第 3 章中“有序分类变量”下的案例

让我们再看另一个例子，来说明类别排序如何改变我们看到了什么（见图 3-3）。国家选举研究调查中记录了个人婚姻状况的变量（marstat），该变量有六种不同的类别，没有明显的顺序：已婚（Married）、分居（Separated）、离婚（Divorced）、丧偶（Widowed）、单身（Single）和同居（Domestic partnership）。为了便于展示，首先我将“Domestic Partnership（同居）”标签缩写为“Dom. Part.”（见代码块 3-3a）。

代码块 3-3a

```
levels(nes$marstat)[levels(nes$marstat)=="Domestic Partnership"] <-
"Dom. Part."
```

然后，我根据生活中的经历，将 marstat 变量中的类别重新排序：单身、已婚、同居、分居、离婚或丧偶，并创建了另一个变量 nes$g。为了说明另一种重新排列变量中类别的方法，在代码块 3-3b 中我使用了与每个类别对应的数字，而非实际的类别名称。我建议你尝试使用这个命令改变数字的顺序，看看它对类别的顺序影响如何。要查看类别的顺序如何变化，先运行 levels(nes$marstat)。有了顺序不同的新变量 nes$g 后，再运行 levels(nes$g)。

代码块 3-3b

```
nes$g <- factor(nes$marstat,
                levels(nes$marstat)[c(5, 1, 6, 2, 3, 4)])
```

在这个例子中，假设我们对婚姻状况和个人对同性恋看法之间的关系感兴趣。与上一个例子一样，我绘制了两张图分别展示无序和有序版本的 marstat 变量（见代码块 3-3c）。

代码块 3-3c

```
fig1 <- ggplot(nes, aes(x=marstat, y=ftgay)) +
  stat_summary(fun.y=mean, geom="bar", fill="#0000ff",
               aes(group=1))  +
  ylab(" 未排序 ") +
  xlab("") +
  theme_minimal() +
  coord_flip(ylim=c(45,70)) +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))

fig2 <- ggplot(nes, aes(x=g, y=ftgay)) +
  stat_summary(fun.y=mean, geom="bar", fill="#0000ff",
               aes(group=1))  +
  ylab(" 已排序 ") +
```

```
  xlab("") +
  theme_minimal() +
  coord_flip(ylim=c(45,70)) +
  theme(plot.title = element_text(size = 8, face = "bold"),
        axis.title = element_text(size = 8, face = "bold"))
```

再次使用 grid.arrange() 函数将定义好的两张条形图并排放置在页面上（见代码块 3-3d）。

代码块 3-3d

```
grid.arrange(fig1, fig2, ncol=2,
             top = textGrob("图 3-3:婚姻状况和对同性恋情感量表的无序和有序视图",
                            gp=gpar(fontsize=10,fontface = "bold")))
```

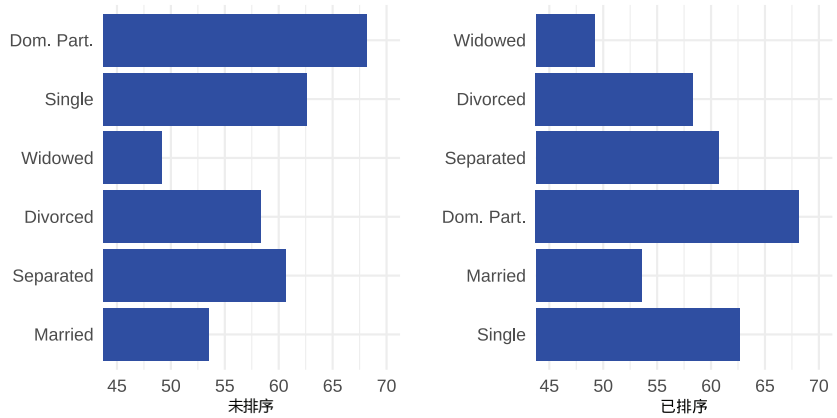

图 3-3：婚姻状况和对同性恋情感量表的无序和有序视图

如果我们不关心类别如何排序，而只是简单地计算并用条形图来观察每个类别的平均情感量表得分，那么相对而言，就很难理解婚姻状况和对同性恋看法之间的关系（左侧的条形图）。我们看到丧偶者（Widowed）对同性恋的看法得分最低，而同居者（Dom. Part.）对同性恋的看法得分最高，但这两个类别为何相邻没有深层的原因。

观察一下，如果我们根据个人在一生中经历的不同婚姻状态来重新排列类别：单身（Single）、已婚（Married）、同居（Domestic Partnership）、分居（Separated）、离婚（Divorced）和丧偶（Widowed）（暂时不考虑离婚后再婚的情况），会发生什么情况？在对类别重新排序绘制的条形图中，我们发现了不同的模式（右侧的条形图）。右侧的条形图显示，年长的受访者对同性恋的看法得分较低，已婚的和同居的受访者之间的变化也非常显著。对类别的不同排序方式，可以将我们的目光吸引到不同的关系上，并提醒我们注意其他因素。就婚姻状况和对同性恋的看法而言，不同的观点表明年龄和宗教（表现为传统婚姻或同居关系）可能很重要。

第 5 章中“箱线图（双变量）”下的案例

下面的例子说明了为什么箱线图这么有用。其关键属性在于方框：表明了中间那一半数据（四分位距）是否随类别变化。它们不仅表明了中位数的情况，而且还展示了分布是如何变化的。

在之前的箱线图代码中，我只想总结一个变量的情况。在本例中（见代码块 5-18），我想为每类政党认同都创建一个箱线图（见图 5-15）。因此，在图形属性函数 `aes()` 中，我指定了两个变量：`pid7` 和 `ftgay`。

代码块 5-18

```
ggplot(subset(nes, pid7!="NA" & pid7!="Not sure"),
       aes(pid7, ftgay)) +
  geom_boxplot(col="#0000bf") +
  theme_minimal() +
  theme(axis.text.x = element_text(size=8, angle=45, vjust=0.7)) +
  ggtitle("图 5-15：党派影响了看法") +
  ylab("同性恋情感量表") +
  xlab("政党认同") +
  coord_flip() +
  theme(plot.title = element_text(size = 8, face = "bold")) +
  theme(axis.title = element_text(size = 8, face = "bold"))
```

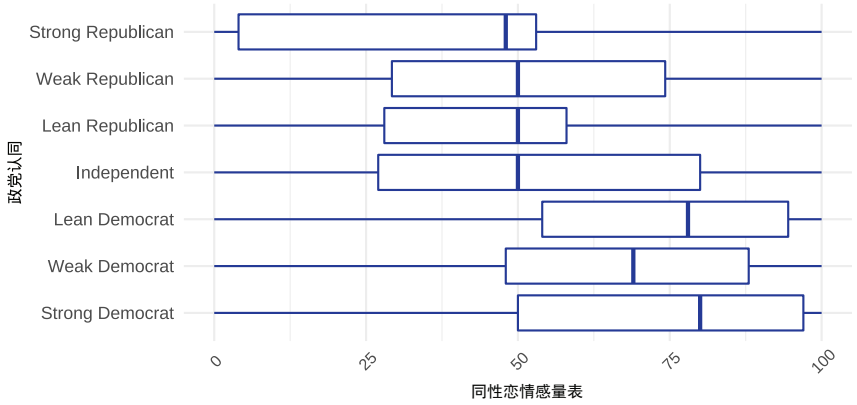

图 5-15：党派影响了看法

图 5-15 说明了当按政党认同（Party ID）来观察对同性恋的支持时，中位数和四分位距是如何变化的[1]。请注意，在比较无党派人士和共和党人时，中位数的数值并没

1 国家选举研究（NES）数据中的许多变量都是情感量表——用 0 ~ 100 的变量来表示调查对象对某些人或政策的感受。例如，同性恋情感量表，要求受访者用 0 ~ 100 来表示他们对同性恋的看法。

有发生变化。如果只看中位数，我们就会得出没有区别的结论。然而，箱线图讲述了一个非常不同的故事：与无党派人士相比，坚定的共和党人在情感量表上的四分位距得分要低得多。

再看另一个例子（见代码块 5-19）。world 数据集中一个令人关注的变量记录了中小学女孩与男孩的入学比率。比率越大，女孩相对于男孩的教育成果越好。在比较中位数时，图 5-16 中的箱线图表明威权政体〔文官独裁（Civilian Dictatorship）和军事独裁（Military Dictatorship）〕和民主政体〔议会制民主（Parliamentary Democracy）或总统制民主（Presidential Democracy）〕之间存在的差异相对较小。

代码块 5-19

```
ggplot(world, aes(regime, gtbeduc, na.rm=TRUE)) +
    geom_boxplot(col="#0000bf", na.rm=TRUE) +
    theme_minimal() +
    ggtitle("图 5-16：威权政体下的巨大差异") +
    ylab("女孩与男孩受教育程度之比")+
    xlab("") +
    coord_flip() +
    theme(plot.title = element_text(size = 8, face = "bold")) +
    theme(axis.title = element_text(size = 8, face = "bold"))
```

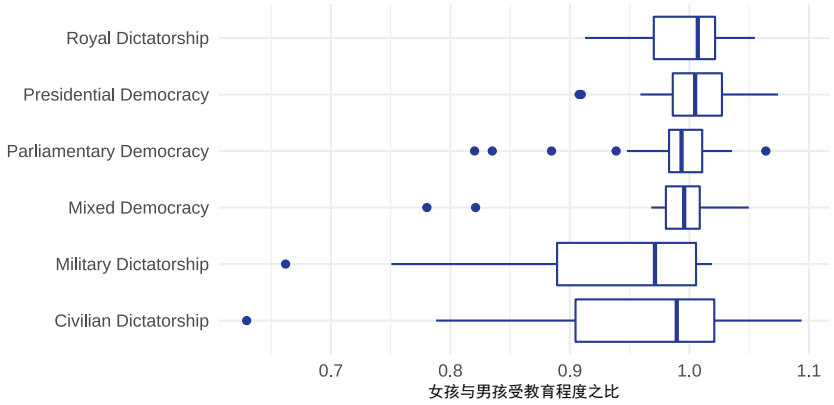

图 5-16：威权政体下的巨大差异

然而，当我们比较分布时，就能发现明显的差异。具体而言，在威权政体中下四分位数下降了。显然，威权政体无法保证女孩获得足够的教育。

让我们从箱（box）里面找出“罪魁祸首”。为了构建图 5-17，我使用 R 中的 subset() 函数创建了一个新的数据集，将样本限制在两种政权类型上〔文官独裁（Civilian Dictatorship）和军事独裁（Military Dictatorship）〕。然后，我让 R 抖动它们

的位置，让标签不会重叠。还要注意的是，scale_fill_manual() 函数用来准确指定想要的颜色（见代码块 5-20）。

代码块 5-20

```
regimedata <- subset(world, regime=="Civilian Dictatorship" |
                     regime=="Military Dictatorship")

ggplot(regimedata, aes(regime, gtbeduc, na.rm=TRUE)) +
     geom_boxplot(outlier.colour=NA, fill=NA, colour="grey20") +
     geom_point(aes(fill = regime), size=5, shape=21, colour="grey",
                position=position_jitter(seed = 1)) +
     geom_text_repel(size = 2.5, aes(label = iso3c), col = "grey",
                position = position_jitter(seed = 1),
                hjust=-0.5, vjust=1) +
     theme_minimal() +
     ggtitle("图 5-17：独裁统治对女孩特别不利") +
     ylab("女孩与男孩受教育程度之比")+
     xlab("") +
     scale_fill_manual(values=c("#0000bf", "#bf0000")) +
     theme(plot.title = element_text(size = 8, face = "bold")) +
     theme(axis.title = element_text(size = 8, face = "bold")) +
     theme(legend.position = "none")
```

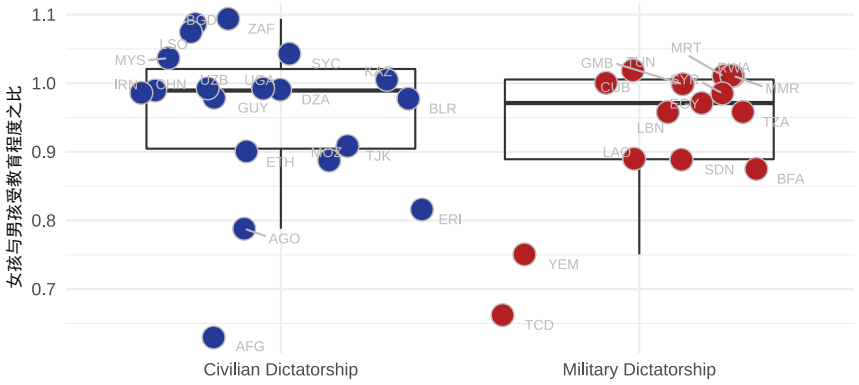

图 5-17：独裁统治对女孩特别不利

阿富汗（AFG）、安哥拉（AGO）、厄立特里亚（ERI）和莫桑比克（MOZ）代表表现欠佳的文官独裁政权。在军事独裁统治下，表现不佳的有布基纳法索（BFA）、老挝（LAO）、苏丹（SDN）、也门（YEM）和乍得（TCD）。

回顾一下，在双变量语境中，箱线图为每个类别的变量提供了 5 种不同的信息：第 25、第 50 和第 75 百分位数（箱）以及最小值和最大值（须）。标记箱线图可以获得额外的信息。因此，一系列并排的箱线图使我们更容易找出重要的差异或相似性。

然而，在极端情况下箱线图就没那么有用了。假设我对受访者对穆斯林和奥巴马的看法或“感觉”之间的关系感兴趣。为了说明，我首先将 ftobama 变量从数值变量转换为因子类型，这会产生 100 个不同类别的因子（见代码块 5-21）。然后为每个级别（100 个）的奥巴马情感量表绘制一个箱线图（见图 5-18）。虽然我们可以观察到正相关的趋势，但由于箱线图的数量太多了，很难发现什么具体的特征。而这么做也是有问题的，因为在奥巴马情感量表的某些级别上可能只有几个观测。鉴于箱线图有 5 个具体特征（两个须、两个枢和一个中位数），那些只有不到 5 个观测的类别是有问题的。在这些情况下，无法形成完整的箱线图。

代码块 5-21

```
nes$f.obama <- as.factor(nes$ftobama)

ggplot(nes, aes(f.obama, ftmuslim)) +
  geom_boxplot(col="#0000bf") +
  theme_minimal() +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) +
  ggtitle("图 5-18：类别过多") +
  ylab("穆斯林情感量表") +
  xlab("奥巴马情感量表") +
  theme(plot.title = element_text(size = 8, face = "bold")) +
  theme(axis.title = element_text(size = 8, face = "bold"))
```

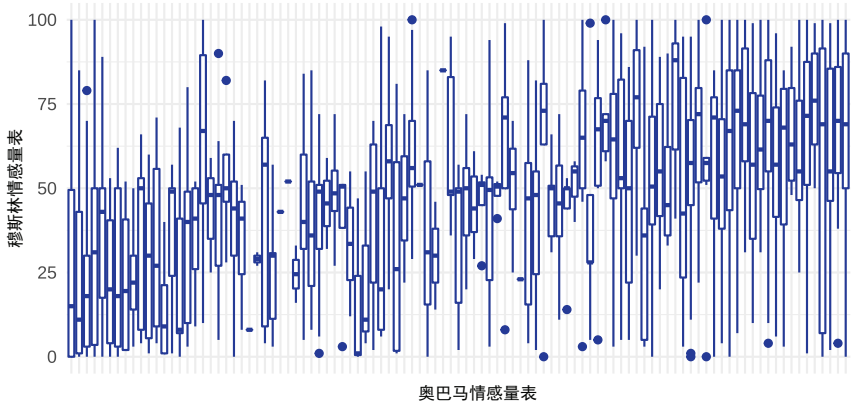

图 5-18：类别过多

如前所述，在比较连续变量和分类变量时，可以使用箱线图。我们看到了两个用例：①对同性恋的看法和政党认同之间的关系；②女孩与男孩受教育程度之比和政权类型的关系。然而，很多时候我们想研究的是两个分类变量之间的关系。在研究两个分类变量之间的关系时，使用马赛克图和交叉表。

第 5 章中“交叉表”下的案例

另一个例子也说明了与交叉表相比，马赛克图更加直观的性质。我们来提问，政党认同与受访者是否担心恐怖主义之间是否存在关系。在代码块 5-26 中，通过创建新的变量 var2 去掉了 terror_worry 变量中的类别“未询问”（Not asked）。

代码块 5-26

```
nes$var2 <- droplevels(nes$terror_worry, "Not asked")
```

现在我们准备好创建马赛克图了。在代码块 5-27 中，我让 R 取出两个变量都不含缺失值（NA）的案例子集。

代码块 5-27

```
ggplot(data = subset(nes, pid3.new!="NA" & var2 !="NA")) +
      geom_mosaic(aes(x = product(var2, pid3.new),
                      fill=var2,
                      na.rm=TRUE)) +
      xlab("") +
      ylab("") +
      ggtitle("图 5-20：民主党人不太担心恐怖主义") +
      theme_minimal() +
      scale_fill_brewer(palette="Blues") +
      theme(legend.position = "none") +
      theme(plot.title = element_text(size = 8, face = "bold")) +
      theme(axis.title = element_text(size = 8, face = "bold"))
```

回顾一下，国家选举研究的调查对象被问及他们对恐怖主义的担心程度。他们可以在从“完全不担心”（Not at all worried）到“非常担心”（Extremely worried）的答案中进行选择。是民主党人、无党派人士还是共和党人更担心恐怖主义？由于比较的是两个分类变量，因此马赛克图符合我们的需求（见图 5-20）。

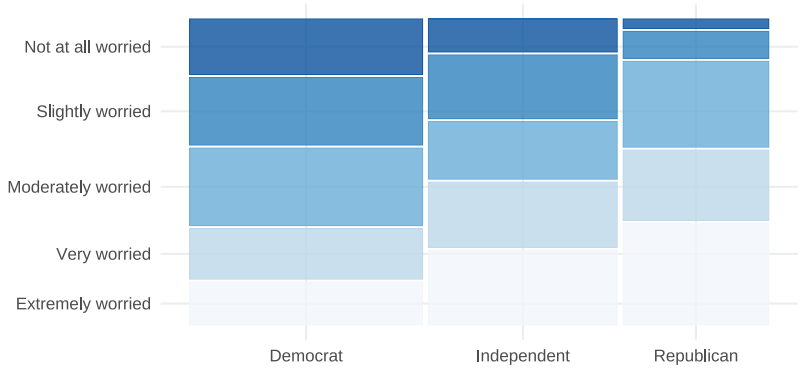

图 5-20：民主党人不太担心恐怖主义

图 5-20 中出现了一种有趣的模式。随着意识形态从自由主义转向保守主义，担心程度呈系统性上升。更准确地说，民主党人最不担心恐怖袭击，其次是无党派人士，最后是共和党人。马赛克图凸显了这一点。它还通过列的粗细霍然展示了民主党人、无党派人士和共和党人的相对规模（人数）在参与调查的人中所占的比例。共和党人更担心下一次恐怖袭击，这是一个精彩的结论。虽然我不会进一步探讨，但这个结果确实引出了一个问题：为什么？

现在再看看用交叉表展示的相同信息（见代码块 5-28 和表 5-4）。

代码块 5-28

```
CrossTable(nes$var2, nes$pid3.new,
           main=" 表 5-4：对恐怖主义的担心与政党认同的交叉表 ",
           prop.chisq=FALSE)
```

表 5-4：对恐怖主义的担心与政党认同的交叉表

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

=============================================================================
                     nes$pid3.new
nes$var2             Democrat    Independent    Republican    Total
-----------------------------------------------------------------------------
Extremely worried           43             61            64      168
                         0.256          0.363         0.381    0.233
                         0.145          0.254         0.346
                         0.060          0.085         0.089
-----------------------------------------------------------------------------
Very worried                51             53            44      148
                         0.345          0.358         0.297    0.205
                         0.172          0.221         0.238
                         0.071          0.074         0.061
-----------------------------------------------------------------------------
Moderately worried          78             47            54      179
                         0.436          0.263         0.302    0.248
                         0.264          0.196         0.292
                         0.108          0.065         0.075
-----------------------------------------------------------------------------
Slightly worried            68             52            17      137
                         0.496          0.380         0.124    0.190
                         0.230          0.217         0.092
```

```
                         0.094         0.072        0.024
-------------------------------------------------------------------
Not at all worried          56            27            6       89
                         0.629         0.303        0.067    0.123
                         0.189         0.112        0.032
                         0.078         0.037        0.008
-------------------------------------------------------------------
Total                      296           240          185      721
                         0.411         0.333        0.257
===================================================================
```

尽管交叉表提供了每个类别的确切数字，但整体信息更难辨别。

第 12 章中“二元回归的例子”下的案例

多族裔国家是否更暴力

因变量经过了对数变换

这里的因变量是凶杀率，即每个国家每 10 万人口中凶杀案受害者数量。由于原始形式的变量是正偏态的，我们取 `log(homicide+1)`。通过加 1，我们在对数中加入了常数。请注意代码中使用的是凶杀案受害者数量变量，所以没有零值[1]。如果对 0 取对数，结果是未定义的。未定义的个案会从回归中删除，并导致错误，使 R 无法进行所需的计算。

自变量是民族语言碎片化程度（通常被称为 ELF 分数），它表明随机选择的两个人说相同语言或有相同宗教信仰的可能性有多大。标度的范围从 0 到 1，1 表示语言极为多样化。请注意，在代码块 12-9 中，我将 ELF 分数变换为 0 和 100 之间的范围，以便更容易解读。当因变量为对数时，ELF 分数每增加 1 个单位，每 10 万人口中的凶杀案受害者数量比例就会相应地变化 $\left(e^{(\text{beta})}-1\right)\times 100$。表 12-4 中的估计表明 beta 为 0.009。因此，ELF 分数每增加 1 点，相应的每 10 万人口中凶杀案受害者数量就会变化 0.9%。

请注意，在代码块 12-9 中，变量是在 `lm()` 函数中进行的变换。这是一个非常有用的特性，因为这样可以省去额外创建新变量的步骤。此处还使用了不同的方法来标记一些选定的案例。这里没有在 `ifelse()` 函数中使用竖条（|）运算符，而是使用 `%in%` 运算符指定了一个国家列表：美国（USA）、洪都拉斯（HND）、阿富汗（AFG）和德国（DEU）。如果标签变量（`world$iso3c`）满足条件，将根据 `world$country` 将其国家名称标记出来。

代码块 12-9

```
world$elf <- world$ethfrac * 100

murder.lm <- lm(log(homicide + 1) ~ elf, data = world)

stargazer(murder.lm, type = "text", title = "表 12-4:民族异质性和暴力有关",
          header = FALSE)

ggplot(world, aes(elf, log(homicide + 1))) +
    geom_point(col = "#bf0000") +
    geom_text_repel(size = 3, aes(label =
                    ifelse(world$iso3c %in% c("USA", "HND", "AFG", "DEU"),
```

1 如果原始数据是偏态的，但范围在 0 和 1 之间，我们可以用 0.001 作为常数，而不是 1。

```
                        as.character(world$country), ""),
                        size = 1, hjust = 0, vjust = 1.25), col = "grey",
                        show.legend = FALSE) +
  ggtitle("图 12-7：多样化的人口刺激凶杀案的发生") +
  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("每 10 万人口中凶杀案受害者数量") +
  xlab("民族语言碎片化")
```

表 12-4：民族异质性和暴力有关

```
================================================
                          Dependent variable:
                      ---------------------------
                           log(homicide + 1)
------------------------------------------------
elf                              0.009**
                                 (0.004)

Constant                        1.223***
                                 (0.168)

------------------------------------------------
Observations                       122
R2                                0.046
Adjusted R2                       0.038
Residual Std. Error         0.974 (df = 120)
F Statistic              5.835** (df = 1; 120)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

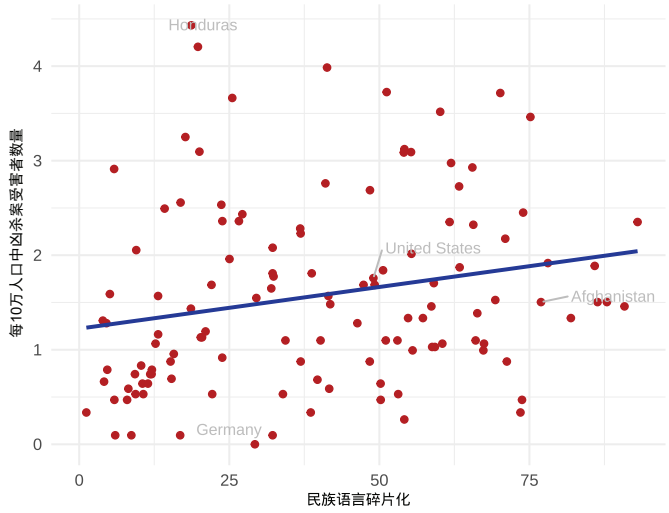

图 12-7：多样化的人口刺激凶杀案的发生

该估计的 *t*- 比率为 2.15，具有统计显著性：当关系实际上为 0 时，得到系数 0.009 的可能性相当小。R^2 统计量表明这解释了 5% 的变化（准确地说是 0.046）。因此，虽然 ELF 分数与凶杀案受害者数量之间的关系具有统计显著性，但我们确实无法解释世界各地凶杀率的大部分差异。这显然还有很多工作要做。

第 14 章中“二元虚拟变量回归”下的案例

让我们从一个简单的模型开始，并在此基础上构建模型。假设我们感兴趣的是美国人对穆斯林的看法。具体而言，我们使用 NES 调查中对穆斯林的看法的情感量表——从 0 到 100 的连续变量。我们的问题是，所有受访者的情感量表平均分数是多少，以及男性与女性之间是否存在差异？

$$Y = a_1 + e$$
$$Y = a_1 + b_1\ (\text{Gender}) + e$$

第一个方程式，截距 a_1 给出了所有受访者（男性和女性）Y 的平均数。第二个方程式，加上了虚拟变量。我们给女性赋值为 1，男性赋值为 0。在本例中，由于男性被赋值为 0，所以男性是**参考类别**[1]（**reference category**）。由于虚拟变量在受访者是女性时记录为 1，是男性时记录为 0，所以当受访者是男性时，b_1 就会丢失——b_1 乘以 0 等于 0，留下 a_1。因此，当受访者是男性时，a_1 项提供了 Y 的平均数。当受访者是女性时，性别变量记录为 1。因此，要计算女性 Y 的平均数，用 $a_1 + b_1$ 即可。请注意，系数 b_1 带来了男性和女性对穆斯林的看法差异。意外的收获是，回归分析还表明了男性和女性之间的差异（b_1）是否具有统计显著性。请看下面相应的回归表（见表 14-1）。

代码块 14-1 中的代码看起来应该很熟悉。我定义了两个线性模型，其中一个包含常数 1，另一个包含虚拟变量 `nes$gender`。然后使用 `stargazer()` 函数展示两个回归的结果。

代码块 14-1

```
genmus1 <- lm(ftmuslim ~ 1, data=nes)
genmus2 <- lm(ftmuslim ~ gender, data=nes)

stargazer(genmus1, genmus2, type = "text",
          title = "表 14-1：女性对穆斯林的看法更积极",
          header=FALSE)
```

表 14-1：女性对穆斯林的看法更积极

```
==========================================================
                              Dependent variable:
                    ------------------------------------
                                   ftmuslim
                           (1)                  (2)
----------------------------------------------------------
genderFemale                                   5.892***
```

1 虚拟变量中被赋值为零的类别。

```
                                                    (1.749)

Constant                        45.439***           42.383***
                                 (0.877)             (1.259)

-------------------------------------------------------------------
Observations                      1,178               1,178
R2                                0.000               0.010
Adjusted R2                       0.000               0.009
Residual Std. Error 30.118 (df = 1177)    29.986 (df = 1176)
F Statistic                              11.355*** (df = 1; 1176)
===================================================================
Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

首先，请注意性别变量为 genderFemale，表明性别的女性类别被设置为 1。这是 R 自动生成的：赋值为 1 的类别名称会附在变量名称的后面。男性和女性之间的差异（b_1）是 5.89。平均而言，对穆斯林的看法女性比男性高 5.9 分。差异具有统计显著性：我们有 99.9% 的信心认为男性与女性之间的差异不为零。为了完整起见，模型 2 中的常数给出了男性（参考类别）的穆斯林情感量表平均分数。因此，在回归中我们发现，男性的平均分数是 42.38，而男性和女性之间的差异是 5.89，这意味着女性的平均分数是 42.38+5.89=48.27。

第 14 章中“多元回归与虚拟变量”下的案例

在这里的例子中，在控制了可能解释对穆斯林看法的其他变量的情况下，我们考察了性别对穆斯林看法的影响。性别显然不是影响对穆斯林看法的唯一变量，多元回归分析让我们能够考虑到这一点，最终更准确地估计性别的影响。

当向模型中添加变量时，解读保持不变，同时还要加上约束短语——“保持其他一切不变”。让我们在二元模型中再添加一些控制因素：教育和政治意识形态。教育变量的范围是从 1 到 7，其中 1 表示没有高中文凭，7 表示有研究生学位。政治意识形态变量的范围也是从 1 到 7，其中 1 表示最自由，7 表示最保守。

为了说明问题，在代码块 14-2 中，这两个变量都被转换成了连续变量。表示为连续变量后，就将其加入线性模型中。

代码块 14-2

```
nes$educ.n <- as.numeric(nes$educ)
nes$pid7.n <- as.numeric(nes$pid7)

muslim.lm <- lm(ftmuslim ~ gender + educ.n + pid7.n, data=nes)
```

```
stargazer(muslim.lm,
          title = "表 14-2：性别影响对穆斯林的看法",
          header = FALSE, type="text")
```

表 14-2：性别影响对穆斯林的看法

```
================================================
                        Dependent variable:
                    ----------------------------
                               ftmuslim
------------------------------------------------
genderFemale                  5.039***
                               (1.607)

educ.n                        3.179***
                               (0.526)

pid7.n                        -4.978***
                               (0.358)

Constant                      51.058***
                               (2.526)

------------------------------------------------
Observations                    1,164
R2                              0.178
Adjusted R2                     0.176
Residual Std. Error      27.335 (df = 1160)
F Statistic          83.873*** (df = 3; 1160)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

尽管政治意识形态和教育都与对穆斯林的看法密切相关，但将它们纳入模型中并没有显著改变性别虚拟变量的系数。在控制了政治意识形态和教育的情况下，性别与对穆斯林的看法的联系似乎具有统计显著性。

在继续下一个例子之前，让我们考虑一下，如果在回归中再添加一个虚拟变量会怎样？不使用政治意识形态作为连续变量，而是使用 `nes$partyid3` 将受访者分为民主党人、无党派人士和共和党人。将具有三个水平的分类变量加入模型中时，R 会自动创建两个虚拟变量（本例中，一个是无党派人士，另一个是共和党人）。接下来，我们将讨论得到的方程式和回归输出结果。

数据可视化的艺术与实践
R 中的虚拟变量

在 R 中，将分类变量加入回归中时，程序会创建 n–1 个虚拟变量，其中 n 等于变量中类别的数量。未包含在回归中的类别被称为参考类别，由常数体现。

本例中，省略的类别或者说参考类别是“民主党人”，因为该类别并未作为虚拟变量包含在回归中。因此，回归输出结果中有性别变量和另外两个虚拟变量：一个是共和党人，另一个是无党派人士。在第三个方程式中，当受访者是民主党人（被省略的类别或参考类别）时，b_2 和 b_3 将为 0。因此，回归中的常数给出了那些认为自己是民主党人的男性的穆斯林情感量表的平均分数——记住，男性是性别虚拟变量中被省略的类别。民主党人和男性是两个参考类别。

$$Y = a_1 + e$$
$$Y = a_1 + b_1\ (\text{Gender}) + e$$
$$Y = a_1 + b_1\ (\text{Gender}) + b_2\ (\text{Independent}) + b_3\ (\text{Republican}) + e$$

在将政党认同变量添加到回归中之前，我们有必要整理一下（见代码块 14-3）。首先，我们将“其他（Other）”和“不确定（Not sure）”的回答转换为“NA”，因为想把样本限制为回答民主党人、共和党人或无党派人士的人。我们使用 `ifelse()` 函数完成了此项操作。把 R 代码翻译出来就是，如果 `nes$pid3` 等于“其他（Other）”，则将其赋值为“NA”；如果 `nes$pid3` 等于“不确定（Not sure）”，则将其赋值为“NA”；否则使用 `nes$pid3` 原来的值。然后，将其设置为因子，按民主党人、无党派人士、共和党人的顺序标注因子水平。

代码块 14–3

```
nes$pid3.new <- ifelse(nes$pid3 == "Other", NA,
   ifelse(nes$pid3 %in% c("Not Sure", "Not sure"), NA, nes$pid3))

nes$pid3.new <- as.factor(nes$pid3.new)

levels(nes$pid3.new)=c("Democrat", "Republican", "Independent")

nes$pid3.new = factor(nes$pid3.new,levels(nes$pid3.new)[c(1,3,2)])
```

现在，在准确指定了政党认同变量之后，我们就准备好将其纳入回归了。为了说明该变量所造成的差异，我们先定义一个不包含它的回归（`muslim3.lm`），再定义一个包含它的回归（`muslim4.lm`）。然后，将这两个模型放在 `stargazer()` 函

数中（见代码块 14-4）。

代码块 14-4

```
muslim3.lm <- lm(ftmuslim ~ gender, data=nes)
muslim4.lm <- lm(ftmuslim ~ gender + pid3.new, data=nes)

stargazer(muslim3.lm, muslim4.lm, type = "text",
          title = "表 14-3：对穆斯林的看法", header = FALSE)
```

表 14-3：对穆斯林的看法

```
==========================================================================
                                  Dependent variable:
                      ---------------------------------------------
                                         ftmuslim
                                (1)                      (2)
--------------------------------------------------------------------------
genderFemale                 5.892***                 4.345**
                             (1.749)                  (1.718)

pid3.newIndependent                                   -12.953***
                                                      (1.991)

pid3.newRepublican                                    -25.449***
                                                      (2.163)

Constant                     42.383***                54.391***
                             (1.259)                  (1.668)

--------------------------------------------------------------------------
Observations                  1,178                    1,100
R2                            0.010                    0.122
Adjusted R2                   0.009                    0.119
Residual Std. Error     29.986 (df = 1176)       28.253 (df = 1096)
F Statistic         11.355*** (df = 1; 1176) 50.590*** (df = 3; 1096)
==========================================================================
Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

本例中，在保持性别不变的情况下，共和党人录得的情感量表分值比民主党人少 25.45 分，无党派人士则比民主党人少 12.95 分。