班级	xxx	学号	xxx	姓名	xxx

结合课程上机 (R 语言), 完成以下上机作业问题:

注 1 基本要求:

1) 针对题目要求给出解答, 给出核心关键代码, 不必要粘贴所有源代码):

2) 简要概述你通过编程解决此问题所遇到的难点及收获的 R 语言或课程理论等方面的心得。

注 2 评价依据:

1) 解答的完整性、正确性: 是否缺少内容、是否计算无误;

2) 难点分析: 是否记录了解决问题过程中的难点和心得;

3) 核心关键代码是否有恰当的注释;

4) 雷同抄袭等: 超期作业评价不超过 60 分, 雷同抄袭一律 0 分。

1 习题 1.3

数据:

[1]:
```
data <- read.table("./ex_1_3.txt", header=TRUE); data
```

A data.frame: 22 × 4

Year <int>	Nationwide <int>	Rural <int>	Urban <int>
1978	184	138	405
1979	207	158	434
1980	236	178	496
⋮	⋮	⋮	⋮
1997	2834	1876	5796
1998	2972	1895	6217
1999	3180	1973	6651

[2]:
```
summary(data)
```

```
      Year        Nationwide         Rural            Urban       
 Min.   :1978   Min.   : 184.0   Min.   : 138.0   Min.   : 405.0  
 1st Qu.:1983   1st Qu.: 321.8   1st Qu.: 255.2   1st Qu.: 617.8  
 Median :1988   Median : 727.5   Median : 530.5   Median :1499.5  
```

```
 Mean   :1988   Mean   :1117.0   Mean   : 747.9   Mean   :2336.4
 3rd Qu.:1994   3rd Qu.:1642.2   3rd Qu.:1052.2   3rd Qu.:3675.0
 Max.   :1999   Max.   :3180.0   Max.   :1973.0   Max.   :6651.0
```

[3]:
```
attach(data)
```

方便直接使用 Year, Nationwide, Rural, Urban 这几个变量。

1.1 (1)

求均值、方差、标准差、变异系数、偏度、峰度

均值:

[4]:
```
c(mean(Nationwide), mean(Rural), mean(Urban))
```

1. 1117 2. 747.863636363636 3. 2336.40909090909

或者，R 提供有一个更方便的函数 `colMeans`，可以直接计算 data frame 各列均值 (`data[-1]` 排除了第一列 Year):

[5]:
```
colMeans(data[-1])
```

Nationwide 1117 **Rural** 747.863636363636 **Urban** 2336.40909090909

colMeans 函数等价于如下 apply:

[6]:
```
apply(data[-1], MARGIN=2, FUN=mean)
```

Nationwide 1117 **Rural** 747.863636363636 **Urban** 2336.40909090909

注: `MARGIN = 2` 就是取列。

方差:

[7]:
```
apply(data[-1], 2, var)
```

Nationwide 1031680.28571429 **Rural** 399673.837662338 **Urban** 4536136.44372294

标准差:

[8]:
```
apply(data[-1], 2, sd)
```

Nationwide 1015.71663652531 **Rural** 632.197625479832 **Urban** 2129.82075389525

变异系数:

[9]:
```r
cv <- function(x) sd(x)/mean(x)
apply(data[-1], 2, cv)
```

Nationwide 0.90932554747118 **Rural** 0.845338100076357 **Urban** 0.91157869663422

偏度（skewness）：

可以调用一个库完成计算 [?]，也可以照着书上写公式 [?]：

[10]:
```r
# install.packages("psych")
library(psych)
```

[11]:
```r
# g1 <- function(x) skew(x, type=2)  # g1、g2 是用 type=2: seehelp(skew)
# or 按照书上（P6）的手写这个函数
g1 <- function(x) {
    n <- length(x)
    A <- n / ((n-1) * (n-2))
    B <- 1 / sd(x)^3
    S <- sum((x - mean(x))^3)
    A * B * S
}
apply(data[-1], 2, g1)
```

Nationwide 1.02484718945818 **Rural** 1.01256119786818 **Urban** 0.970464107448573

峰度：

[12]:
```r
# g2 <- function(x) kurtosi(x, type=2)  # help(kurtosi)
g2 <- function(x) {
    n <- length(x)
    A <- (n * (n+1)) / ((n-1) * (n-2) * (n-3))
    B <- 1 / sd(x)^4
    S <- sum((x - mean(x))^4)
    C <- (3 * (n-1)^2) / ((n-2) * (n-3))
    A * B * S - C
}
apply(data[-1], 2, g2)
```

Nationwide -0.457241207817378 **Rural** -0.451444093083178 **Urban** -0.573162098915409

注：实际上，psych 包提供了一个 `describe` 函数，可以一次性得到各种常用值：

[13]:
```
describe(data[-1], type=2)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Nationwide	1	22	1117.0000	1015.7166	727.5	1001.7222	673.8417	184	3180	2996	1.0248472	-0.4572412	216.5515
Rural	2	22	747.8636	632.1976	530.5	682.7222	469.9842	138	1973	1835	1.0125612	-0.4514441	134.7850
Urban	3	22	2336.4091	2129.8208	1499.5	2094.1111	1379.5593	405	6651	6246	0.9704641	-0.5731621	454.0793

1.2 (2)

中位数:

[14]:
```
apply(data[-1], 2, median)
```

Nationwide 727.5 **Rural** 530.5 **Urban** 1499.5

四分位距:

[15]:
```
apply(data[-1], 2, quantile)
```

A matrix: 5 × 3 of type dbl

	Nationwide	Rural	Urban
0%	184.00	138.00	405.00
25%	321.75	255.25	617.75
50%	727.50	530.50	1499.50
75%	1642.25	1052.25	3675.00
100%	3180.00	1973.00	6651.00

五数:

[16]:
```
fn <- apply(data[-1], 2, fivenum)
fn
```

A matrix: 5 × 3 of type dbl

Nationwide	Rural	Urban
184.0	138.0	405.0
311.0	246.0	603.0
727.5	530.5	1499.5
1746.0	1118.0	3891.0
3180.0	1973.0	6651.0

四分位极差:

[17]:
```
R1 <- function(Q3, Q1) Q3 - Q1
R1(Q3=fn[4,], Q1=fn[2,])
```

Nationwide 1435 **Rural** 872 **Urban** 3288

三均值:

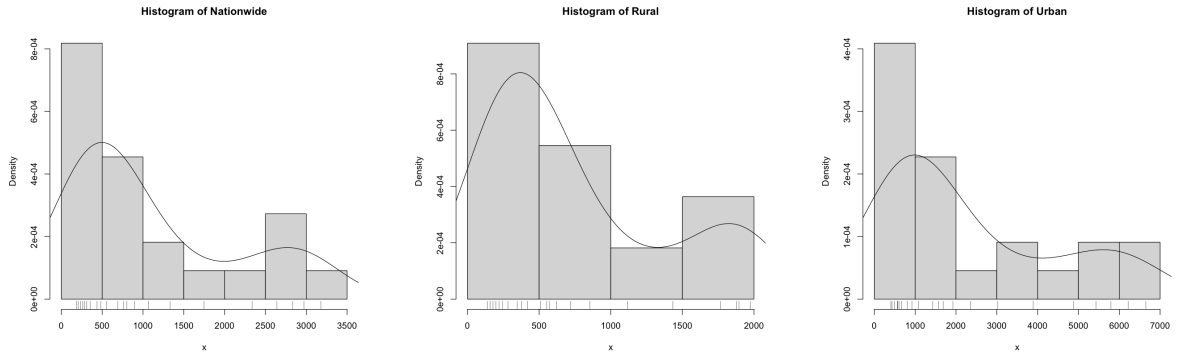
[18]:
```
M3 <- function(Q1, M, Q3) Q1/4 + M/2 + Q3/4
M3(Q1=fn[2,], M=fn[3,], Q3=fn[4,])
```

Nationwide 878 **Rural** 606.25 **Urban** 1873.25

1.3 (3) 直方图

[19]:
```
histogram <- function(x, xname="x") {
    hist(x, prob=TRUE, main=paste("Histogram of" , xname))
    lines(density(x))
    rug(x) # show the actual data points
}
```

[20]:
```
# layout(matrix(c(1,2,3), nr=1, byrow=T))
histogram(Nationwide, xname="Nationwide")
histogram(Rural, xname="Rural")
histogram(Urban, xname="Urban")
```



1.4 (4) 茎叶图

[21]:
```
stem(Nationwide)
```

```
  The decimal point is 3 digit(s) to the right of the |
```

```
  0 | 22233344567889
  1 | 137
  2 | 368
  3 | 02
```

[22]:
```
stem(Rural)
```

```
  The decimal point is 3 digit(s) to the right of the |

  0 | 1222223344
  0 | 566679
  1 | 14
  1 | 899
  2 | 0
```

[23]:
```
stem(Urban)
```

```
  The decimal point is 3 digit(s) to the right of the |

  0 | 44566678914679
  2 | 409
  4 | 948
  6 | 27
```

1.5 (5) 异常值

[24]:
```
abnormal <- function(x) {
    fn <- fivenum(x);
    Q1 <- fn[2];  Q3 <- fn[4];
    R1 <- Q3 - Q1
    QD <- Q1 - 1.5 * R1
    QU <- Q3 + 1.5 * R1
    x[(x < QD) | (x > QU)]
```

```
}
```

[25]:
```
apply(data[-1], 2, abnormal)
```

结果为空：没有异常值。

[26]:
```
detach(data)
```

2 习题 1.4

[1]:
```
data <- read.csv("./ex_1_4.csv")
cat(names(data))
```

```
X.序号 省市区 X11 月 X1.11 月
```

设「11 月」为 X_1，「1～11 月」为 X_2：

[2]:
```
data <- data[-1]  # remove "序号" col
names(data) <- c("Province", "X1", "X2")
data
```

A data.frame: 31 × 3

Province	X1	X2
<chr>	<dbl>	<dbl>
北京	35.22	499.80
天津	10.41	161.37
河北	17.22	273.29
⋮	⋮	⋮
青海	1.21	18.30
宁夏	2.31	23.81
新疆	3.24	103.81

[3]:
```
attach(data)
```

2.1 (1) 均值、方差、标准差、变异系数、偏度、峰度

调用库完成：

[4]:
```
library(psych)
describe(data[-1], type=2)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
X1	1	31	19.16645	19.79977	14.77	15.7252	8.258082	0.77	99.32	98.55	2.515352	8.266989	3.556143
X2	2	31	246.19323	232.97210	179.41	210.6356	123.856404	6.08	1080.26	1074.18	1.915957	4.385233	41.843024

手动实现:

[5]:
```
describes <- function(df) {
    cv <- function(x) sd(x)/mean(x)  # 变异系数
    g1 <- function(x) {  # 偏度
        n <- length(x)
        A <- n / ((n-1) * (n-2))
        B <- 1 / sd(x)^3
        S <- sum((x - mean(x))^3)
        A * B * S
    }
    g2 <- function(x) {  # 峰度
        n <- length(x)
        A <- (n * (n+1)) / ((n-1) * (n-2) * (n-3))
        B <- 1 / sd(x)^4
        S <- sum((x - mean(x))^4)
        C <- (3 * (n-1)^2) / ((n-2) * (n-3))
        A * B * S - C
    }
    itm <- matrix(c("均值", "方差", "标准差", "变异系数", "偏度", "峰度"), 6, 1)
    res <- apply(df, 2,
                 function(x) c(mean(x), var(x), sd(x), cv(x), g1(x), g2(x)))
    cbind(itm, res)
}
```

[6]:
```
describes(data[-1])
```

		X1	X2
A matrix: 6 × 3 of type chr	均值	19.1664516129032	246.193225806452
	方差	392.030750322581	54275.9982492473
	标准差	19.7997664209096	232.972097576614
	变异系数	1.0330428824697	0.946297757842323
	偏度	2.51535182567297	1.91595698889706
	峰度	8.26698939080861	4.38523271345801

2.2 (2) 中位数、上下四分位数、四分位极差

五数：minimum, lower-hinge, median, upper-hinge, maximum

[7]:
```
fn <- apply(data[-1], 2, fivenum)
fn
```

	X1	X2
A matrix: 5 × 2 of type dbl	0.770	6.080
	8.265	105.350
	14.770	179.410
	20.080	270.745
	99.320	1080.260

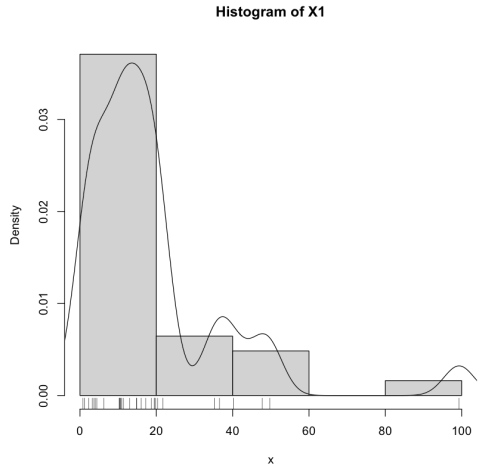
[8]:
```
# 四分位极差
R1 <- function(Q3, Q1) Q3 - Q1
R1(Q3=fn[4,], Q1=fn[2,])
```

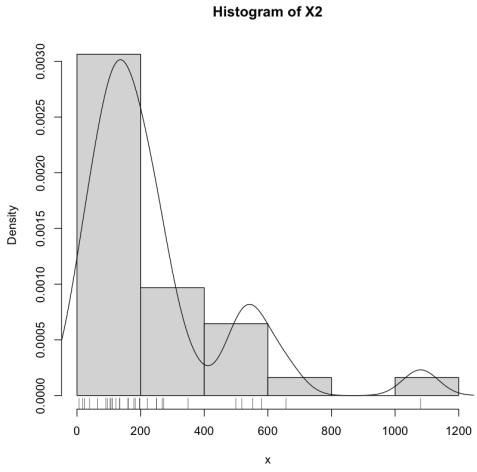
X1	11.815
X2	165.395

2.3 (3) 直方图

[9]:
```
histogram <- function(x, xname="x") {
    hist(x, prob=TRUE, main=paste("Histogram of" , xname))
    lines(density(x))
    rug(x) # show the actual data points
}
```

[10]:
```
histogram(X1, "X1")
histogram(X2, "X2")
```



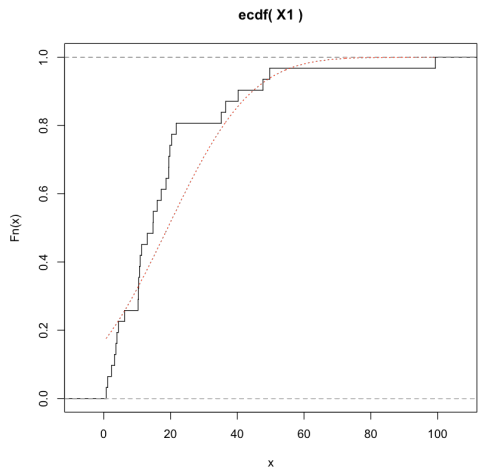


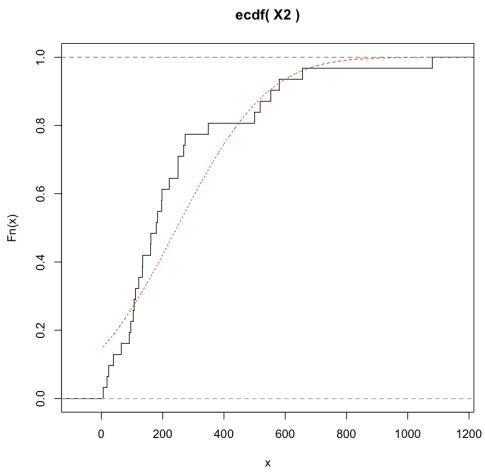
2.4 (4) 经验分布函数图

[11]:
```
plot_ecdf <- function(x, xname="x") {
    plot(ecdf(x), do.points=FALSE, verticals=TRUE, main=paste("ecdf(" , xname,␣
 ↪")"))
    xs <- seq(min(x), max(x), 1/sqrt(length(x)))
    lines(xs, pnorm(xs, mean=mean(x), sd=sd(x)), lty=3, col="red")
}
```

[12]:
```
plot_ecdf(X1, "X1")
```

[13]:
```
plot_ecdf(X2, "X2")
```

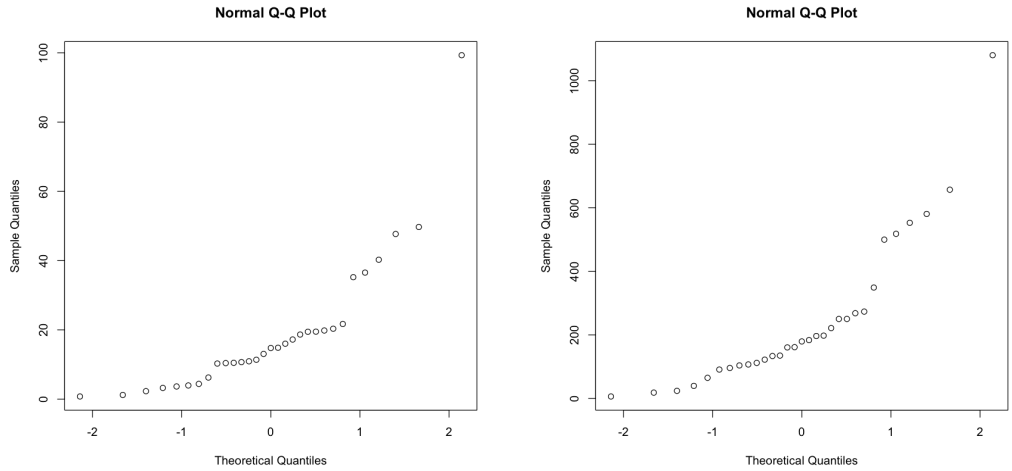




正态 QQ 图:

[14]:
```
qqnorm(X1)
```

[15]:
```
qqnorm(X2)
```



2.5 (5) X_1、X_2 的 Pearson 相关系数与 Spearman 相关系数

[16]:
```
cor.test(X1, X2, method="pearson")
```

```
        Pearson's product-moment correlation

data:  X1 and X2
t = 24.265, df = 29, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9508176 0.9886055
sample estimates:
      cor
0.9762474
```

[17]:
```
cor.test(X1, X2, method="spearman")
```

```
        Spearman's rank correlation rho

data:  X1 and X2
```

```
S = 358, p-value = 9.781e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9278226
```

[18]:
```
detach(data)
```

3 习题 2.4

[1]:
```
data <- read.table("./ex_2_4.txt", header=TRUE); data
```

A data.frame: 15 × 3

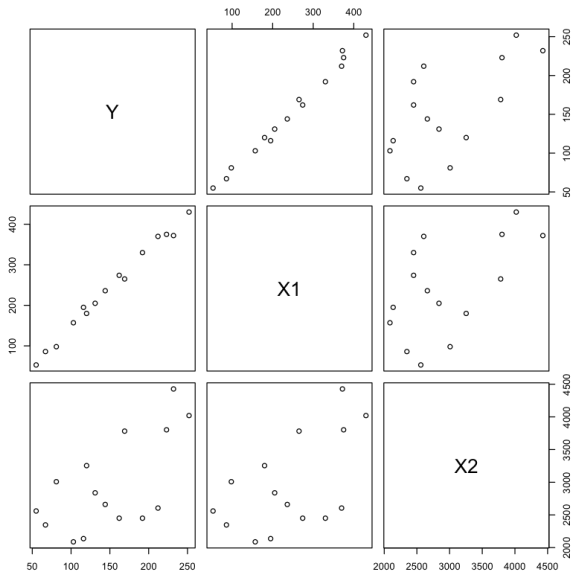
Y	X1	X2
<int>	<int>	<int>
162	274	2450
120	180	3254
223	375	3802
⋮	⋮	⋮
144	236	2660
103	157	2088
212	370	2605

[2]:
```
attach(data)
```

回归之前应该先看一下变量之间的相关关系如何，可以借助散点图矩阵来实现。

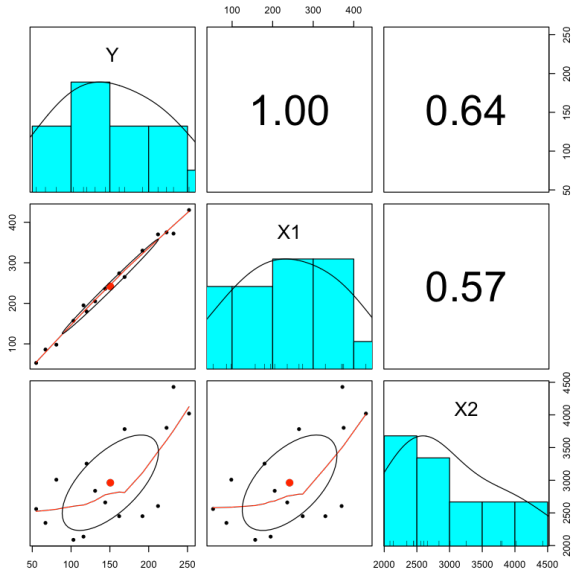
散点图矩阵（scatterplot matrix）：每个行与列的交叉点所在的散点图表示其所在的行与列的两个变量的相关关系：

[3]:
```
pairs(data)
```



用 psych 包里的 `pairs.panels` 可以作出有更多信息的图：

[4]:
```
library(psych)
pairs.panels(data)
```



- 对角线上方：相关系数
- 对角线：每个特征的数值分布直方图
- 下方：散点图
  - 相关椭圆：

    * 中心点：两个变量的均值所确定的点
    * 椭圆形状：两个变量的相关性：椭圆越被拉伸，其相关性越强
  - 局部回归平滑曲线：x 轴和 y 轴变量之间的一般关系

假设 Y 与 X_1, X_2 之间满足线性回归关系

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, 2, \cdots, 15$$

其中 $\epsilon_i (i = 1, 1, \cdots, 15)$ 独立同分布于 $N(0, \sigma^2)$.

3.1 (1)

求回归系数 $\beta_0, \beta_1, \beta_2$ 的最小二乘估计和误差方差 σ^2 的估计，写出回归方程并对回归系数作解释

回归:

[5]:
```
fm <- lm(Y ~ X1 + X2, data)
summary(fm)
```

```
Call:
lm(formula = Y ~ X1 + X2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8320 -1.2044 -0.2406  1.4888  3.3092

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.4526128  2.4306505   1.420    0.181
X1          0.4960050  0.0060544  81.924  < 2e-16 ***
X2          0.0091991  0.0009681   9.502  6.2e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.177 on 12 degrees of freedom
Multiple R-squared:  0.9989,        Adjusted R-squared:  0.9988
F-statistic:  5679 on 2 and 12 DF,  p-value: < 2.2e-16
```

Coefficients 是回归得到的系数，由此得到回归方程：

$$\hat{Y} = 0.496005X_1 + 0.009199X_2 + 3.452613$$

Residual standard error 即残差标准差，$\sigma = 2.177$, 由此得到误差方差的估计：

$$\sigma^2 = 4.739$$

3.2 (2)

求出方差分析表，解释对现行回归关系显著性检验的结果。求复相关系数的平方 R^2 的值并解释其意义

模型方差分析表:

[6]:
```
anova(fm)
```

A anova: 3 × 5

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
X1	1	53416.71863	53416.718634	11268.64354	3.270351e-19
X2	1	427.99780	427.997800	90.28923	6.201181e-07
Residuals	12	56.88357	4.740297	NA	NA

aov 输出更友好，可以直接显示出结果：

[7]:
```
summary(aov(fm))
```

```
            Df Sum Sq Mean Sq  F value  Pr(>F)
X1           1  53417   53417 11268.64 < 2e-16 ***
X2           1    428     428    90.29 6.2e-07 ***
Residuals   12     57       5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R^2 可以由 summary(fm) 输出的 `Multiple R-squared` 看出：

[8]:
```
summary(fm)$r.squared
```

0.998944677605876

$\therefore R^2 = 0.9989$

3.3 (3)

分别求 β_1 和 β_2 的置信度为 95% 的置信区间

[9]:
```
# 模型参数的置信区间
confint(fm, level = 0.95)
```

A matrix: 3 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	-1.843319690	8.74854527
X1	0.482813482	0.50919647
X2	0.007089742	0.01130842

得到所求置信区间:

$$\beta_1: \quad (0.482813482, 0.50919647)$$
$$\beta_2: \quad (0.007089742, 0.01130842)$$

3.4 (4)

对 $\alpha = 0.05$，分别检验人数 X_1 和收入 X_2 对销量 Y 的影响是否显著，利用回归系数有关的一般假设检验方法检验 X_1 和 X_2 的交互作用 (X_1X_2) 对 Y 的影响是否显著。

在 `summary(fm)` 输出的 Coefficients 段即可看出 X_1、X_2 对 Y 影响的显著性:

[10]:
```
summary(fm)#$coefficients
```

…

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.4526128  2.4306505   1.420    0.181
X1          0.4960050  0.0060544  81.924  < 2e-16 ***
X2          0.0091991  0.0009681   9.502  6.2e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

这一段输出里有:

- `t value`: T 检验的值

- Pr(>|t|): 表示 T 检验判定 P 值，后面有显著性标记（* 号）
- 显著性标记：* 个数对应显著性水平。

这里 X_1、X_2 都是有 *** 的，或者看 P 都在 0.0001 以下，所以对于 $\alpha = 0.05$，认为人数 X_1 和收入 X_2 对销量 Y 的影响显著。

注：这里可以做约简模型，求 $SSE(R)$、$SSE(T)$... 去分析，下面的函数会有帮助：（由于麻烦，这里不采用这种方法。）

[11]:
```
sse <- function(model, y) sum((fitted(model) - y)^2)
ssr <- function(model, y) sum((fitted(model) - mean(y))^2)
sst <- function(model, y) ssr(model, y) + sse(model, y)
ss <- function(model, y) {
    cbind(matrix(c("sse", "ssr", "sst"), 3, 1),
          c(sse(model, y), ssr(model, y), sst(model, y)))
    }
# R^2
# ssr(fm, Y)/sst(fm, Y)
```

下面研究 X_1、X_2 交互作用对 Y 的影响:

[12]:
```
fm1 <- update(fm, . ~ . + X1:X2)
summary(fm1)
```

```

Call:
lm(formula = Y ~ X1 + X2 + X1:X2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9094 -1.2010 -0.1811  1.5072  3.2141

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.901e+00  8.539e+00   0.574    0.578
X1          4.911e-01  2.832e-02  17.344 2.45e-09 ***
X2          8.674e-03  3.124e-03   2.777    0.018 *
X1:X2       1.698e-06  9.556e-06   0.178    0.862
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.271 on 11 degrees of freedom
Multiple R-squared:  0.9989,    Adjusted R-squared:  0.9987
F-statistic:  3481 on 3 and 11 DF,  p-value: < 2.2e-16
```

从拟合的结果里，`X1:X2` 的 P-value=0.862 比较大，说明交互作用对 Y 的影响不显著，没有必要引入交叉项。

3.5 (5)

(5) 该公司欲在一个适宜使用该化妆品的人数 x_{01} = 220，人均月收入 x_{02} = 2 500 的新的城市中销售该化妆品，求其销量的预测值及其置信度为 95% 的置信区间；

新数据：

[13]:
```
x01 = 220
x02 = 2500
newdata <- data.frame(X1 = c(x01), X2 = c(x02))
newdata
```

A data.frame: 1 × 2

X1	X2
<dbl>	<dbl>
220	2500

代入模型进行预测：

[14]:
```
predict(fm, newdata)
```

1: 135.571409702671

要获取置信区间需要再传入几个参数：

[15]:
```
predict(fm, newdata, interval="prediction", levels=0.95)
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	135.5714	130.5998	140.543

即得到预测值 $\hat{y}_0 = 135.5714$，置信度为 95% 的置信区间 $(130.5998, 140.543)$。

3.6 (6)

(6) 求 Y 的拟合值,残差及学生化残差. 根据对学生化残差正态性的频率检验及正态 QQ 图检验说明模型误差项的正态性假定是否合理,有序学生化残差与相应标准正态分布的分位数的相关系数是多少?作出各种残差图,分析模型有关假定的合理性.

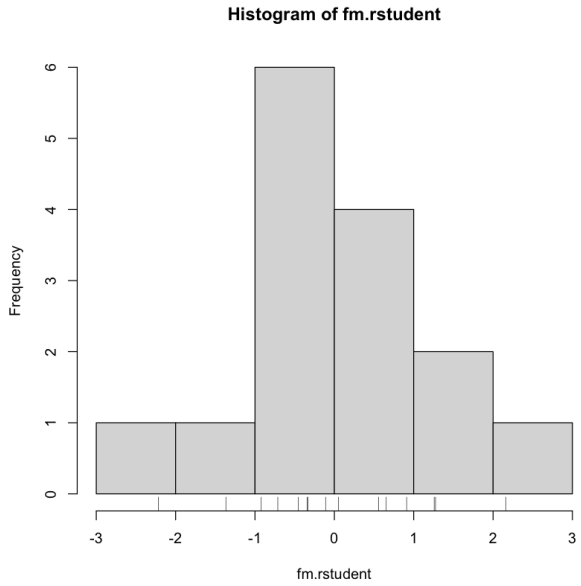
[16]:
```
# 拟合值:
fm.fitted <- fitted(fm)
# 残差:
fm.residuals <- residuals(fm)
# 学生化残差:
fm.rstudent <- rstudent(fm)
fmr <- data.frame(
    `拟合值`=fm.fitted,
    `残差`=fm.residuals,
    `学生化残差`=fm.rstudent
); fmr
```

A data.frame: 15 × 3

	拟合值 <dbl>	残差 <dbl>	学生化残差 <dbl>
1	161.89572	0.1042756	0.04973473
2	122.66732	-2.6673176	-1.36670170
3	224.42938	-1.4293843	-0.71265041
4	131.24062	-0.2406244	-0.11000544
5	67.69928	-0.6992835	-0.34443368
6	169.68486	-0.6848553	-0.33365032
7	79.73194	1.2680643	0.65018013
8	189.67200	2.3279970	1.25776219
9	119.83202	-3.8320189	-2.21655274
10	53.29052	1.7094765	0.91079624
11	253.71506	-1.7150576	-0.92397237
12	228.69079	3.3092051	2.16085046
13	144.97934	-0.9793423	-0.45379850
14	100.53307	2.4669251	1.27498091
15	210.93806	1.0619404	0.55945704

对于学生化残差用频率检验法:

[17]:
```
hist(fm.rstudent)
rug(fm.rstudent)
```



- 有 $\frac{10}{15} \approx 0.68$ 落在 $(-1.0, 1.0)$ 内,
- 有 $\frac{13}{15} \approx 0.87$ 落在 $(-1.5, 1.5)$ 内,
- 有 $\frac{15}{15} = 1.00$ 落在 $(-2.4, 2.4)$ 内,

可见，学生化残差与上述个区间的频率在 $N(0,1)$ 分布的相应概率相差不大，所以模型误差项的正态性假设是合理的。

进一步，可以做一个 ks 检测:

[18]:
```
ks.test(fm.rstudent, "pnorm", 0, 1)
```

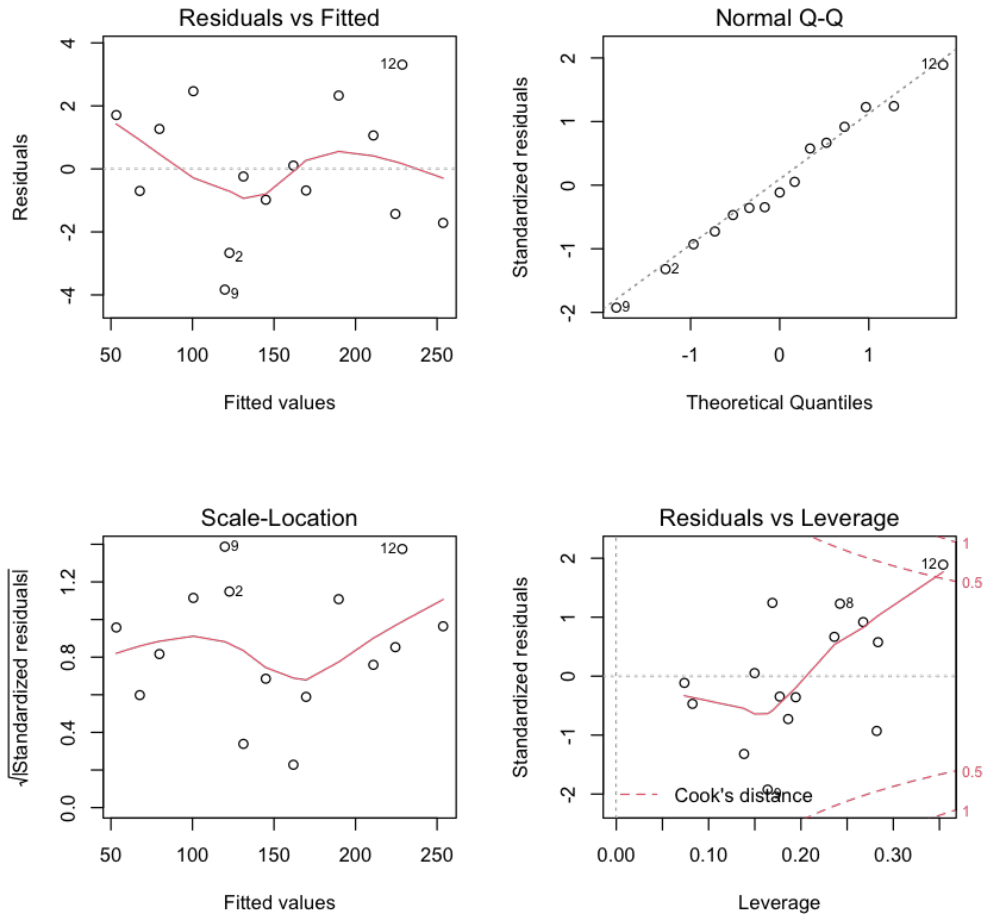
```
        One-sample Kolmogorov-Smirnov test

data:  fm.rstudent
D = 0.11208, p-value = 0.9808
alternative hypothesis: two-sided
```

检测结果也说明正态分布。

下面作出各种残差图:

[19]:
```
par(mfrow=c(2,2))
plot(fm)
```



- 左上图：**残差-拟合**：残差和拟合值之间，数据点均匀分布在 y=0 两侧，呈现出随机的分布，没有明显的形状特征，说明残差数据表现比较好。
- 右上图：**标准化残差 Q-Q**：数据点按对角直线排列，趋于一条直线，并被对角直接穿过，直观上符合正态分布。
- 左下图：**标准化残差-拟合**：数据随机分布与左上图蕾丝，无明显的形状特征。
- 右下图：**标准化残差杠杆图**：可以看出离群点、高杠杆值点和强影响点。（这里不做讨论）

由这些图，可以认为相应的线性回归模型以及误差的独立正态分布的假设是合理的。

[20]:
```
detach(data)
```

4 习题 2.5

[1]:
```
data <- read.csv("./ex_2_5.csv"); data
```

A data.frame: 20 × 2

x <dbl>	y <dbl>
0.05	5.9421
0.15	5.4691
0.25	5.8724
⋮	⋮
1.75	5.6500
1.85	6.0256
1.95	5.5350

[2]:
```
attach(data)
```

4.1 (1)

(1) 首先拟合 Y 关于 X 的线性回归模型，结果如何?通过残差分析(尤其是残差图分析)并参考 Y 与 X 的散点图，选择你认为合理的回归函数形式. 拟合你所选择的回归模型，再通过残差分析考察所设定的模型的合理性. 最后，将你所拟合的回归方程与真实模型($Y = 5 + (X - 1)^2 + \varepsilon, \varepsilon \sim N(0, 0.625)$)比较，你是否给出了正确的模型形式.

首先尝试拟合线性回归模型

$$\hat{Y} = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim N$$

[3]:
```
linear_model <- lm(y ~ x, data)
summary(linear_model)
```

```
Call:
lm(formula = y ~ x, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72831 -0.17147 -0.08433  0.16405  0.70025
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.319696   0.165202  32.201   <2e-16 ***
x           0.003054   0.143114   0.021    0.983
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3691 on 18 degrees of freedom
Multiple R-squared:  2.53e-05,        Adjusted R-squared:  -0.05553
F-statistic: 0.0004554 on 1 and 18 DF,  p-value: 0.9832
```

从拟合结果的检验可以看出，R^2 趋于 0，说明该模型拟合效果不好；同时 x 对应的 P 值非常高，X 对 Y 没有显著影响。

方差分析：

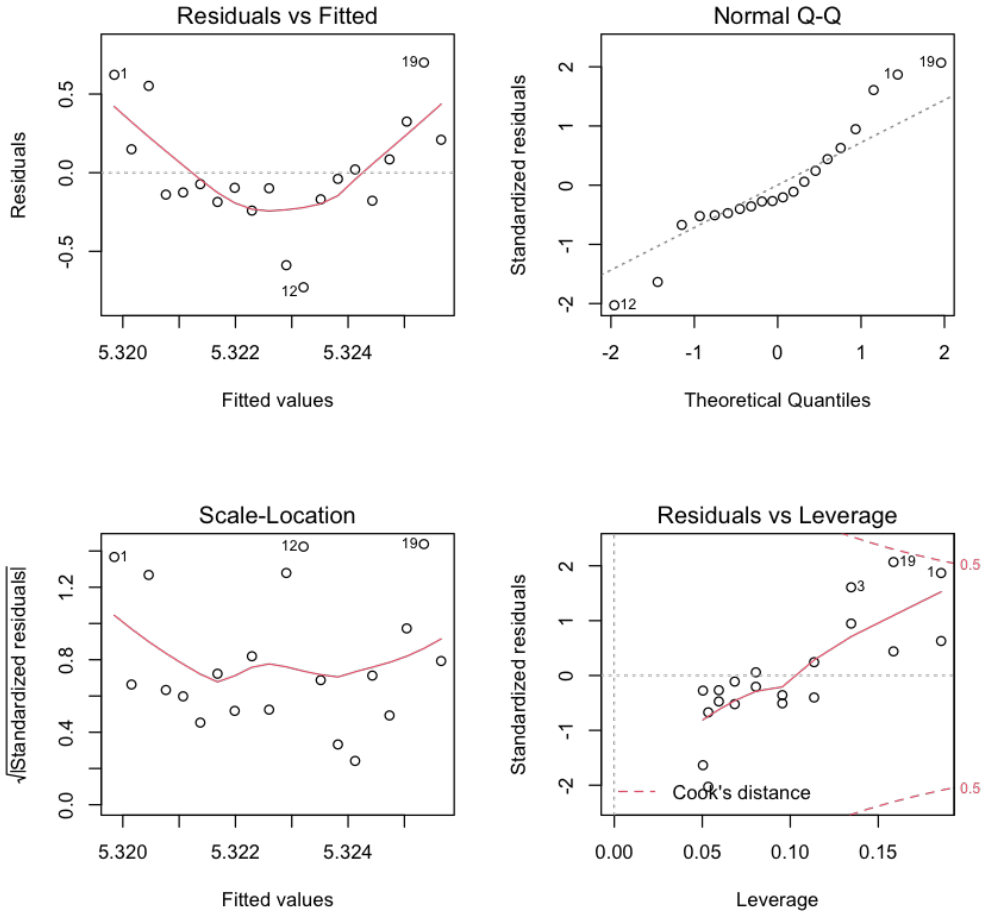
[4]:
```
summary(aov(linear_model))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
x            1 0.0001 0.00006       0  0.983
Residuals   18 2.4516 0.13620
```

这也证实了前面的结论。

下面作出各种残差图：

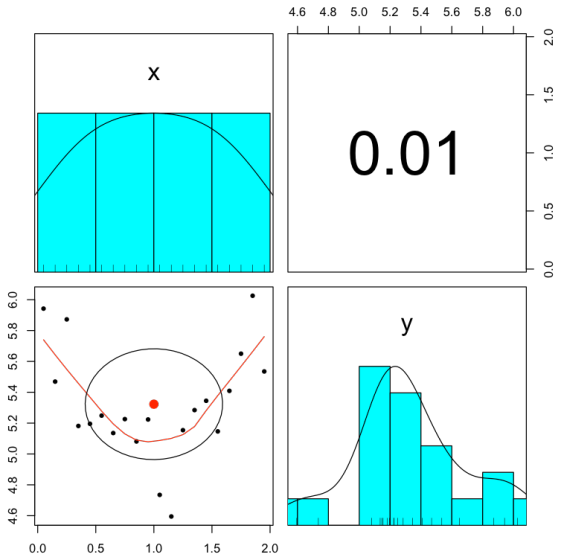
[5]:
```
par(mfrow=c(2,2))
plot(linear_model)
```



可以看出残差还是近似正太分布的。

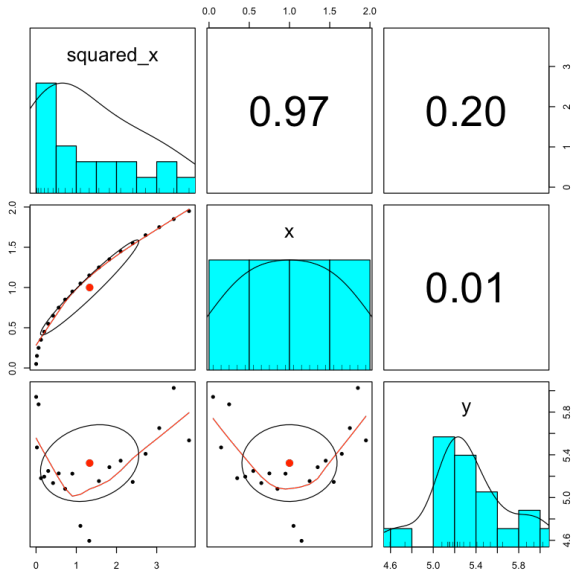
下面作出 x、y 的散点图：

[6]:
```
library(psych)
pairs.panels(data)
```



感觉上，x、y 呈现二次关系，再尝试作出 y 与 x^2 及 x 的散点图：

[7]:
```
pairs.panels(data.frame(`squared_x`=x^2, `x`=x, `y`=y))
```



可以看出 y 与 x^2 的相关性要优于线性。所以下面尝试做带有二次项的拟合 [?]：

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon, \quad \epsilon \sim N$$

[8]:
```
fm <- lm(y ~ x + I(x^2), data)
summary(fm)
```

```
Call:
lm(formula = y ~ x + I(x^2), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43449 -0.16752  0.05168  0.14782  0.33394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9524     0.1529  38.942  < 2e-16 ***
x            -1.8926     0.3535  -5.354 5.26e-05 ***
I(x^2)        0.9478     0.1712   5.537 3.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2268 on 17 degrees of freedom
Multiple R-squared:  0.6433,        Adjusted R-squared:  0.6013
F-statistic: 15.33 on 2 and 17 DF,  p-value: 0.0001565
```

得到回归方程:

$$\hat{Y} = 5.9524 - 1.8926X + 0.9478X^2$$

这一次拟合效果有了显著提升，从检验可以看出，加入的二次项 X^2 对 Y 有显著影响。模型 p 值小于 0.05，可以认为假设比较合理。

下面进一步做方差分析:

[9]:
```
summary(aov(fm))
```

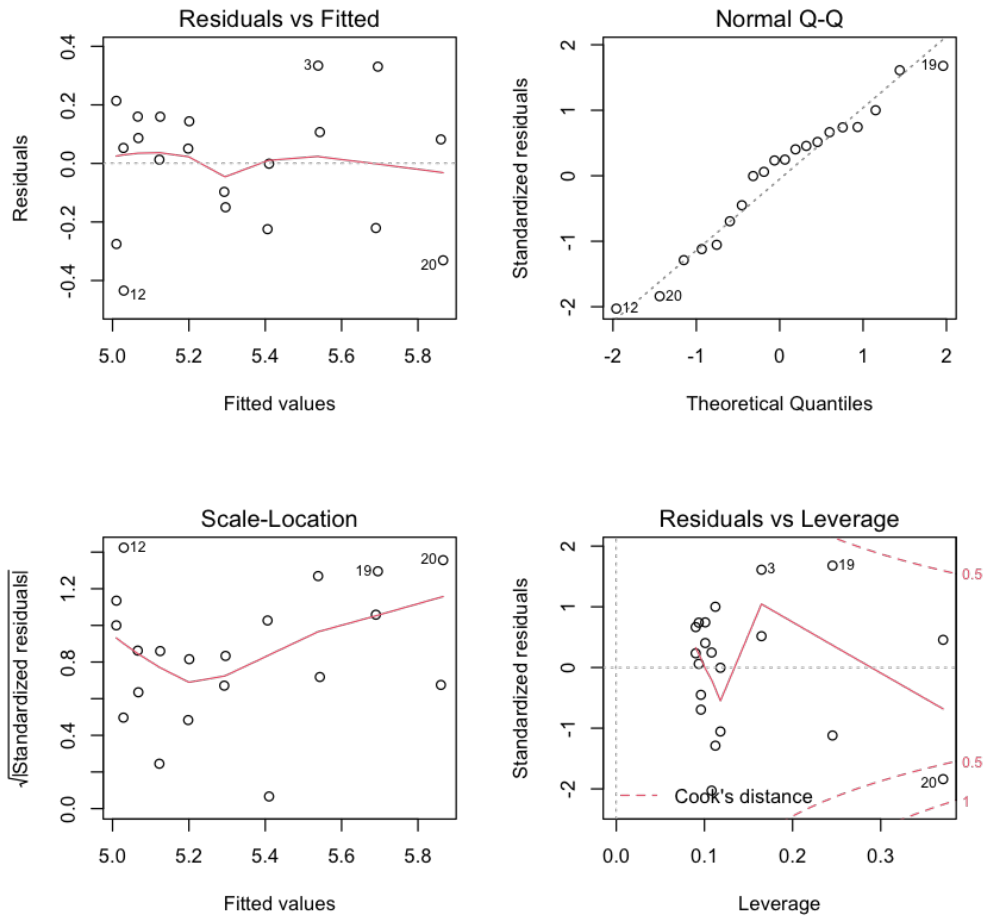
```
            Df Sum Sq Mean Sq F value   Pr(>F)
x            1 0.0001  0.0001   0.001    0.973
I(x^2)       1 1.5771  1.5771  30.658 3.62e-05 ***
```

```
Residuals   17 0.8745  0.0514
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

可以看到 X^2 对 Y 的影响显著，进一步证明了模型的合理性。

再作出各种残差图：

[10]:
```
par(mfrow=c(2,2))
plot(fm)
```



- 左下图：**标准化残差-拟合**：数据随机分布，无明显的形状特征，说明残差数据表现比较好。
- 右上图：**标准化残差 Q-Q**：数据点按对角直线排列，直观上符合正态分布。

由这些图，可以认为相应的线性回归模型以及误差的独立正态分布的假设是合理的。[?]

进一步，对残差做 Shapiro-Wilk 正态检验：

[11]:
```
shapiro.test(residuals(fm))
```

```
        Shapiro-Wilk normality test
data:  residuals(fm)
W = 0.95731, p-value = 0.4917
```

原假设 H_0 是数据服从正态分布，这里 $p > 0.05$，接受原假设，认为数据服从正态分布。

最后，拟合出的回归方程

$$\hat{Y} = 5.9524 - 1.8926X + 0.9478X^2$$

与题目给出的真实模型

$$\begin{aligned} Y &= 5 + (X-1)^2 \\ &= 6 - 2X + X^2 \end{aligned}$$

二者相比，形式一致，系数相差不大。拟合的效果比较好，模型选择正确。

4.2 (2)

(2) 如果对因变量作 Box-Cox 变换，求变换参数 λ 的值，拟合变换后的变量关于 X 的简单线性回归模型，结果如何？你对 Box-Cox 变换有何新的认识？

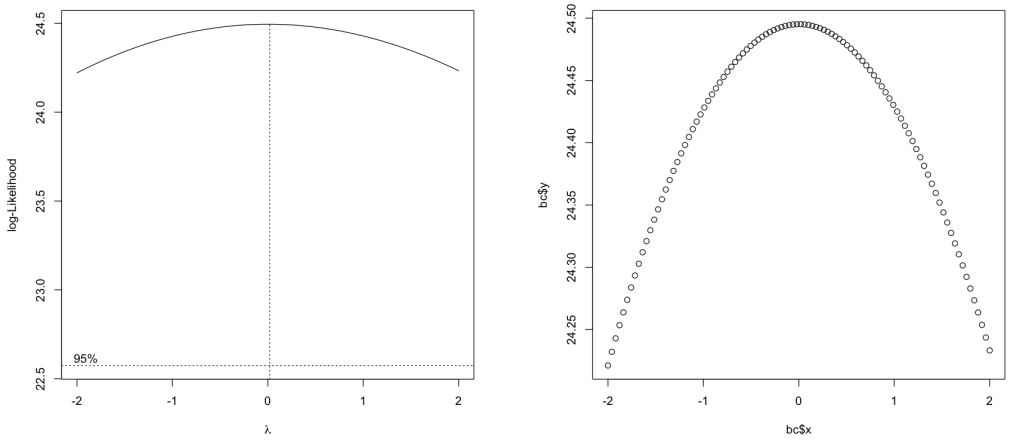
在 R 中实现 Box-Cox 变换，需要通过 MASS 包：

[12]:
```
library(MASS)
```

变换过程需要分两步 [?]：

1. 第一步：拟合 Box-Cox 模型，得到 lambda 值

[13]:
```
bc <- boxcox(y ~ ., data=data)  # . 表示除因变量外的所有变量
plot(bc)
```



[14]:
```
idx <- which(bc$y==max(bc$y))
lambda <- bc$x[idx]
lambda
```

0.0202020202020203

得到 $\lambda = 0.0202020202020203$，即 $\lambda -$ logLikelihood 图中的最高点。

2. 将上一步 Box-Cox 变换的 λ 值代入

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases}$$

拟合变换后的变量关于 X 的简单线性回归模型:

[15]:
```
fmbc <- lm((y^lambda-1)/lambda ~ x, data)
summary(fmbc)
```

```
Call:
lm(formula = (y^lambda - 1)/lambda ~ x, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14976 -0.03164 -0.01436  0.03356  0.13019

Coefficients:
```

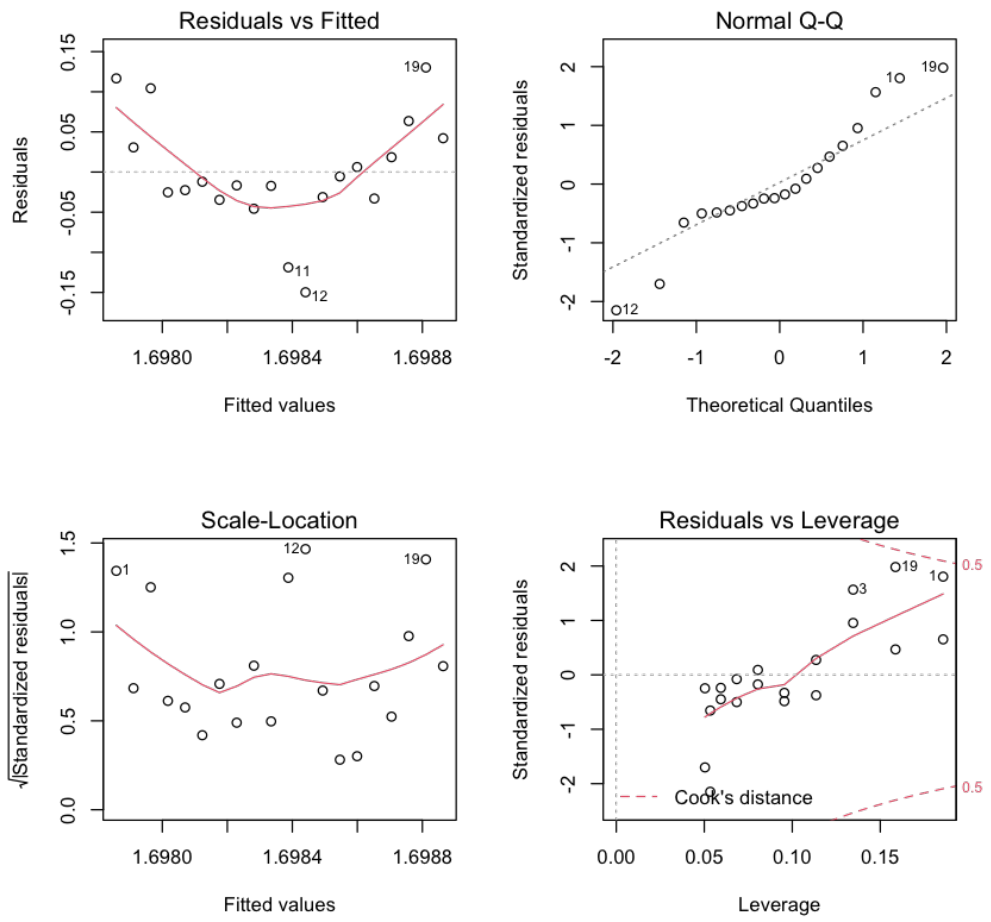
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.6978320  0.0320650  52.950   <2e-16 ***
x           0.0005291  0.0277778   0.019    0.985
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07163 on 18 degrees of freedom
Multiple R-squared:  2.016e-05,        Adjusted R-squared:  -0.05553
F-statistic: 0.0003628 on 1 and 18 DF,  p-value: 0.985
```

[16]:
```
summary(aov(fmbc))
```

```
            Df  Sum Sq  Mean Sq F value Pr(>F)
x            1 0.00000 0.000002       0  0.985
Residuals   18 0.09236 0.005131
```

[17]:
```
par(mfrow=c(2,2))
plot(fmbc)
```



[18]:
```
shapiro.test(residuals(fmbc))
```

```
        Shapiro-Wilk normality test
data:  residuals(fmbc)
W = 0.94268, p-value = 0.2693
```

从上面拟合结果回归分析、方差分析、残差图、残差正态性检验结果可以看出，Box-Cox 变化没有带来明显的改进。

做 Box-Cox 变换的意义在于使满足线性模型的正态性假设，而由前面第 (1) 题的数据拟合结果，已经满足正态性，这里再作 Box-Cox 变换没有意义。[?, ?]

[19]:
```
detach(data)
```

5　习题 2.6

[1]:
```
data <- read.csv("./ex_2_6.csv"); data
```

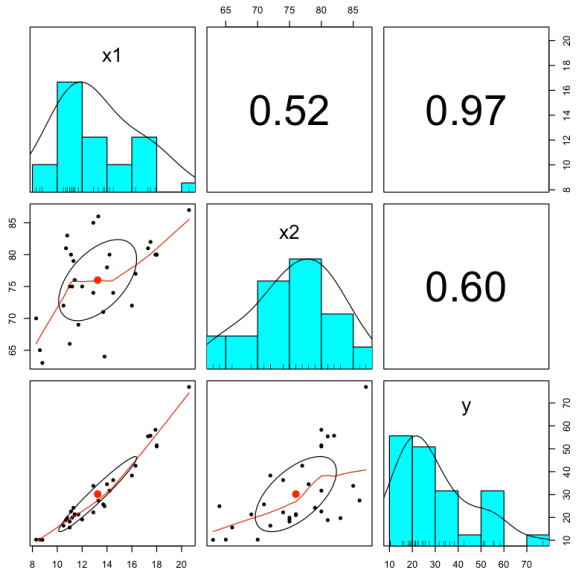
A data.frame: 31 × 3

直径.x1.	高度.x2.	体积.y.
<dbl>	<int>	<dbl>
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
⋮	⋮	⋮
18.0	80	51.5
18.0	80	51.0
20.6	87	77.0

[2]:
```
names(data) <- c("x1", "x2", "y")
attach(data)
```

在开始之前，先看一看数据的特征及相关关系：

[3]:
```
library(psych)
pairs.panels(data)
```



可以看到 x_1、y 有很强的线性相关性；x_2 对 y 的相关性略差，但也足够高。

5.1 (1)

(1) 首先拟合线性回归模型 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$，通过残差分析考察模型的合理性，是否需要对数据作变换？

[4]:
```
linear_model <- lm(y ~ x1 + x2, data=data)
summary(linear_model)
```

```
Call:
lm(formula = y ~ x1 + x2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
x1            4.7082     0.2643  17.816  < 2e-16 ***
x2            0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,        Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

分析结果中模型 $p < 0.05$，假设合理，$R^2 > 0.94$ 拟合效果较好。但注意到相比于 x_1，x_2 项对 y 没有非常显著的影响（$0.01 < p < 0.05$）。

用 aov 作出方差分析表：

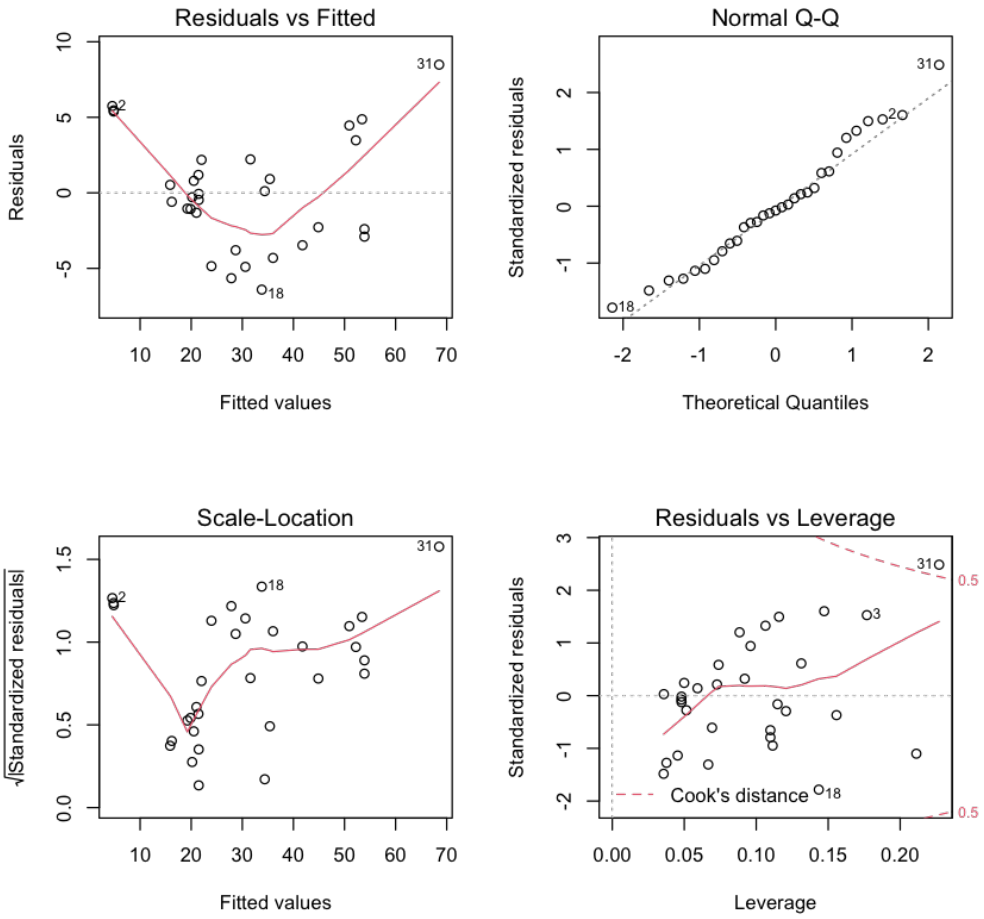
[5]:
```
summary(aov(linear_model))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
x1           1   7582    7582 503.150 <2e-16 ***
x2           1    102     102   6.794 0.0145 *
Residuals   28    422      15
```

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

下面作出各种残差诊断图:

[6]:
```
par(mfrow=c(2,2))
plot(linear_model)
```



从残差拟合图中可以看出，学生化残差明显不在同一条直线上，考虑对 Y 作 Box_Cox 变换。

5.2 (2)

(2) 对因变量 Y 作 Box-Cox 变换，确定变换参数 λ 的值. 对变换后的因变量重新拟合与 X_1, X_2 的线性回归模型并作残差分析，Box-Cox 变换的效果如何?

下面对数据进行 Box-Cox 变换后重新拟合:

[23]:
```
library(MASS)
bc <- boxcox(y ~ ., data=data)
idx <- which(bc$y==max(bc$y))
lambda <- bc$x[idx]
lambda
```

0.303030303030303

得到了 $\lambda = 0.303030303030303$，下面代入变换，重新拟合：

[24]:
```
fmbc <- lm((y^lambda-1)/lambda ~ ., data)
summary(fmbc)
```

```
Call:
lm(formula = (y^lambda - 1)/lambda ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42600 -0.14274 -0.01468  0.18705  0.36851

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.733542   0.500080  -5.466 7.77e-06 ***
x1           0.409448   0.015299  26.764  < 2e-16 ***
x2           0.039685   0.007535   5.267 1.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2247 on 28 degrees of freedom
Multiple R-squared:  0.9775,        Adjusted R-squared:  0.9759
F-statistic: 609.6 on 2 and 28 DF,  p-value: < 2.2e-16
```

拟合效果有所提升，

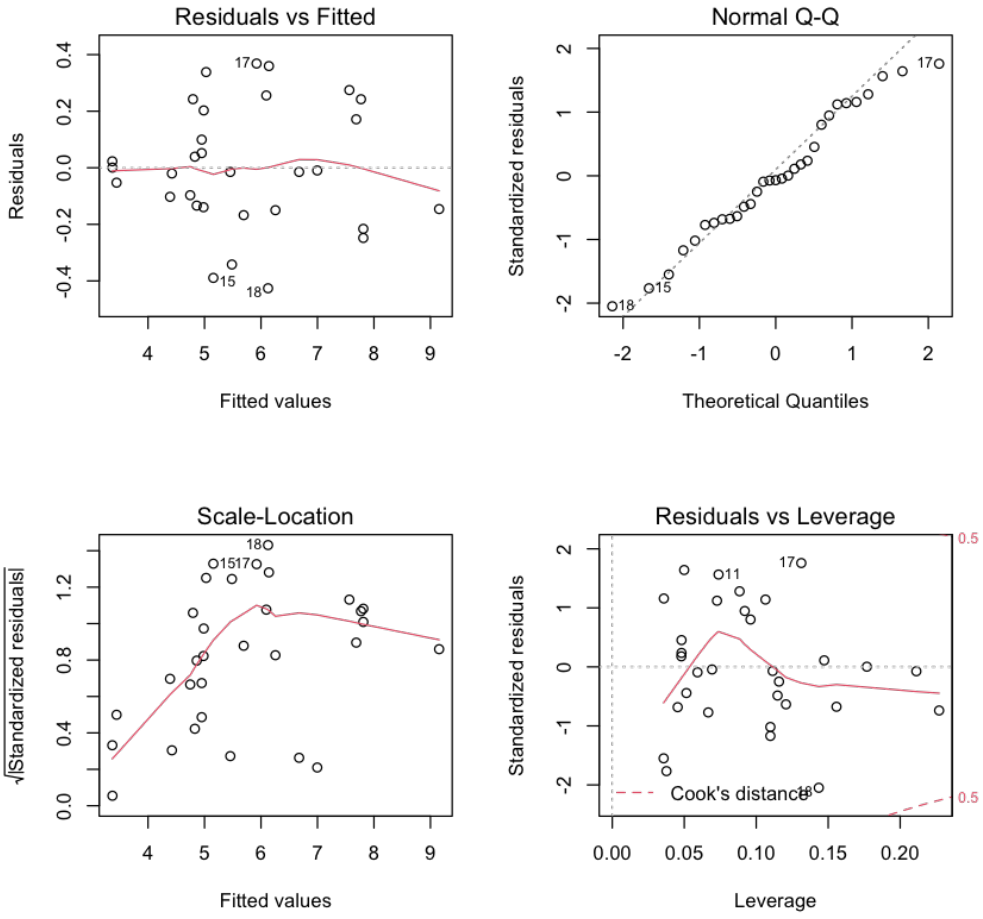
[26]:
```
summary(aov(fmbc))
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
x1           1  60.17   60.17 1191.45  < 2e-16 ***
```

```
x2           1   1.40    1.40   27.74 1.34e-05 ***
Residuals   28   1.41    0.05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

检验 $p < 0.001$，x_1、x_2 对变换后的 y 有显著影响，认为变换后的 Y 与 X_1、X_2 之间的线性关系较为合理。

[25]:
```
par(mfrow=c(2,2))
plot(fmbc)
```



从图中可以看出，无论是学生化残差的正态 Q-Q 图还是变换后因变量的拟合值都有明显的改观。

还可以进一步作正态性检验：

[19]:
```
ks.test(residuals(fmbc), "pnorm")
```

```
        One-sample Kolmogorov-Smirnov test

data:  residuals(fmbc)
D = 0.35625, p-value = 0.0005008
alternative hypothesis: two-sided
```

综上，认为数据满足正态性。Box-Cox 变换效果显著。

[29]:
```
detach(data)
```

6 习题 3.4

[1]:
```
raw_data <- read.csv("./ex_3_4.csv"); raw_data
```

催化剂		产	品	得	率	
	X.1	X.2	X.3	X.4	X.5	X.6
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
A1	0.88	0.85	0.79	0.86	0.85	0.83
A2	0.87	0.92	0.85	0.83	0.90	0.80
A3	0.84	0.78	0.81	0.80	0.85	0.83
A4	0.81	0.86	0.90	0.87	0.78	0.79

为方便后续分析，先将数据处理成 $< A, x >$ 的序对形式：

[2]:
```
A = c()
x = c()
for (row in 1:nrow(raw_data)) {
    for (col in 2:ncol(raw_data)) {
        A = c(A, raw_data[row, 1])
        x = c(x, raw_data[row, col])
    }
}
data = data.frame(A, x)
data
```

A data.frame: 24 × 2

A	x
<chr>	<dbl>
A1	0.88
A1	0.85
A1	0.79
⋮	⋮
A4	0.87
A4	0.78
A4	0.79

6.1 正态性检验

首先，第一步，需要对数据进行正态性检验：

用 shapiro.test 检验 4 组数据是否服从正态分布：

[3]:
```
shapiro.test(x[1:6])
shapiro.test(x[7:12])
shapiro.test(x[13:18])
shapiro.test(x[19:24 ])
```

```
        Shapiro-Wilk normality test

data:  x[1:6]
W = 0.92835, p-value = 0.5674


        Shapiro-Wilk normality test

data:  x[7:12]
W = 0.9831, p-value = 0.9659


        Shapiro-Wilk normality test

data:  x[13:18]
W = 0.96579, p-value = 0.8631


        Shapiro-Wilk normality test

data:  x[19:24]
W = 0.92097, p-value = 0.5124
```

A1、A2、A3、A4 四个组的 shapiro.test 检验 p 值依次为 0.5674、0.9659、0.8631、0.5124，均大于 0.05，认为原假设成立（H_0：假设数据服从正态分布），表明 4 组数据均来自正态分布总体。

6.2 方差齐次检验

接下来进行方差齐次检验：

用 bartlett.test 检验 4 个分组数据方差是否一致：

[4]:
```
bartlett.test(x ~ A, data = data)
```

```
        Bartlett test of homogeneity of variances

data:  x by A
Bartlett's K-squared = 2.2257, df = 3, p-value = 0.5269
```

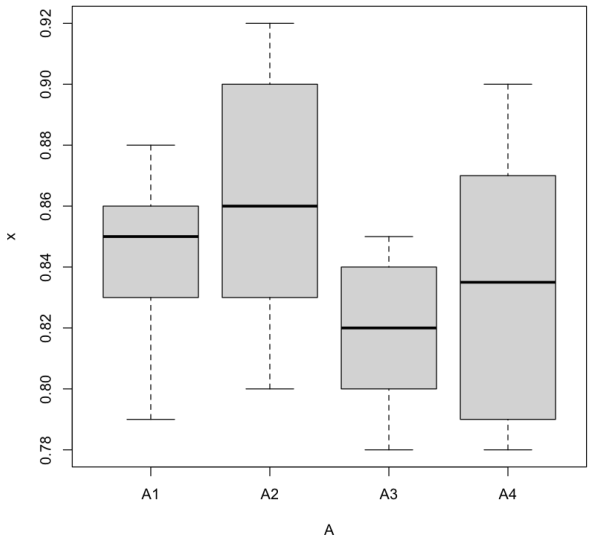
得到 p-value $= 0.5269 > 0.05$，认为原假设成立（H_0：假设 4 组数据方差相等），表明 4 组数据的方差齐次。

通过了上面两个检验，下面就可以开始做单因素方差分析了。

6.3 图形可视化

开始之前，还可以考虑先用箱图观察一下几组数据的分布：

[5]:
```
boxplot(x ~ A, data = data)
```



6.4 单因素方差分析

使用 aov 函数完成方差分析:

[6]:
```
dfc <- aov(x~A, data=data)
summary(dfc)
```

```
            Df   Sum Sq  Mean Sq F value Pr(>F)
A            3 0.005846 0.001949   1.306    0.3
Residuals   20 0.029850 0.001492
```

从方差分析表中看到 p-value $= 0.3 > 0.05$，接收原假设 H_0，认为四种不同催化剂对产品的得到率无显著影响。[?]

7 习题 3.5

[1]:
```
raw_data <- read.csv("./ex_3_5.csv"); raw_data
```

科研经费投入			生	产	能	力	提	高	量			
	x.0	x.1	x.2	x.3	x.4	x.5	x.6	x.7	x.8	x.9	x.10	x.11
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>
低	7.6	8.2	6.8	5.8	6.9	6.6	6.3	7.7	6			NA
中	6.7	8.1	9.4	8.6	7.8	7.7	8.9	7.9	8.3	8.7	7.1	8.4
高	8.5	9.7	10.1	7.8	9.6	9.5						NA

首先还是把数据处理成两列:

[2]:
```
t_raw_data = t(raw_data)  # 转置，方便取数据
dict = c("low", "mid", "high")
names(dict) <- c("低", "中", "高")
investment = c()   # 科研经费投入
improvement = c()  # 生产能力提高量
for (i in 1:3) {
    b = dict[[t_raw_data[[1,i]]]]   # 科研经费投入
    a = array(t_raw_data[-1,i])   # 生产能力提高量: 下面 2 行将数据转化为 dobule
↪型，并清除 NA、空值
    a = apply(a[!is.na(a)], 1, as.double)  # 这里会有一些产生 NA 的 Warning，这个
问题来自: as.double("")，不必在意
    a = a[!is.na(a)]
    for (p in a) {
        investment = c(investment, b)
```

```
        improvement = c(improvement, p)
    }
}
data = data.frame(investment, improvement);data
```

```
Warning message in apply(a[!is.na(a)], 1, as.double):
"强制改变过程中产生了 NA"
```

A data.frame: 27 × 2

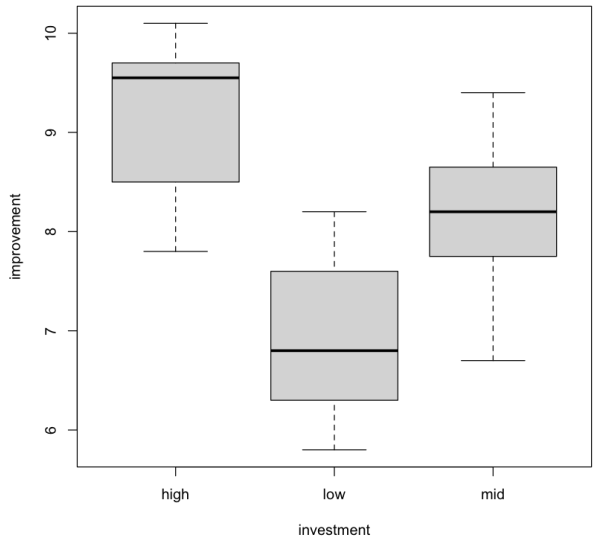
investment <chr>	improvement <dbl>
low	7.6
low	8.2
low	6.8
⋮	⋮
high	7.8
high	9.6
high	9.5

7.1 (1)

(1) 建立方差分析表,在显著水平 $\alpha = 0.05$ 下检验过去三年科研经费投入的不同是否对当年生产力的提高有显著影响.

首先，通过箱线图，直观感觉是有显著差异的:

[3]:
```
boxplot(improvement ~ investment, data = data)
```



接下来建立方差分析表:

[4]:
```
dfc <- aov(improvement ~ investment, data=data)
summary(dfc)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
investment   2  20.12   10.06   15.72 4.33e-05 ***
Residuals   24  15.36    0.64
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

检验 $p < 0.05$，拒绝原假设，认为在显著水平 $\alpha = 0.05$ 下过去三年科研经费投入的不同对当年生产力的提高有显著影响。

7.2 (2)

(2) 分别以 μ_L、μ_M 和 μ_H 记在过去三年科研经费投入为低、中、高情况下当年生产能力提高量的均值，分别给出 μ_L、μ_M 和 μ_H 的置信度为 95% 的置信区间以及差值 $\mu_L - \mu_M$、$\mu_L - \mu_H$ 和 $\mu_M - \mu_H$ 的置信度不小于 95% 的 Bonferroni 同时置信区间. 是否过去三年科研经费投入越高，当年生产能力的改善越显著?

首先为了方便，提取出几个组:

[5]:
```
group_low  <- data[investment=="low",]
group_mid  <- data[investment=="mid",]
```

```
group_high <- data[investment=="high",]
```

利用 t.test 容易获取到均值的置信区间 [?]:

[6]:
```
t.test(group_low$improvement)
```

```
        One Sample t-test

data:  group_low$improvement
t = 25.361, df = 8, p-value = 6.261e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 6.252390 7.503166
sample estimates:
mean of x
 6.877778
```

其中 mean of x 为均值，confidence interval 即置信区间。

为了方便，封装如下函数，用来提取这些信息:

[7]:
```
mean_confin <- function(x, ...) {    # mean and confidence interval of x by t.
 ↪test
    t.res <- t.test(x, ...)
    mean_val <- t.res$estimate[["mean of x"]]
    mean_conf.in <- t.res$conf.in
    res <- c(mean_val, mean_conf.in)
    names(res) <- c("mean", "conf.left", "conf.right")
    res  # ret
}
```

调用上面封装的函数可以很方便地得到一个表格:

[8]:
```
mu <- (function() {
    L <- mean_confin(group_low$improvement)
    M <- mean_confin(group_mid$improvement)
    H <- mean_confin(group_high$improvement)
```

```
    rbind(L, M, H)
})()
mu
```

A matrix: 3 × 3 of type dbl

	mean	conf.left	conf.right
L	6.877778	6.252390	7.503166
M	8.133333	7.652239	8.614427
H	9.200000	8.289951	10.110049

得到三年经费投入为低、中、高情况下当年生产能力提高量的均值为:

$$\mu_L = 6.877778$$
$$\mu_M = 8.133333$$
$$\mu_H = 9.2$$

各 95% 置信区间如下:

$$\mu_L \in (6.252390, 7.503166)$$
$$\mu_M \in (7.652239, 8.614427)$$
$$\mu_H \in (8.289951, 10.110049)$$

当然，这里也可以手动实现计算过程，但比较麻烦:

```
# 分类汇总: means of improvements
grouped_means <- aggregate(improvement, by=list(investment), FUN=mean)
# 从方差分析表中提取出 MSE
mse <- anova(dfc)["Residuals", "Mean Sq"]
# 带入公式，用 qt 计算 t 分位数，得到置信区间
...
```

要求 Bonferroni 同时置信区间，R 好像没有内置的实现，同样手写麻烦，所以考虑调用第三方包 [?]:

[9]:
```
# install.packages("DescTools")
require(DescTools)
PostHocTest(dfc, method = "bonferroni")
```

```
Loading required package: DescTools
```

```
  Posthoc multiple comparisons of means : Bonferroni
    95% family-wise confidence level

$investment
              diff     lwr.ci      upr.ci    pval
low-high -2.322222 -3.4074430 -1.23700140 3.5e-05 ***
mid-high -1.066667 -2.0961975 -0.03713579  0.0405 *
mid-low   1.255556  0.3475947  2.16351644  0.0048 **

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

得到 $\mu_L - \mu_M$，$\mu_L - \mu_H$ 和 $\mu_M - \mu_H$ 的置信度不小于 95% 的 Bonferroni 同时置信区间：

$$\mu_L - \mu_M \in (-2.16351644, -0.3475947)$$
$$\mu_L - \mu_H \in (-3.4074430, -1.23700140)$$
$$\mu_M - \mu_H \in (-2.0961975, -0.03713579)$$

从 $\mu_L - \mu_M$，$\mu_L - \mu_H$ 和 $\mu_M - \mu_H$ 的 Bonferroni 同时置信区间都位于负值区间可知，随着三年科研经费的投入越高，当年生产能力的改善越显著。

8 习题 3.6

[1]:
```
data <- read.table("./ex_3_6.meaningfulize.txt", header=TRUE); data
```

A data.frame: 108 × 3

FeIon	Dose	Retention
<chr>	<chr>	<dbl>
Fe3	high	0.71
Fe3	high	1.66
Fe3	high	2.01
⋮	⋮	⋮
Fe2	low	19.87
Fe2	low	21.60
Fe2	low	22.25

[2]:
```
attach(data)
```

8.1 (1)

(1) 由 SAS 系统 proc anova 过程的"means"语句(或者其他方法)求出各组合水平上的观测值的样本均值和标准差,各水平组合上的标准差(从而样本方差)差异是否明显?你认为假定误差的等方差性是否合理?

首先，求出各组合观测值的样本均值、标准差。这里可以利用 aggregate 做分类汇总:

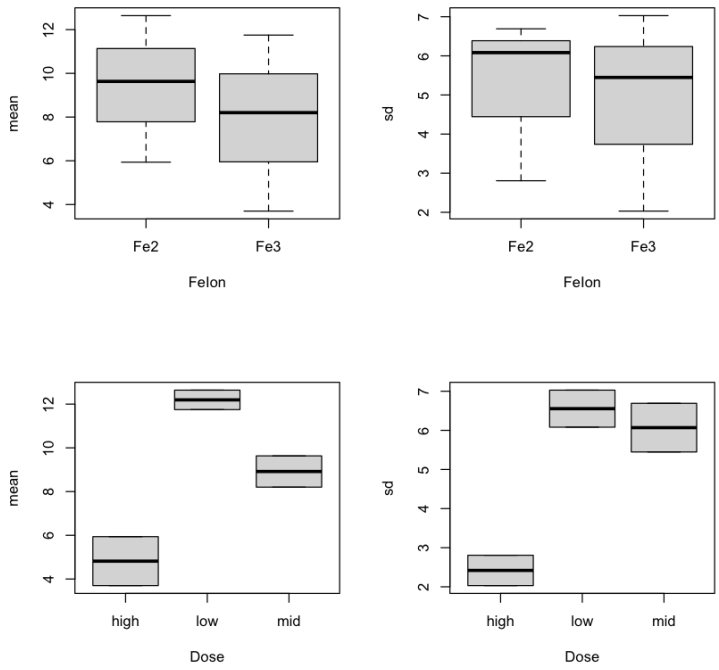
[3]:
```
grouped_means <- aggregate(Retention, by=list(FeIon, Dose), FUN=mean)
grouped_sds  <- aggregate(Retention, by=list(FeIon, Dose), FUN=sd)
# 下面几行代码将结果整合到一个表格，方便查看:
grouped_means_sds <- cbind(grouped_means, grouped_sds["x"])
names(grouped_means_sds) <- c("FeIon", "Dose", "mean", "sd")
grouped_means_sds
```

A data.frame: 6 × 4

FeIon <chr>	Dose <chr>	mean <dbl>	sd <dbl>
Fe2	high	5.936667	2.806778
Fe3	high	3.698889	2.030870
Fe2	low	12.639444	6.082089
Fe3	low	11.750000	7.028150
Fe2	mid	9.632222	6.691215
Fe3	mid	8.203889	5.447386

为方便观察，可以画出箱线图来比较:

[6]:
```
par(mfrow=c(2,2))
boxplot(`mean` ~ `FeIon`, data=grouped_means_sds)
boxplot(`sd` ~ `FeIon`, data=grouped_means_sds)
boxplot(`mean` ~ `Dose`, data=grouped_means_sds)
boxplot(`sd` ~ `Dose`, data=grouped_means_sds)
```



从比较结果来看，高剂量组标准差明显异于其他两组，认为假定误差的等方差性不太合理。所以不能直接进行方差分析。

8.2 (2)

(2) 对观测数据作自然对数变换，再进行(1)中分析. 此时，各组合水平上的标准差是否趋于一致？

自然对数变换，把变换后的数据列叫做 `lnRetention`:

[7]:
```
lnRetention <- log(Retention)
data <- cbind(data, lnRetention)
```

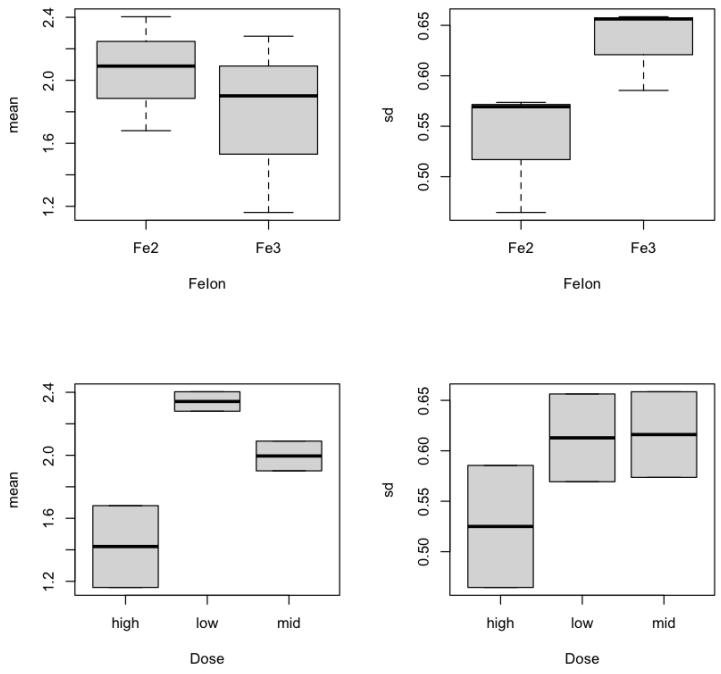
计算变换后的分组均值、标准差:

[8]:
```
grouped_ln_means <- aggregate(lnRetention, by=list(FeIon, Dose), FUN=mean)
grouped_ln_sds   <- aggregate(lnRetention, by=list(FeIon, Dose), FUN=sd)
# 下面省略和前面类似的代码，将结果整合到一个表格，方便查看:
...
```

A data.frame: 6 × 4

FeIon	Dose	mean	sd
<chr>	<chr>	<dbl>	<dbl>
Fe2	high	1.680129	0.4645464
Fe3	high	1.160924	0.5854773
Fe2	low	2.403389	0.5693701
Fe3	low	2.279981	0.6563113
Fe2	mid	2.090045	0.5736511
Fe3	mid	1.901225	0.6585116

作图比较（代码和前面类似，省略了）：



可以看到，现在个组标准差趋于一致，各族间标准差差异不大。可以利用变换之后的数据进行方差分析了。

8.3 (3)

(3) 对变换后的数据进行方差分析，建立方差分析表. 在显著水平 $\alpha = 0.05$ 下，因素的交互效应是否显著？各因素的影响是否显著？

用变换后的数据，重新做 aov：

[10]:
```
lnRetention.aov <- aov(lnRetention ~ FeIon + Dose + FeIon:Dose, data=data)
summary(lnRetention.aov)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
FeIon         1   2.07   2.074   5.993   0.0161 *
Dose          2  15.59   7.794  22.524 7.91e-09 ***f
FeIon:Dose    2   0.81   0.405   1.171   0.3143
Residuals   102  35.30   0.346
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

从结果中可以看出，在显著水平 $\alpha = 0.05$ 下，铁离子种类因素（Fe^{2+}、Fe^{3+}）和剂量因素（剂量低、中、高）对存留量的影响均显著（检验 p 值都小于 0.05）。即说明两种铁离子存留量是有显著差异的，不同剂量水平下存留量也是有显著差异的。

同时，可以看到在该水平下，交叉因子（FeIon:Dose 项）对存留量影响不显著（$p = 0.3143 > 0.05$），认为两种铁离子存留量在不同剂量水平下可认为是相同的。

8.4 (4)

(4) 根据(3)中分析，分别求各因素在其不同水平上的均值的置信度为95% 置信区间以及两两均值之差的置信度不小于 95% 的 Bonferroni 同时置信区间，并解释其结果.

先求各因素在不同水平下的均值以及估计区间，可以复用 3.5 中封装的 mean_confin 函数。

按离子:

[12]:
```
(function() {
    Fe2 <- mean_confin(data[FeIon=="Fe2",]$lnRetention)
    Fe3 <- mean_confin(data[FeIon=="Fe3",]$lnRetention)
    rbind(Fe2, Fe3)
})()
```

A matrix: 2 × 3 of type dbl

	mean	conf.left	conf.right
Fe2	2.057854	1.892251	2.223458
Fe3	1.780710	1.568011	1.993409

按剂量:

[13]:
```
(function() {
    Low  <- mean_confin(data[Dose=="low",]$lnRetention)
    Mid  <- mean_confin(data[Dose=="mid",]$lnRetention)
    High <- mean_confin(data[Dose=="high",]$lnRetention)
    rbind(Low, Mid, High)
})()
```

A matrix: 3 × 3 of type dbl

	mean	conf.left	conf.right
Low	2.341685	2.135709	2.547661
Mid	1.995635	1.787163	2.204107
High	1.420526	1.223052	1.618001

利用 DescTools 包，求 Bonferroni 同时置信区间：[?]

[14]:
```
# install.packages("DescTools")
library(DescTools)
PostHocTest(lnRetention.aov, method = "bonferroni")
```

```
  Posthoc multiple comparisons of means : Bonferroni
    95% family-wise confidence level

$FeIon
              diff     lwr.ci      upr.ci   pval
Fe3-Fe2 -0.2771441 -0.5016931 -0.05259515 0.0161 *

$Dose
               diff     lwr.ci       upr.ci    pval
low-high  0.9211588  0.5836659  1.258651627 4.6e-09 ***
mid-high  0.5751084  0.2376156  0.912601307 0.00021 ***
mid-low  -0.3460503 -0.6835432 -0.008557451 0.04251 *

$`FeIon:Dose`
                         diff      lwr.ci     upr.ci    pval
Fe3:high-Fe2:high -0.5192055 -1.10861287 0.07020194  0.1408
Fe2:low-Fe2:high   0.7232596  0.13385220 1.31266701  0.0055 **
Fe3:low-Fe2:high   0.5998524  0.01044505 1.18925985  0.0425 *
Fe2:mid-Fe2:high   0.4099156 -0.17949182 0.99932299  0.5859
```

```
Fe3:mid-Fe2:high   0.2210958 -0.36831157 0.81050323  1.0000
Fe2:low-Fe3:high   1.2424651  0.65305767 1.83187247 9.7e-08 ***
Fe3:low-Fe3:high   1.1190579  0.52965051 1.70846531 1.7e-06 ***
Fe2:mid-Fe3:high   0.9291210  0.33971365 1.51852845  0.0001 ***
Fe3:mid-Fe3:high   0.7403013  0.15089389 1.32970869  0.0040 **
Fe3:low-Fe2:low   -0.1234072 -0.71281456 0.46600024  1.0000
Fe2:mid-Fe2:low   -0.3133440 -0.90275142 0.27606338  1.0000
Fe3:mid-Fe2:low   -0.5021638 -1.09157118 0.08724362  0.1785
Fe2:mid-Fe3:low   -0.1899369 -0.77934426 0.39947054  1.0000
Fe3:mid-Fe3:low   -0.3787566 -0.96816402 0.21065078  0.8427
Fe3:mid-Fe2:mid   -0.1888198 -0.77822716 0.40058764  1.0000

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

参考文献

[1] 萌弟. R 语言实战之回归分析 [EB/OL]. (2020-08-15)[2021-05-30]. https://zhuanlan.zhihu.com/p/184923047

[2] fitzgerald0. R 中的 Box-Cox 变换 [EB/OL]. (2017-07-16)[2021-06-1]. https://blog.csdn.net/fitzgerald0/article/details/75212215

[3] 吴健. 基于 R 语言进行 Box-Cox 变换 [EB/OL]. (2018-11-19)[2021-06-01]. https://ask.hellobi.com/blog/R_shequ/18371

[4] jinzhao. 如何数据正态化 Box-Cox 变换 [EB/OL]. (2018-07-19)[2021-06-02]. https://zhuanlan.zhihu.com/p/40125782

[5] 数据小兵. 用 R 语言做单因素方差分析及多重比较 [EB/OL]. (2019-07-05)[2021-06-03]. http://www.datasoldier.net/archives/1315

[6] 刘小芬. R 语言计算置信区间 [EB/OL]. (2018-04-15)[2021-06-04]. https://zhuanlan.zhihu.com/p/35713329

[7] Maurits Evers. Bonferroni Simultaneous Confidence Intervals of differences in means [EB/OL]. (2021-02-18)[2021-06-04]. https://stackoverflow.com/questions/48572619/bonferroni-simultaneous-confidence-intervals-of-differences-in-means

[8] psych. psych: Procedures for Psychological, Psychometric, and Personality Research[EB/OL]. (2021-03-27)[2021-05-31]. https://cran.r-project.org/web/packages/psych/index.html

[9] 梅长林范金城. 数据分析方法 [M]. 高等教育出版社, 2006.

[10] R-project. An Introduction to R [M/OL]. https://cran.r-project.org/doc/manuals/r-release/R-intro.html