第 5 章　R 的变量相关性分析和统计图形

数据基本分析起步于单个变量的描述统计，进一步，本章将关注两个或多个变量间的相关性研究，旨在揭示变量取值间相互影响的特点和相互作用的程度。变量相关性分析有两个层面：第一，样本相关性层面，即视数据集为随机样本，选择恰当的描述统计量，刻画样本中两个变量间相关性的强弱；第二，总体相关性层面，即基于样本相关性对样本来自的总体相关性进行推断。从统计学的方法体系上看，第二个层面的研究属于推断统计的范畴。再有，不同类型变量间相关性分析的方法也有所不同，表现在所采用的描述统计量不同，可视化统计图形也不同。

本章将继续围绕第 4 章所列大数据分析案例中的问题，就变量相关性研究问题，基于样本相关性层面，分章节地分别对两分类型变量间的相关性、两数值型变量间的相关性以及案例进行讲解。

5.1　分类型变量相关性的分析

分类型变量相关性分析的研究对象是两个或多个分类型变量，主要研究目的是考察一个分类型变量的取值是否与另一个分类型变量的取值有关。

例如，美食餐馆食客评分案例中，考察是否存在某些区域仅经营某几种主打菜，其他主打菜则主要集中在除此之外的其他区域的情况。可将该研究视为探讨餐馆的分布区域(分类型变量)与主打菜(分类型变量)分布，即两分类型变量是否具有相关性的问题。

两分类型变量的相关性进一步可细分为三种情况：两类别型变量的相关性、两顺序型变量的相关性、一类别型与一顺序型变量的相关性。针对不同情况将采用不同的分析方法。

5.1.1　分类型变量相关性的描述

1．编制列联表

研究两分类型变量相关性的常见方法是编制列联表，且列联表均适用于上述三种情况。列联表中的内容一般包括：两分类型变量类别值交叉分组下的实际观测频数，表各行或各列的频数合计，各频数占所在行或列合计的百分比等。

例如，表 5-1 是基于美食餐馆食客评分数据编制的餐馆主打菜和餐馆区域分布的列联表。其中，每个单元格中的第一个数值为观测频数，第二和第三个数值为频数占所在行的百分比以及占所在列的百分比。

R 中编制列联表可采用表 3-6 中的 table 函数。table 函数不仅可编制两分类型变量的二维列联表，还可编制高维列联表。也可以调用 gmodels 包中的 CrossTable 函数。gmodels 包首次使用时须下载安装，并加载到 R 的工作空间中。将在后续的案例中说明列联表的具体含义。

表 5-1 餐馆主打菜和区域的列联表

			区域		
			北太平庄	五道口	合计
主打菜	川菜	频数	37	39	76
		区域百分比	48.7%	51.3%	100.0%
		主打菜百分比	19.5%	17.2%	18.2%
	淮扬菜	频数	3	0	3
		区域百分比	100.0%	.0%	100.0%
		主打菜百分比	1.6%	.0%	.7%
	火锅	频数	32	32	64
		区域百分比	50.0%	50.0%	100.0%
		主打菜百分比	16.8%	14.1%	15.3%
	咖啡厅	频数	18	28	46
		区域百分比	39.1%	60.9%	100.0%
		主打菜百分比	9.5%	12.3%	11.0%
	寿司/简餐	频数	3	9	12
		区域百分比	25.0%	75.0%	100.0%
		主打菜百分比	1.6%	4.0%	2.9%
	西式简餐	频数	5	19	24
		区域百分比	20.8%	79.2%	100.0%
		主打菜百分比	2.6%	8.4%	5.8%
	湘菜	频数	14	17	31
		区域百分比	45.2%	54.8%	100.0%
		主打菜百分比	7.4%	7.5%	7.4%
	小吃	频数	78	83	161
		区域百分比	48.4%	51.6%	100.0%
		主打菜百分比	41.1%	36.6%	38.6%
合计		频数	190	227	417
		区域百分比	45.6%	54.4%	100.0%
		主打菜百分比	100.0%	100.0%	100.0%

2．基于列联表卡方统计量的相关性描述

基于列联表可计算得到一个名为 Pearson 卡方的统计量。Pearson 卡方的定义：$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}^{o}-f_{ij}^{e})^2}{f_{ij}^{e}}$。其中，$r$ 为列联表的行数，c 为列联表的列数，f_{ij}^{o} 为列联表第 i 行第 j 列单元格的实际观测频数，f_{ij}^{e} 为列联表第 i 行第 j 列单元格的期望频数，体现了列联表中两变量不相关下的理论分布。

以表 5-1 为例，表中餐馆在北太平庄和五道口区域的分布依次是 45.6%和 54.4%。若假设不存在某些主打菜只在北太平庄区域的餐馆经营而很少在五道口区域的餐馆经营的现象，即主打菜与区域无关，则 76 家主打川菜的餐馆，理论上应有 76×45.6%=76×190/471=34.6 家餐馆在北太平庄区域，76×54.4%=76×227/471=41.4 家餐馆在五道口区域。这两个数值依次是第一行两个单元格的期望频数。其他单元格均可同理计算出各自的期望频数 f_{ij}^{e}。

进一步，计算 Pearson 卡方统计量。可见，Pearson 卡方统计量的值越大，说明整体上实际观测频数和期望频数差距越大。因期望频数体现的是列联表中两变量不相关下的理论分

布，所以卡方值越大，表明列联表中两变量不相关的可能性越小，两变量越可能具有一定的相关性。

再进一步，Pearson 卡方统计量的数值受到样本量和列联表单元格数目的影响，应剔除这些因素，为此得到以下刻画列联表中两分类型变量相关性的系数。

(1) phi 系数

phi 系数适用于如表 5-2 所示的 2×2(2 行 2 列)的列联表。

表 5-2　2×2 列联表

	类 1	类 2	合　计
类 1	A_{11}	A_{12}	R_1
类 2	A_{21}	A_{22}	R_2
合计	C_1	C_2	n

在 Pearson 卡方统计量基础上，phi 系数定义为 $\phi=\sqrt{\frac{\chi^2}{n}}=\frac{A_{11}A_{22}-A_{12}A_{21}}{\sqrt{R_1R_2C_1C_2}}$，$n$ 为样本量。

若假设列联表中的两变量独立，由于 $\frac{A_{11}}{C_1}=\frac{A_{12}}{C_2}$，可知 $A_{11}A_{22}=A_{12}A_{21}$，代入 phi 系数，计算，有 $\phi=0$；若假设列联表中的两变量完全相关，且 $A_{12}=A_{21}=0$，有 $\phi=1$。或 $A_{11}=A_{22}=0$ 时，有 $\phi=-1$。因分类型变量的类别编码可以互换，所以 ϕ 前的正负符号没有意义。可见，ϕ 的绝对值越接近 1，表明两分类型变量的相关性越强；绝对值越接近 0，表明两分类型变量的相关性越弱。

(2) 列联(Contingency) 系数

列联系数适用于 2×2 以上的列联表。在 Pearson 卡方统计量基础上，列联系数定义为 $C=\sqrt{\frac{\chi^2}{\chi^2+n}}$，其取值范围在[0~1)之间。越接近 1，表明卡方值足够大而弱化了样本量在分母中的作用，两分类型变量的相关性越强；反之，越接近 0，表明卡方值非常小而强化了样本量在分母中的作用，两分类型变量的相关性越弱。

(3) Cramer's V 系数

在 Pearson 卡方统计量基础上，Cramer's V 系数定义为 $V=\sqrt{\frac{\chi^2}{n\min[(r-1)(c-1)]}}$，其中，min[(*r*–1)(*c*–1)]表示取(*r*–1)和(*c*–1)中的最小值(*r*、*c* 分别表示列联表的行数和列数)。

Cramer's V 系数在考虑样本量影响的同时，还兼顾到了列联表单元格数目的影响。在 2×2 的列联表中，V 系数与 phi 系数相等。可以证明，V 系数的取值在 0~1 之间，越接近 1 表明两分类型变量间的相关性越强。

可利用 vcd 包中的 assocstats 函数，计算上述基于卡方统计量的相关性度量统计量。基本书写格式：

assocstats(列联表对象)

3．等级相关系数

研究两分类型变量相关性的另一种常见方法是计算等级相关系数。该方法只适用于两顺

序型变量间的相关性研究。常见的等级相关系数有 Spearman 等级相关系数和 Kendall-τ等级相关系数。

(1) Spearman 等级相关系数

Spearman 等级相关系数用来度量两顺序型变量间的线性相关关系。首先，得到两变量中各数据的升序排序名次或称秩(排名并列时一般以各自顺序位次的均值为秩)，分别记作 U_i、V_i；然后，计算 D_i^2 和 $\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (U_i - V_i)^2$；最后，计算 Spearman 等级相关系数：$R = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$。

Spearman 等级相关系数并非直接基于顺序型变量的水平值，而是基于秩。

这里以分析学历水平和年薪水平两顺序型变量的相关性为例，说明 Spearman 等级相关系数的意义。通常学历水平和年薪水平呈正相关性，表现在学历水平的秩增大时，年薪水平的秩也会随之增大，反之学历水平秩减小时，年薪水平的秩也会随之减小。在这样的情况下，每个人的 D_i^2 值都会很小，$\sum D_i^2$ 也会很小。当学历水平和年薪水平呈完全正相关时：U_i=V_i，$\sum D_i^2 = 0$，Spearman 等级相关系数 R 等于 1。此外，当两变量呈完全负相关时：$\sum D_i^2 = \frac{1}{3} n(n^2-1)$，Spearman 等级相关系数 R 等于–1。可见，Spearman 等级相关系数的绝对值越接近 1，两变量的相关性越强。正号表示正相关，负号表示负相关。

Spearman 等级相关系数可延伸应用到一个数值型和一个顺序型变量的相关性研究上，须对数值型变量计算秩。

(2) Kendall-τ等级相关系数

Kendall-τ等级相关仍基于秩。首先，计算一致对数目(U)和非一致对数目(V)；然后，计算 Kendall-τ相关系数：$\tau = (U - V)\frac{2}{n(n-1)}$。

仍以分析学历水平和年薪水平两顺序型变量的相关性为例，说明 Kendall-τ等级相关系数的意义。若学历水平和年薪水平呈完全正相关性，学历水平的秩增大时，年薪水平的秩也随之增大；学历水平秩为升序时，年薪水平秩(d_i)也严格为升序，即 $d_j > d_i (j > i)$，则称其为一个一致对。若样本量为 n，则存在 $U = \frac{1}{2} n(n-1)$ 个一致对 $d_j > d_i (j > i)$，和 $V = 0$ 个非一致对 $d_j < d_i (j > i)$，此时 Kendall-τ相关系数等于 1。此外，当两变量呈完全负相关时，存在 $U = 0$ 个一致对和 $V = \frac{1}{2} n(n-1)$ 个非一致对，Kendall-τ相关系数等于–1。可见，Kendall-τ等级相关系数的绝对值越接近 1，两变量的相关性越强。正号表示正相关，负号表示负相关。

Kendall-τ等级相关系数也可延伸应用到一个数值型和一个顺序型变量的相关性研究上，须对数值型变量计算秩。

可利用 cor 函数计算等级相关系数。基本书写格式：

cor(矩阵或数据框列号, use=缺失值处理方式, method="spearman/kendall")

cor 函数用于计算指定列变量间的相关系数，返回结果为相关系数矩阵。其中：参数 use 用于指定计算相关系数时缺失值的处理方法，可取值为 all.obs，表示默认数据中不存在缺失值，如果存在则自动给出提示信息；everything 表示若某变量存在缺失值，则它与其他

变量相关系数为 NA，即不计算；complete.obs 表示采用删除法做缺失值处理后再计算相关系数；pairwise.complete.obs 表示采用成对删除法做缺失值处理后再计算相关系数。参数 method 用于指定相关系数类型，取 spearman 或 kendall 表示计算 Spearman 等级相关系数和 Kendall-τ等级相关系数。

5.1.2　分类型变量相关性的统计图形

常用的分类型变量相关性的基本统计图形为马赛克图，因图中格子的排列形似马赛克而得名。绘制马赛克图的 R 函数 mosaicplot 函数，基本书写格式：

mosaicplot（~分类型域名 1+分类型域名 2+…, data=数据框名）

其中，数据组织在指定数据框中，分类型域名应为因子。马赛克图的具体含义在应用案例中详细说明。

5.1.3　大数据分析案例：餐馆的区域分布与主打菜分布是否具有相关性

基于美食餐馆食客评分数据，首先，利用列联表、基于卡方统计量的相关性描述统计量以及马赛克图，研究餐馆的区域分布与主打菜分布是否具有相关性；其次，利用等级相关系数研究人气热度(顺序型变量)与人均消费金额(数值型变量)、平均打分(数值型变量)的相关性。具体代码和部分结果如下。

```
> MyData<-read.table(file="美食餐馆食客评分数据.txt",header=TRUE,sep=" ",
        stringsAsFactors=FALSE)
> (CrossT<-table(MyData$food_type,MyData$region))      #编制列联表
           北太平庄  五道口
  川菜           37     39
  淮扬菜          3      0
  火锅           32     32
  咖啡厅         18     28
  寿司/简餐       3      9
  西式简餐        5     19
  湘菜           14     17
  小吃           78     83
> addmargins(round(prop.table(CrossT,1)*100,2),2)      #计算行百分比
           北太平庄  五道口     Sum
  川菜        48.68  51.32  100.00
  淮扬菜     100.00   0.00  100.00
  火锅        50.00  50.00  100.00
  咖啡厅      39.13  60.87  100.00
  寿司/简餐   25.00  75.00  100.00
  西式简餐    20.83  79.17  100.00
  湘菜        45.16  54.84  100.00
  小吃        48.45  51.55  100.00
> library("vcd")
> assocstats(CrossT)                  #计算基于卡方的相关性描述统计量
                    X^2 df P(> X^2)
Likelihood Ratio 15.409  7 0.031103
Pearson          13.663  7 0.057502
```

```
Phi-Coefficient   : NA
Contingency Coeff.: 0.178
Cramer's V        : 0.181
> mosaicplot(~food_type+region,data=MyData,main="主打菜和区域的马赛克图")
> cor(MyData$heat,MyData$cost_avg,method="spearman")  #计算人均消费与热度的
                                                       Spearman 等级相关系数
[1] 0.3101741
> cor(MyData$heat,MyData$score_avg,method="kendall")   #计算平均打分与热度的
                                                       Kendall-τ等级相关系数
[1] 0.1189485
```

说明：

① 本例首先利用 table 函数编制列联表，之后利用 prop.table 和 addmargins 函数计算行百分比且合并到列联表中。

② 本例的 Pearson 卡方统计量等于 13.663，基于 Pearson 卡方计算的列联系数和 Cramer's V 系数分别等于 0.178 和 0.181，主打菜和区域变量的相关性很弱。

③ 图 5-1 为主打菜和区域的马赛克图。马赛克图的数据来自列联表。本例中，图的行向为区域，列向为主打菜，分别以矩形的高和宽体现餐馆数量的多少。可见，北太平庄和五道口地区均以经营小吃为主，餐馆数量都是最多的，其次均为川菜和火锅；淮扬菜仅在北太平庄地区有，五道口区域的简餐餐馆数量明显多于北太平庄区域，等等。

④ 计算人气热度与人均消费金额和平均打分的两种等级相关系数(R 可自动计算各变量的秩)，相关程度均不高，但相比之下，人气与人均消费金额的相关性略高些。

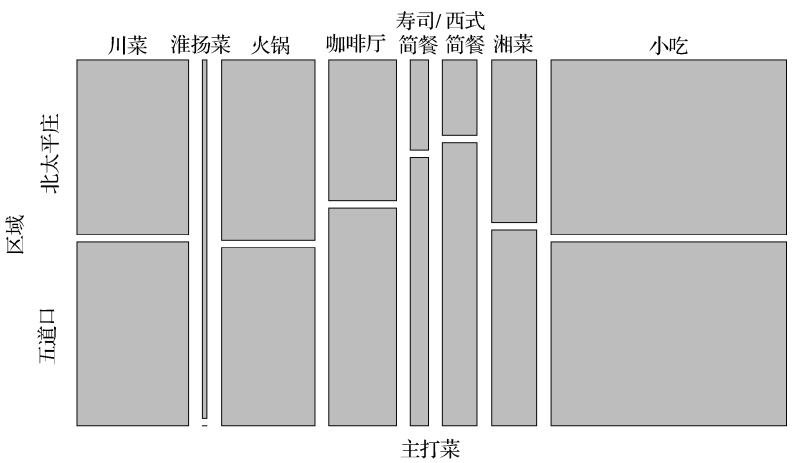


图 5-1 主打菜与区域的马赛克图

5.2 数值型变量相关性的分析

5.2.1 数值型变量相关性的描述

数值型变量相关性分析的研究对象是两个或多个数值型变量，主要研究目的是考察两个数值型变量取值的相关性强弱。

例如，美食餐馆食客评分案例中，考察美食餐馆的评分是否与人均消费金额相关，即为两数值型变量相关性研究问题。

可通过简单相关系数刻画两数值型变量的线性相关性。假设 x_i 和 y_i 分别为两数值型变量的变量值，有 n 个数据对，简单相关系数也称 pearson 相关系数，定义：

$$r=\frac{\sum_{1=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

简单相关系数的取值范围在–1～+1 之间，绝对值越接近 1，线性相关性越强；越接近 0，线性相关性越弱。正号表示正相关，负号表示负相关。

计算两数值型变量相关系数的 R 函数是 cor，基本书写格式：

cor(矩阵或数据框列号, use=缺失值处理方式, method="pearson")

参数含义详见 5.1.1 节。

5.2.2　数值型变量相关性的统计图形

散点图是展示两个或多个数值型变量相关性特征的最常用工具，包括简单散点图、三维散点图、气泡图、矩阵散点图等。以下仅列出各种图形的简单说明和实现的 R 函数，图形的具体含义将在案例中说明。

1．简单散点图

简单散点图将观测数据点绘制在一个二维平面中，通过数据点分布的形状可粗略展示两数值型变量间的相关性特点。若数据点大致分布在一条直线的周围，表示两变量具有一定的线性相关性。

绘制简单散点图的函数是 plot，基本书写格式：

plot(*x*=数值型向量名 1, y=数值型向量名 2)

或

plot(域名 2~域名 1, data=数据框名)

其中，数值型向量 1(域名 1)和数值型向量 2(域名 2)分别作为散点图的横坐标和纵坐标。第一种格式较为直接，容易理解；第二种格式采用了 R 公式的写法，“~”符号前的作为纵坐标，“~”符号后的作为横坐标，数据组织在 data 参数指定的数据框中。

2．三维散点图

三维散点图在展示两数值型变量相关性的同时，还希望体现第三个变量的取值状况。绘制三维散点图的函数是 scatterplot3d 包中的 scatterplot3d 函数，首次使用该包时应下载安装，并加载到 R 的工作空间中。scatterplot3d 的基本书写格式：

scatterplot3d(向量名 1, 向量名 2, 向量名 3)

其中，向量名 1、向量名 2、向量名 3 分别对应 *x* 轴、*y* 轴、*z* 轴的变量。

3．气泡图

三维散点图对第三个变量取值大小的体现并不十分清晰，对此可引入气泡图。气泡图即在绘制两个变量的散点图时，各个数据点的大小取决于第三个变量的取值。第三个变量取值不同，数据点的大小也就不同，形如大小不一的一组气泡。

绘制气泡图的函数是 symbols，基本书写格式：

symbols(向量名 1, 向量名 2, circle=向量名 3, inches=计量单位, fg=绘图颜色, bg=填充色)

其中，向量名 1，向量名 2 分别对应横坐标和纵坐标上的变量；气泡大小由参数 circle 指定的变量决定；参数 inches 指定气泡大小的计量单位，默认为英寸；参数 fg 指定绘制气泡的颜色；参数 bg 指定气泡的填充色。

4、矩阵散点图

矩阵散点图用于在一幅图上同时展示多对数值型变量的相关性。绘制矩阵散点图的函数是 pairs，基本书写格式：

pairs(~域名 1+域名 2+…+域名 *n*, data=数据框名)

其中，第一个参数是 R 公式的写法，表示分别对指定域两两绘制散点图，并集成在一幅图中。数据存放在 data 指定的数据框中。

5.2.3 大数据分析案例：餐馆各打分之间、打分与人均消费之间是否具有相关性

基于美食餐馆食客评分数据，对餐馆食客的各打分之间、打分与人均消费之间是否具有相关性进行研究。具体代码和执行结果如下。

```
> MyData<-read.table(file="美食餐馆食客评分数据.txt",header=TRUE,sep=" ",
    stringsAsFactors=FALSE)
> cor(MyData$taste,MyData$score_avg,method="pearson")    #计算简单相关系数
  [1] 0.699294
> par(mfrow=c(1,2),mar=c(6,4,4,4))                       #图形窗口布局
> plot(MyData$taste,MyData$score_avg,main="口味打分和平均打分的散点图",
    xlab="口味打分",ylab="平均打分",cex.main=0.8,cex.lab=0.8)
> plot(MyData$taste,MyData$score_avg,main="口味打分和平均打分的散点图",
    xlab="口味打分",ylab="平均打分",cex.main=0.8,cex.lab=0.8)
> M0<-lm(score_avg~taste,data=MyData)                    #线性回归分析
> abline(M0$coefficients,col=2)                          #添加回归直线
> M.Loess<-loess(score_avg~taste,data=MyData)            #局部加权散点平滑法拟合回归
> Ord<-order(MyData$taste)                               #按 x 轴取值排序后再绘图
> lines(MyData$taste[Ord],M.Loess$fitted[Ord],lwd=1,lty=2,col=3)
> library("scatterplot3d")
> par(mfrow=c(1,2),mar=c(6,4,4,4))
> with(MyData,scatterplot3d(taste,score_avg,cost_avg,main="美食餐馆口味
    打分、平均打分和人均消费的三维散点图", xlab="口味打分",ylab="平均打分",
    zlab="人均消费金额",cex.main=0.8,cex.lab=0.8,cex.axis=0.8))
> with(MyData,symbols(taste,score_avg,circle=cost_avg,inches=0.2,main=
    "美食餐馆口味打分、平均打分和人均消费的汽包图",xlab="口味打分",ylab="平均
```

```
      打分",cex.main=0.8,cex.lab=0.8,cex.axis=0.8,fg="white",bg= "lightblue"))
> pairs(~taste+environment+service+score_avg+cost_avg,data=MyData,main="
      美食餐馆打分和人均消费相关系数矩阵三点图")
> cor(MyData[,5:9])   #计算两两变量间的简单相关系数矩阵
              taste environment   service score_avg  cost_avg
taste       1.0000000   0.4828860 0.6931538 0.6992940 0.2175407
environment 0.4828860   1.0000000 0.8160090 0.4724386 0.4487347
service     0.6931538   0.8160090 1.0000000 0.6256902 0.3409103
score_avg   0.6992940   0.4724386 0.6256902 1.0000000 0.1771265
cost_avg    0.2175407   0.4487347 0.3409103 0.1771265 1.0000000
> library("corrgram")
> corrgram(MyData[,5:9],lower.panel=panel.shade,upper.panel=panel.pie, text.
      panel= panel.txt,main="美食餐馆打分和人均消费相关系数图")
> corrgram(MyData[,5:9],lower.panel=panel.ellipse,upper.panel=panel.pts,
      diag.panel=panel.minmax,main="美食餐馆打分和人均消费相关系数图")
```

说明：

(1)计算简单相关系数，绘制简单散点图

首先利用 cor 函数计算口味打分与平均打分的简单相关系数，近似为 0.7，具有中等强度正相关性。进一步绘制两者的简单散点图，如图 5-2(a)左图所示。其中，横坐标为口味打分，纵坐标为平均打分。直观上，两者存在一定的正相关性，与简单相关系数的计算结果一致。

(2)在简单散点图上添加回归线

为进一步刻画散点图所体现的两变量间的相关关系，可在散点图上添加回归线。为此，须经以下两步实现。

第一步，求解回归线。

求解回归线有以下两种主要方法。

- 一元线性回归法

一元线性回归法非常经典，但因相关理论较为复杂，将在第 8 章集中讨论，这里仅给出实现的 R 函数。一元线性回归法的函数是 lm 函数，基本书写格式：

lm(被解释变量名~解释变量名, data=数据框名)

其中，被解释变量是散点图中纵坐标对应的变量，解释变量是横坐标对应的变量。数据存储在 data 参数指定的数据框中。lm 函数的返回值是个列表，其中包括一个名为 coefficients 的成分，它是一个向量，存储了线性回归直线的截距和斜率。

- 局部加权散点平滑(LOcally WEighted Scatterplot Smoothing，LOWESS)法

局部加权散点平滑法属于非参数统计方法。不同于一般线性回归法中依赖全部观测数据建模，局部加权散点平滑法总是取一定比例的局部数据，在这部分子集中拟合多项式回归曲线，以展示数据的局部规律和趋势。若将散点图中点的局部范围从左往右依次推进，最终一条连续的曲线就可被平滑出来。当无法确定数据之间的相关性是否呈现或是否总是呈现出线性关系时，可采用这种方法得到回归线。R 中的 loess 函数是对局部加权散点平滑法 lowess 函数的修正，更为常用，基本书写格式：

loess(被解释变量名~解释变量名, data=数据框名)

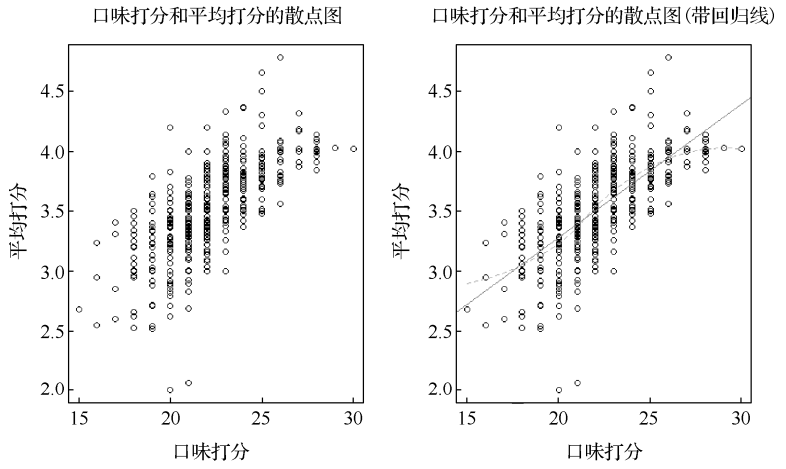
loess 函数的返回值也是个列表，其中名为 fitted 的成分存储了模型计算出的各观测被解释变量的预测值。

第二步，利用 abline 函数将回归线添加到已有的散点图上。

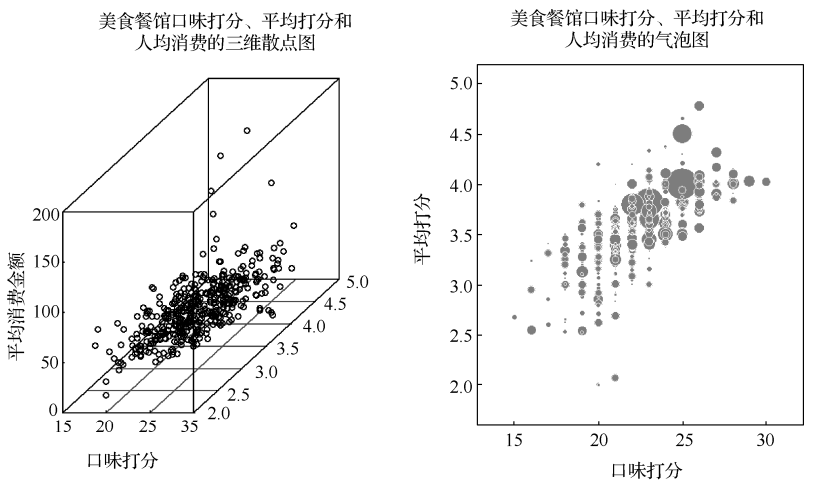
本例结果如图 5-2(a)右图所示。其中直线(实线)为一元线性回归法所得的回归直线，虚线为局部加权散点平滑法所得的回归线。

(3)绘制三维散点图和气泡图

绘制口味打分、平均打分与人均消费金额的三维散点图，如图 5-2(b)左图所示。绘制该图的目的是考察平均打分随口味打分变化的同时，人均消费金额是否也存在某种增大或减小的特点。例如，是否存在打分较高的同时人均消费金额较低的物美价廉现象，图 5-2(b)左图显然并不直观。为此绘制更直观的气泡图。本例中气泡越大表示人均消费金额越高，如图 5-2(b)右图所示。可见，评分较低的餐馆人均消费金额大多偏低，在高评分处也有部分餐馆的人均消费金额较低，高消费的餐馆评分较高。人均消费金额和评分间的线性关系不明显。

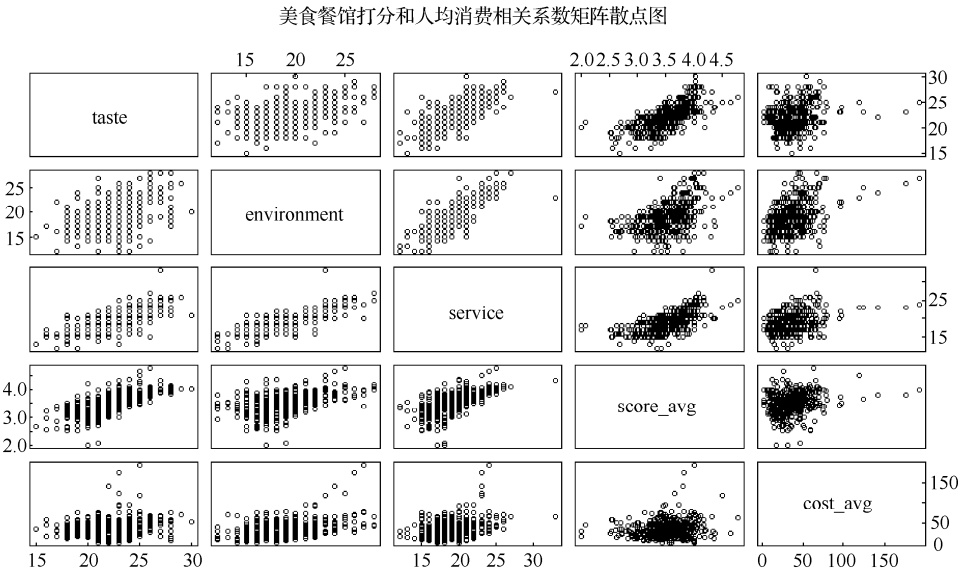


(a) 口味打分与平均打分的简单散点图

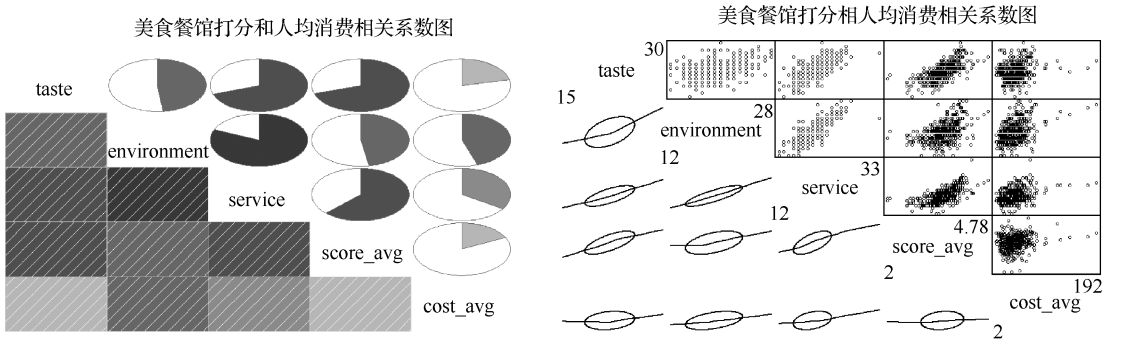


(b) 评分和人均消费金额的三维散点图和气泡图

图 5-2 评分数据的散点图和气泡图



(c) 评分和人均消费金额的矩阵散点图



(d)各打分及人均消费的相关系数图

图 5-2　评分数据的散点图和气泡图(续)

(4) 绘制矩阵散点图

利用矩阵散点图展示打分和人均消费金额两两变量间的相关性强弱，如图 5-2(c)所示，图中对角线单元格中列出了各变量的变量名。可见，就餐环境打分(environment)与服务质量打分(service)的相关性是最强的，人均消费金额(cost_avg)与各打分间的相关性均较弱。

(5) 简单相关系数和相关系数图

为进一步刻画各打分之间以及打分与人均消费金额之间线性相关性的强弱，利用 cor 函数计算简单相关系数矩阵。计算结果表明：就餐环境打分与服务质量打分的简单相关系数为 0.82，具有较最强的正相关。人均消费金额与各打分的简单相关系数均小于 0.5，呈弱相关。

相关系数矩阵虽然可以准确反映两两变量的线性相关性的强弱，但当这个矩阵较大时，分析起来就不太直观。为此，可基于相关系数矩阵绘制相关系数图。如图 5-2(d)所示。

相关系数图由下三角区域、上三角区域、对角区域三部分组成。区域在这里称为面板，三个区域也分别称为下面板、上面板和对角面板。除对角面板外，上下面板以不同形式直观展示相应变量对的相关性强弱。

在图 5-2(d)左边的相关系数图中，下面板通过阴影颜色的深浅表示相关性的强弱。同时，阴影中的斜线，若呈左下至右上，则表示正相关；若呈左上至右下，则表示负相关。上面板以饼图的填充比例展示相关系数的大小。对角面板没有其他信息，仅为变量名。

在右侧的相关系数图中，下面板通过椭圆大致描绘散点图的外围轮廓，中间的红色曲线是采用局部加权散点平滑拟合的回归线。上面板是散点图。对角面板不仅显示变量名，同时显示变量取值的最小值和最大值。

绘制相关系数图的函数是 corrgram 包中的 corrgram 函数，首次使用时应下载安装，并加载到 R 的工作空间中。corrgram 函数的基本书写格式：

corrgram(矩阵或数据框列, lower.panel=面板样式, upper.panel=面板样式,
text.panel=面板样式, diag.panel=面板样式)

其中，lower.panel, upper.panel 分别为下面板和上面板。text.panel 和 diag.panel 均属于对角面板。面板样式中对角面板取值：panel.minmax 表示显示变量的最小值和最大值；panel.text 表示显示变量名。上面板和下面板取值：NULL 表示空白，不显示任何内容；panel.pie 表示显示饼图；panel.shade 表示显示阴影；panel.ellipse 表示显示椭圆等；panel.pts 表示显示散点图。

5.3 大数据分析案例综合：北京市空气质量监测数据的相关性分析

通常认为空气中 PM2.5 的浓度与 CO 和 NO_2 的浓度等有比较密切的关系。现基于北京市 2016 年供暖季空气质量监测数据，利用变量相关性分析的统计图形，对影响 PM2.5 的因素进行直观的初步研究。此外，着重讲解高密度散点图的处理方法以及如何绘制分组散点图。代码和执行结果如下。

```
> MyData<-read.table(file="空气质量.txt",header=TRUE,sep=" ",
      stringsAsFactors=FALSE)
> Data<-subset(MyData,(MyData$date<=20160315|MyData$date>=20161115))
                                        #仅分析供暖季数据
> Data<-na.omit(Data)                   #获得完整观测
> par(mfrow=c(2,2),mar=c(6,4,4,4))
> plot(PM2.5~CO,data=Data,main="PM2.5 和 CO 浓度散点图",xlab="CO", ylab=
      "PM2.5",cex.main=0.8,cex.lab=0.8)
> plot(jitter(PM2.5,factor=1)~jitter(CO,factor=1.5),data=Data,main= "PM2.5
      和CO高密度处理散点图",xlab="CO",ylab="PM2.5",cex.main=0.8, cex.lab=0.8)
> smoothScatter(x=Data$CO,y=Data$PM2.5,main="PM2.5 和 CO 高密度处理散点图",
      xlab="CO",ylab="PM2.5",cex.main=0.8,cex.lab=0.8)
> library("scatterplot3d")
> with(Data,symbols(CO,PM2.5,circle=NO2,inches=0.2,main="CO 、 PM2.5 和
      NO2 浓度气泡图",xlab="CO",ylab="PM2.5",cex.main=0.8,cex.lab=0.8,
      cex.axis=0.8,fg="white",bg="lightblue"))
> Data$SiteTypes<-as.factor(Data$SiteTypes)   #用户自定义函数的功能：一元线
                                               性回归并添加回归直线
> Mypanel.lm<-function(x,y,...){
+  Tmp<-lm(y~x)
```

```
+  abline(Tmp$coefficients,col=2)
+  points(x,y,pch=1)}
> coplot(PM2.5~CO|SiteTypes,panel=Mypanel.lm,data=Data,pch=1,xlab="CO",
      ylab="PM2.5")
```

说明：

① 利用 plot 函数绘制 PM2.5 与 CO 的简单散点图，如图 5-3(a)左上图所示。

因样本量较大并有较多数据点叠加在散点图中，所以图 5-3(a)左上图所示为一种高密度的散点图。显然，高密度散点图不利于展示 PM2.5 与 CO 的相关性特征，为此可做如下两种处理。

第一，增加数据“噪声”，减少数据点的重叠。

为尽可能使数据点不完全叠加，可人为对数据增加“噪声”，即在原变量值上加极小的噪声值。一方面，尽管数据中添加了噪声，但因其极小，并不影响或改变变量间的相关性；另一方面，绘制散点图时，数据点的重叠程度会因噪声的存在得到一定程度的降低。

增加噪声的函数是 jitter，基本书写格式：

jitter(数值型向量, factor = n)

其中，参数 factor 称为扩充因子，n 默认为 1。设原变量值为 x，噪声为 b，添加噪声后的变量值为 $x+b$。噪声 b 是来自均匀分布的一个随机数，该均匀分布的取值范围是(–a,a)，且 a=factor · d/5, d=|x–x 的最近邻|。本例增加噪声后的散点图如图 5-3(a)右上图所示。修正效果不明显，可适度增大扩充因子值 n。

第二，利用色差突出散点图中的数据密集区域，明晰散点图的整体轮廓。

可使用 smoothScatter 函数绘制散点图，基本书写格式：

smoothScatter(x=横坐标向量, y=纵坐标向量)

该函数自动将一定范围内的数据点并为一组，称为分箱。最终数据点将被分成若干个组(箱)。用颜色的深浅表示组(箱)中数据点的多少，如图 5-3(a)左下图所示。

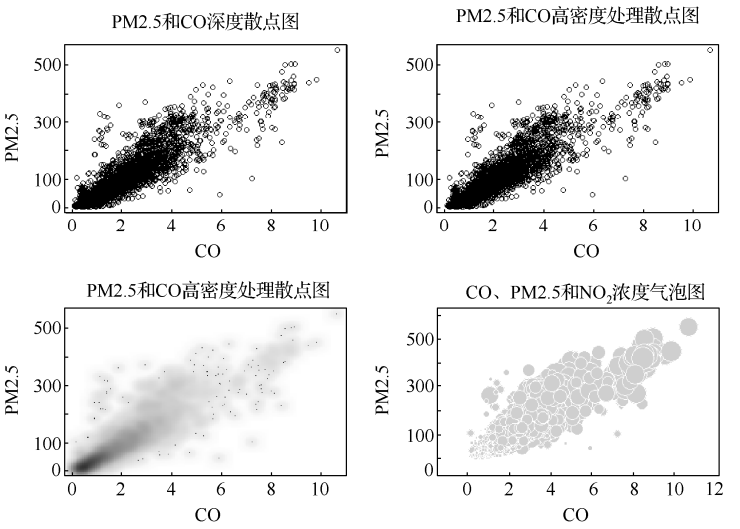
② 绘制 PM2.5、CO、NO_2 的气泡图，如图 5-3(a)右下图所示。图形比较清晰地展示了三者之间的关系：随着 CO 浓度的增加，PM2.5 浓度呈明显的线性增加，同时表示 NO_2 浓度高低的气泡也随之增大。

③ 进一步，基于对不同类型监测点的 PM2.5 和 CO 浓度数据，绘制分组散点图，并添加回归直线，如图 5-3(b)所示。

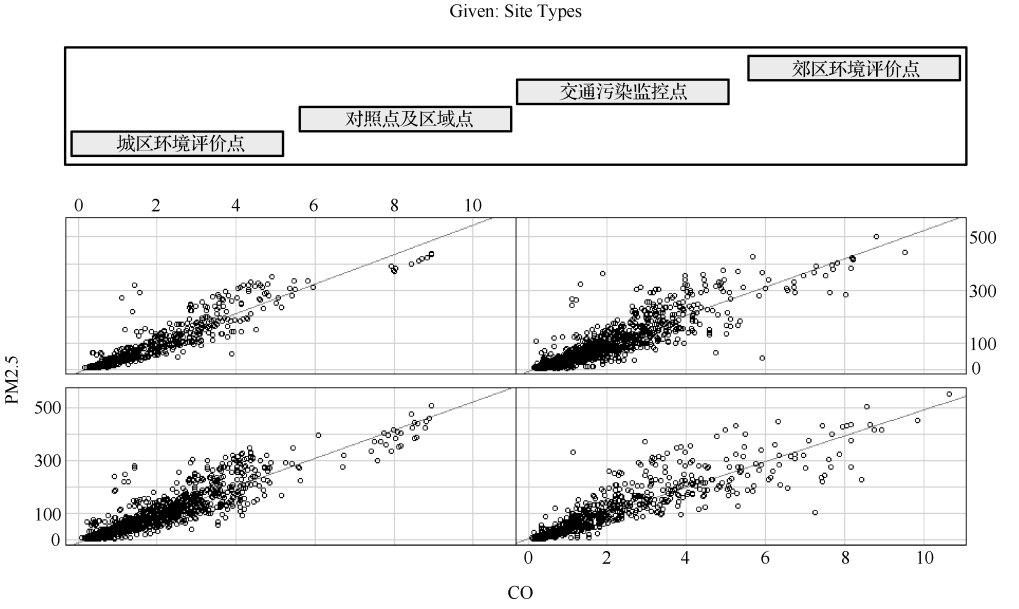
分组散点图也称协同图，用于展示两数值型变量之间的相关性在不同样本组上的差异。可采用 coplot 函数绘图，基本书写格式：

coplot(域名 1~域名 2|分组域名, number=分组数, data=数据框名)

其中，域名 1 和域名 2 分别作为散点图的纵坐标和横坐标，是 R 公式的写法；“|”后跟分组域名，其对应的变量通常是分类型变量(因子)，有时也可以是数值型变量。当分组变量为数值型时，须通过参数 number 指定将数值型变量分成几个有重叠的组。如果省略 number 参数，则默认分成 6 组；数据存储在参数 data 指定的数据框中。该函数首先依据指定的分组变量或分组后的数值型变量，将观测分成若干组，之后分别绘制各个组的散点图。



(a) PM2.5和CO的散点图及NO_2的气泡图



(b) PM2.5与CO的分组散点图

图 5-3　PM2.5 与 CO 的散点图与气泡图

图 5-3(b)中，最上边给出了各监测点类型。图最下一行从左往右从下往上各单元格(也称面板)的散点图，依次是监测点类型分别是城市环境评价点、对照点及区域点、交通污染监控点、郊区环境评价点的 PM2.5 与 CO 浓度的散点图。为在图中添加回归直线，定义面板函数为一个用户自定义函数。

5.4　本章涉及的 R 函数

本章涉及的 R 函数如表 5-3 所示。

表 5-3　本章涉及的 R 函数列表

函　数　名	功　　能
cor(矩阵或数据框列号, use=缺失值处理方式, method=相关系数类型)	计算指定两变量的相关系数
assocstats(列联表对象)	基于 Pearson 卡方统计量计算两分类型变量的相关性
mosaicplot (~分类型域名 1+分类型域名 2+…, data=数据框名)	绘制马赛克图
scatterplot3d(向量名 1,向量名 2,向量名 3)	绘制三维散点图
symbols(向量名 1,向量名 2 circle=向量名 3,inches=计量单位, fg=绘图颜色,bg=填充色)	绘制气泡图
pairs(~域名 1+域名 2+…+域名 *n*, data=数据框名)	绘制矩阵散点图
lm(被解释变量名~解释变量名, data=数据框名)	建立线性回归模型
loess(被解释变量名~解释变量名, data=数据框名)	局部加权散点平滑法
corrgram(矩阵或数据框列, lower.panel=面板样式, upper.panel=面板样式, text.panel=面板样式, diag.panel=面板样式)	绘制相关系数图
jitter(数值型向量,factor = *n*)	添加噪声数据
smoothScatter(*x*=横坐标向量, *y*=纵坐标向量)	高密度散点图处理
coplot(域名 1~域名 2|分组域名, number=分组数, data=数据框名)	绘制分组散点图