第十三章 预测销售额

# 背景

一个连锁餐馆的经营者想要发现一个新的餐馆的位置。利用已有的餐馆的销售总额和两公里半径范围内竞争者的数目以及可以收集到这些餐馆附近的人口、收入数据，来建立一个预测销售额的精确模型，从而利用该模型来发现未来餐馆的位置。

# 解决方案

输入：33个已有餐馆的相关信息（年销售额，竞争者数、人口数、收入）以及3个候选地点的年销售额、竞争者数、收入

输出：三个候选地点的预测销售额

方法：线性回归

# 数据结构

Table C.11. Variables for Studenmund’s Restaurants

|  |  |  |
| --- | --- | --- |
| 变量 | 描述 | 数据类型 |
| sales | 餐馆总的销售额（$/年） | int |
| competition | 餐馆2公里半径范围的直接竞争者的数目 | int |
| population | 餐馆3公里半径范围内居住人口的数目 | int |
| income | 餐馆3公里半径范围内平均家庭收入 | int |

# R代码相关分析

1. **载入相应的packages**

library(car) # 回归工具

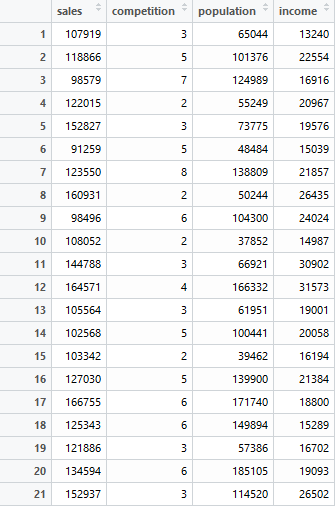
library(lattice) # correlation\_heat\_map（）

load("correlation\_heat\_map.RData") # R通用程序

1. **读入数据，显示**

restdata <- read.csv("studenmunds\_restaurants.csv", header = TRUE)# header = TRUE表示数据第一行为变量注释

View(restdata)



print(summary(restdata))#各个变量的最大值、最小值、平均值以及四分位数

sales competition population

Min. : 91259 Min. :2.000 Min. : 37852

1st Qu.:105564 1st Qu.:3.000 1st Qu.: 57386

Median :122015 Median :4.000 Median : 95120

Mean :125635 Mean :4.394 Mean :103887

3rd Qu.:140791 3rd Qu.:6.000 3rd Qu.:139900

Max. :166755 Max. :9.000 Max. :233844

income

Min. :13240

1st Qu.:16839

Median :19200

Mean :20553

3rd Qu.:22554

Max. :33242

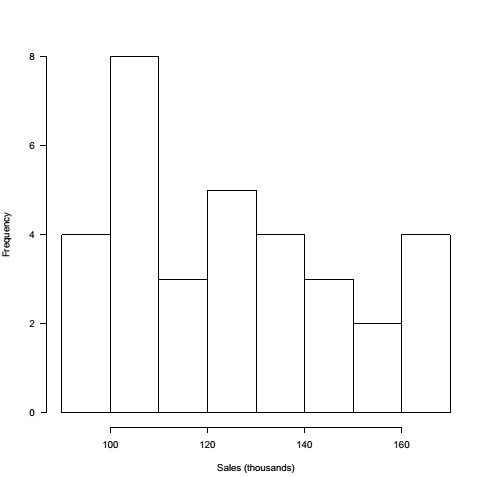
1. **对数据进行分析**

with(restdata, hist(sales/1000,

+ xlab="Sales (thousands)",

+ ylab="Frequency",

+ main = "", las = 1))

#销售总额（sales/1000）的频度直方图，如下图

with(restdata, plot(sort(sales/1000),

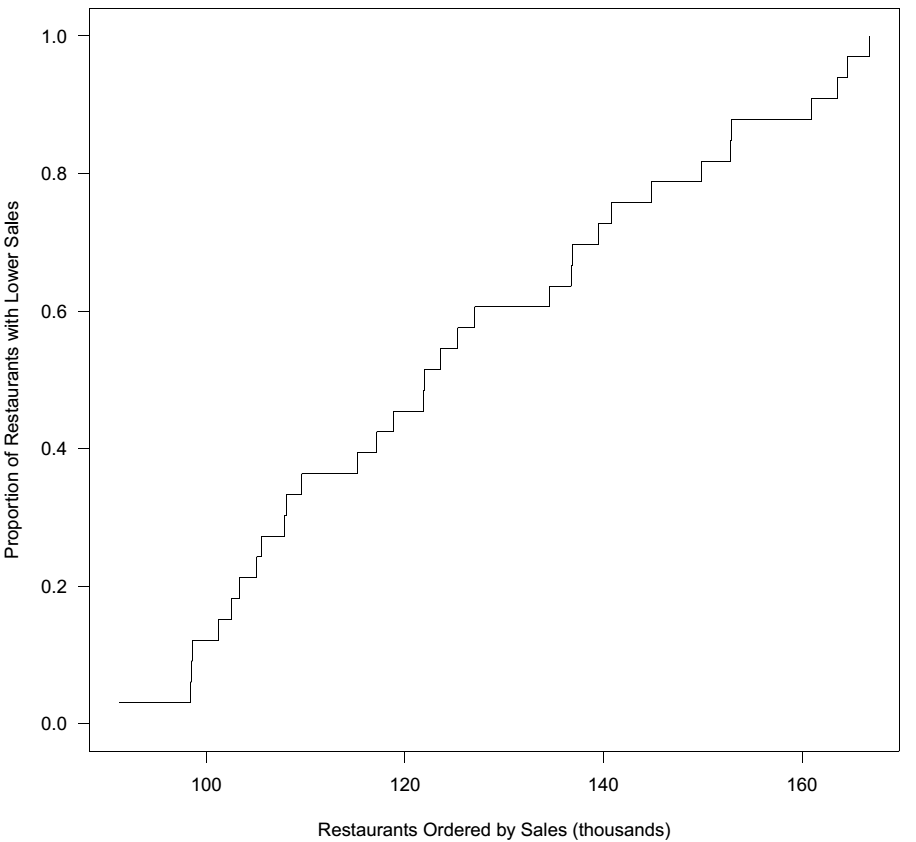
+ (1:length(sales))/length(sales),

+ type="s", ylim=c(0,1), las = 1,

+ xlab="Restaurants Ordered by Sales (thousands)",

+ ylab="Proportion of Restaurants with Lower Sales"))

#低收入餐馆比例图，如下图



pairs(restdata,

+ panel = function(x, y) {

+ points(x, y)

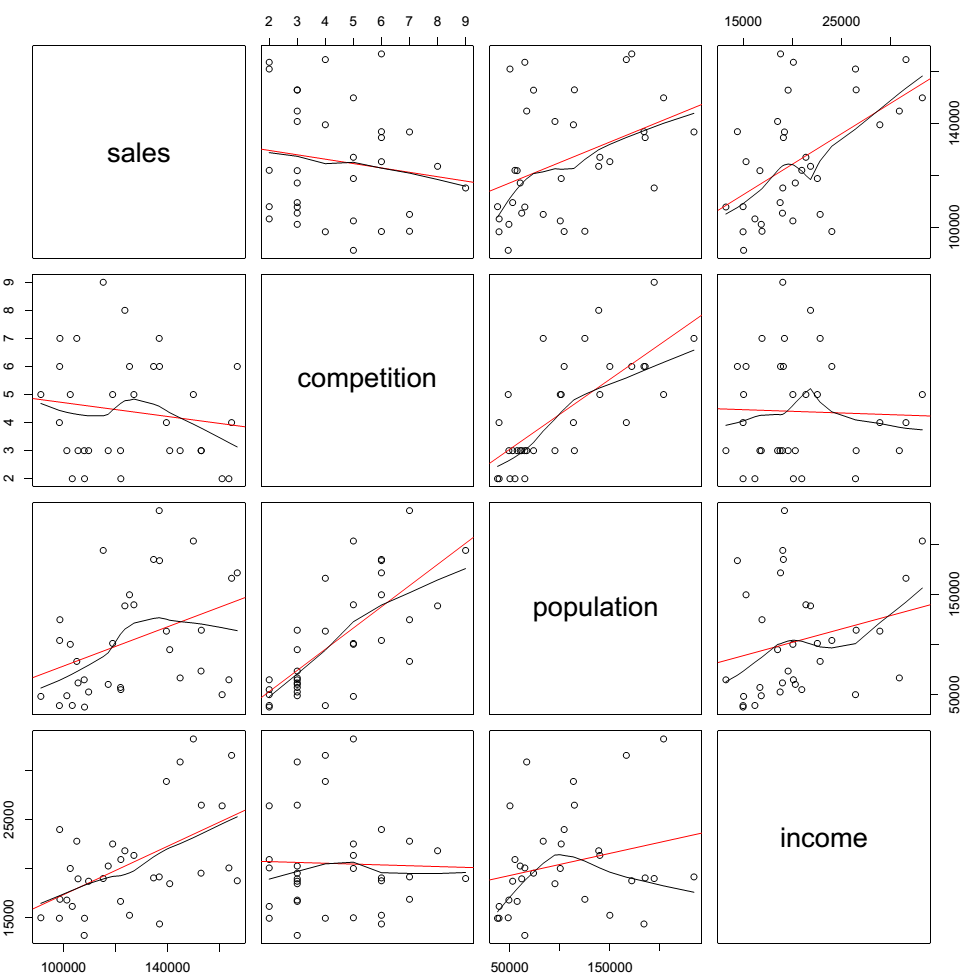
+ abline(lm(y ~ x), lty = "solid", col = "red")

+ lines(lowess(x, y))

+ }

+ )

#四个变量之间的两两关系点图



从上图中可以看出随着竞争增大，销售额会减少；人口越多，销售额越多；收入越多，销售额越多。

restdata\_cormat <-

+ cor(restdata[,c("sales","competition","population","income")])

> correlation\_heat\_map(cormat = restdata\_cormat)

#四个变量之间的两两关系点图



从图中可以看出，销售额与收入的关系为0.537，销售额与人口的关系为0.393，销售额与竞争者数目的关系为-0.144.

1. **线性模型及分析**

restdata\_model <- {sales ~ competition + population + income}

restdata\_fit <- lm(restdata\_model, data = restdata)

# report fitted linear model

print(summary(restdata\_fit))

Call:

lm(formula = restdata\_model, data = restdata)

Residuals:

Min 1Q Median 3Q Max

-21923 -8627 -2956 5328 33887

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.022e+05 1.280e+04 7.984 8.35e-09 \*\*\*

competition -9.075e+03 2.053e+03 -4.421 0.000126 \*\*\*

population 3.547e-01 7.268e-02 4.880 3.54e-05 \*\*\*

income 1.288e+00 5.433e-01 2.371 0.024623 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14540 on 29 degrees of freedom

Multiple R-squared: 0.6182, Adjusted R-squared: 0.5787

F-statistic: 15.65 on 3 and 29 DF, p-value: 3.058e-06

解析：

(Intercept)：截距

Estimate：相应变量的系数

R-squared：模型所能解释的方差的比例。

p-value:在原假设的基础上我们能得到观测数据的概率。

多重共线性可用统计量VIF（Variance Inflation Factor，方差膨胀因子）进行检测。VIF的平方根表示变量回归参数的置信区间能膨胀为与模型无关的预测变量的程度（因此而得名） 。 car包中的vif()函数提供VIF值。一般原则下， vif的平方根 >2就表明存在多重共线性问题。

print(vif(restdata\_fit))

competition population income

2.348446 2.496186 1.180772

可以看出vif的平方根都<2，不存在多重共线性的问题。

正态性 当预测变量值固定时，因变量成正态分布，则残差值也应该是一个均值为0的正态分布。正态Q-Q图（Normal Q-Q，右上）是在正态分布对应的值下，标准化残差的概率图。若满足正态假设，那么图上的点应该落在呈45度角的直线上；若不是如此，那么就违反了正态性的假设。可以看出有几个点（5 33 17）是离群点。

线性 若因变量与自变量线性相关，那么残差值与预测（拟合）值就没有任何系统关联。换句话说，除了白噪声，模型应该包含数据中所有的系统方差。 “残差图与拟合图”Residuals vs Fitted，左上**.**

同方差性 若满足不变方差假设，那么在位置尺度图（Scale-Location Graph，左下）中，水平线周围的点应该随机分布图中的个点是随机分布的。

最后一幅“残差与杠杆图”（Residuals vs Leverage，右下）提供了你可能关注的单个观测点的信息。从图形可以鉴别出离群点、高杠杆值点和强影响点。图中17是离群点，23是高杠杆值，无强影响点。

对最后一幅图的全面解析：

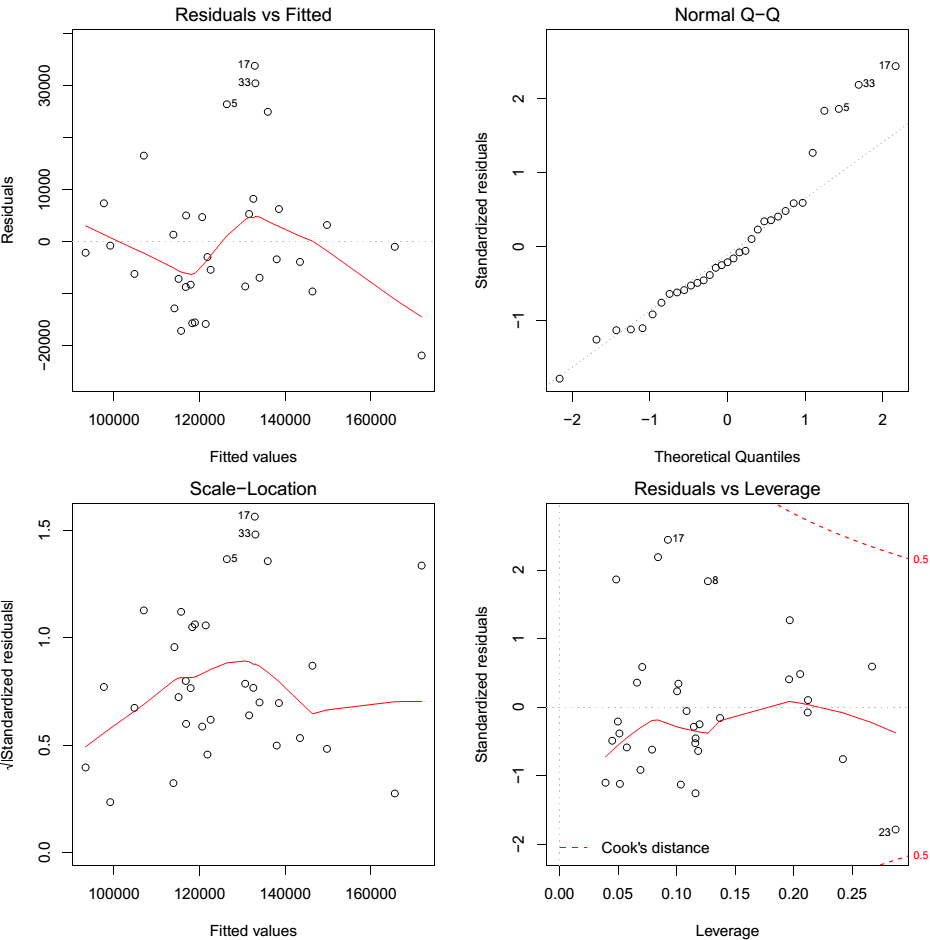
一个全面的回归分析要覆盖对异常值的分析，包括离群点、高杠杆值点和强影响点。

离群点是指那些模型预测效果不佳的观测点。它们通常有很大的、或正或负的残差（Yi−Y预测 ）。正的残差说明模型低估了响应值，负的残差则说明高估了响应值。

高杠杆值观测点，即是与其他预测变量有关的离群点。换句话说，它们是由许多异常的预测变量值组合起来的，与响应变量值没有关系。高杠杆值的观测点可通过帽子统计量（hat statistic）判断。对于一个给定的数据集，帽子均值为p/n，其中p 是模型估计的参数数目（包含截距项）， n 是样本量。一般来说，若观测点的帽子值大于帽子均值的2或3倍，即可以认定为高杠杆值点。

强影响点，即对模型参数估计值影响有些比例失衡的点。例如，若移除模型的一个观测点时模型会发生巨大的改变，那么你就需要检测一下数据中是否存在强影响点了。Cook’s D值大于4/(n−k −1)，则表明它是强影响点，其中n 为样本量大小， k 是预测变量数目。

就右下图来看，纵坐标超过+2或小于−2的州可被认为是离群点，水平轴超过0.2或0.3的有高杠杆值（通常为预测值的组合）。圆圈大小与影响成比例，圆圈很大的点可能是对模型参数的估计造成的不成比例影响的强影响点



1. **预测三个潜在餐馆的销售额**

sites <- data.frame(sales = c(NA, NA, NA),

+ competition = c(2, 3, 5),

+ population = c(50000, 200000, 220000),

+ income = c(25000, 22000, 19000))

> sites$predicted\_sales <- round(predict(restdata\_fit, newdata = sites),0)

> print(sites)

sales competition population income predicted\_sales

1 NA 2 50000 25000 133975

2 NA 3 200000 22000 174236

3 NA 5 220000 19000 159317

# 结论

从预测结构可以看出第二个地点的预测的销售额最高，将会作为未来的餐馆地点。