty-first century*. Belknap Press.

Piketty, T., & Saez, E. (2003). Income inequality in the United States, 1913–1998. *Quarterly Journal of Economics*, *118*(1), 1-41.

Piketty, T., Saez, E., & Zucman, G. (2018). Distributional national accounts: methods and estimates for the United States. *Quarterly Journal of Economics*, *133*(2), 553-609.

Piketty, T., & Zucman, G. (2014). Capital is back: Wealthincome ratios in rich countries 1700–2010. *Quarterly Journal of Economics*, *129*(3), 1255-1310.

Putnam, R. (1993). *Making democracy work: Civic traditions in modern Italy*. Princeton University Press.

Strunk, W., & White, E. B. (1979). *The elements of style* (3rd ed.). Macmillan Publishing Company.

Tufte, E. R. (2006). *Beautiful evidence*. Graphics Press LLC.

Vogler, C. (2007). *The writer's journey: Mythic structure for writers* (3rd ed.). Michael Wiese Productions.

第 8 章

Stigler, S. M. (2016). *The seven pillars of statistical wisdom*. Harvard University Press.

Tijms, H. (2004). *Understanding probability: Chance rules in everyday life*. Cambridge University Press.

第 9 章

Schumacker, R. E. (2015). *Learning statistics using R*. SAGE.

第 10 章

Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. Wiley.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

第 11 章

Karstedt, S. (2006). Democracy, values, and violence: Paradoxes, tensions, and comparative advantages of liberal inclusion. *Annals of the American Academy of Political and Social Science*, *605*(1), 50-81.

Karstedt, S. (2013). *Legitimacy in non-democratic regimes* (J. Tankebe & A. Liebling eds.). Oxford University Press.

LaFree, G., & Tseloni, A. (2006). Democracy and crime: A multilevel analysis of homicide trends in forty-four countries, 1950-2000. *Annals of the American Academy of Political and Social Science*, *605*(1), 25-49.

第 12 章

Tufte, E. R. (1974). *Data analysis for politics and policy*. Prentice-Hall.

第 13 章

Cukier, K., & Mayer-Schönberger, V. (2013). *Big data: A revolution that will transform how we live, work, and think*. Eamon Dolan/Houghton Mifflin Harcourt.

Greene, W. H. (2000). *Econometric analysis* (4th ed). Prentice-Hall.

第 16 章

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed). SAGE.

第 17 章

Pollock, P. H. (2014). *An R companion to political analysis*. SAGE/CQ Press.

配套学习资源

R 下载地址：https://www.r-project.org

RStudio 下载地址：https://www.rstudio.com

R Markdown 速查手册下载地址：https://resources.rstudio.com/rstudio-develop/rmarkdown-2-0

案例数据文件下载地址：https://edge.sagepub.com/brownstats1e/student-resources/code-data-sets