# 数据分析及可视化练习题

以下所有任务，将分析所用R语言代码、绘制的图形以及自己给予的解释，都放置在相关题目的下方。

1. 用某药物治疗10名高血压病人，对每一病人治疗前后的舒张压(mmHg)进行了测量，结果保存在ex1.csv中，试分析该药有无降压作用？给出分析R语言代码，对分析输出进行解释。（假设舒张压服从正态分布）

答：

代码：

df = read.csv("ex1.csv", stringsAsFactors = F, header = T)

t.test(x = df$before, y = df$after, paired = T, alternative = "greater")

结果：

Paired t-test

data: df$before and df$after

t = 2.6455, df = 9, p-value = 0.01334

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

3.070715 Inf

sample estimates:

mean of the differences

10

因为p = 0.01334 < 0.05，

所以接受H1，拒绝H0。

95%置信区间下可以认为药物有降压作用。

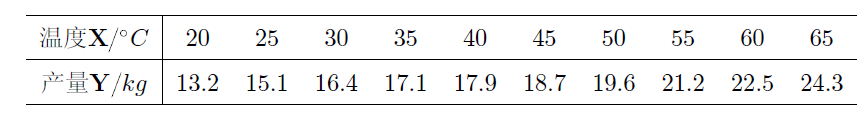
1. 考察温度对产量的影响，测得10组数据，如下表所示。使用R语言完成以下任务：

（1）尝试使用线性回归对问题进行分析，建立回归函数；

（2）对线性回归函数进行显著性检验；

（3）预测X=42度时产量的估计值以及预测区间（置信度为95%）；

（4）尝试使用残差图对回归方程进行诊断。



答：

代码：

# Created by yx

x = seq(20, 65, 5)

y = c(13.2, 15.1, 16.4, 17.1, 17.9, 18.7, 19.6, 21.2, 22.5, 24.3)

df = data.frame(x, y)

mdl = lm(y ~ 1 + x)

lf = function(x){

return(mdl$coefficients[[1]] + mdl$coefficients[[2]] \* x)

}

cat("model: y = ", mdl$coefficients[[1]], " + ", mdl$coefficients[[2]], " \* x\n")

print(summary(mdl))

predPoint = data.frame(x = 42)

print(predict(mdl, predPoint, interval = "prediction", level = 0.95))

op = par(mfrow = c(2, 2))

plot(mdl)

par(op)

分析：

回归函数：y = 9.121212 + 0.2230303 \* x

显著性检验：

Coefficients:

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

(Intercept) 9.12121 0.47708 19.12 5.8e-08 \*\*\*

x 0.22303 0.01063 20.97 2.8e-08 \*\*\*

---

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

两个参数的p值都是\*\*\*，可以认为十分显著。

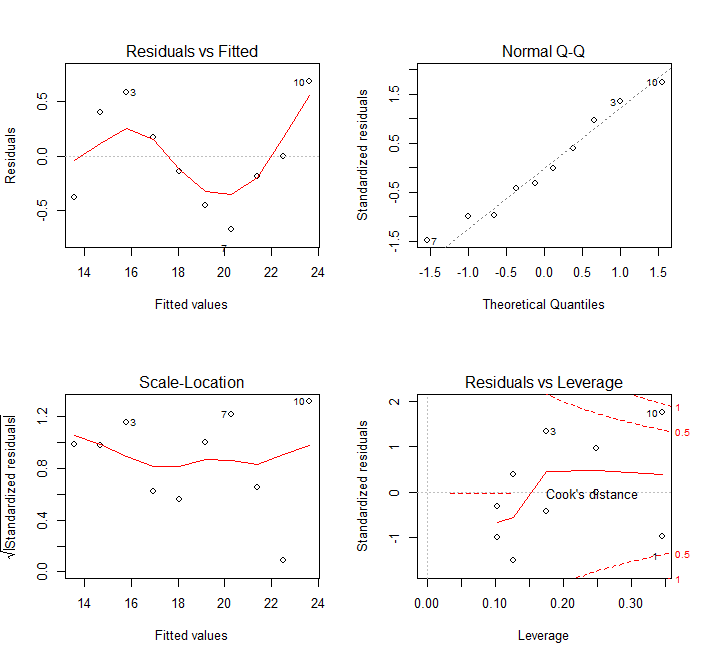
预测x=42:

fit lwr upr

1 18.48848 17.32034 19.65663

预测值：18.48848，置信区间：[17.32034, 19.65663]

残差诊断：

拟合值vs残差：数据点残差较小，基本分布于y=0附近，虽然有一定趋势，但是残差值较小，因此，可以基本判定残差项正常。

Q-Q图：分布趋于一条直线，说明可以认为残差服从正态分布。

标准化残差方根分布：y=24时有残差最大点。

Cook距离：可以推测图像最右侧的残差项在数据改变的情况下改变最大。

1. 基于mtcars数据集，分析汽车各指标之间的相关性，要求：
2. 绘制相关矩阵图
3. 进行相关系数显著性检验并简单解释。

答：

代码：

library(corrplot)

library(psych)

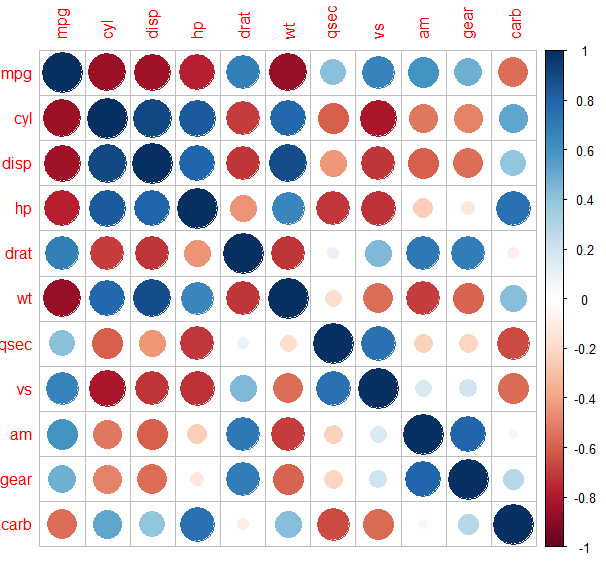
corrMat = cor(mtcars)

corrplot(corrMat)

print(corr.test(mtcars, adjust = "none", use = "complete"))

分析：

相关矩阵图：



相关系数显著性检验：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| p值 | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
| mpg | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.01 | 0.00 |
| cyl | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| disp | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.03 |
| hp | 0.00 | 0 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.18 | 0.49 | 0.00 |
| drat | 0.00 | 0 | 0.00 | 0.01 | 0.00 | 0.00 | 0.62 | 0.01 | 0.00 | 0.00 | 0.62 |
| wt | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.34 | 0.00 | 0.00 | 0.00 | 0.01 |
| qsec | 0.02 | 0 | 0.01 | 0.00 | 0.62 | 0.34 | 0.00 | 0.00 | 0.21 | 0.24 | 0.00 |
| vs | 0.00 | 0 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.36 | 0.26 | 0.00 |
| am | 0.00 | 0 | 0.00 | 0.18 | 0.00 | 0.00 | 0.21 | 0.36 | 0.00 | 0.00 | 0.75 |
| gear | 0.01 | 0 | 0.00 | 0.49 | 0.00 | 0.00 | 0.24 | 0.26 | 0.00 | 0.00 | 0.13 |
| carb | 0.00 | 0 | 0.03 | 0.00 | 0.62 | 0.01 | 0.00 | 0.00 | 0.75 | 0.13 | 0.00 |

p值小于0.05时，可以推断在95%置信水平下，两个变量显著相关。