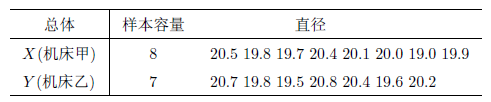
***任务1***

* 甲、乙两台机床分别加工某种轴承, 轴承的直径分别服从正态分布N(μ1，σ12)和N(μ2，σ22)，从各自加工的轴承中分别抽取若干个轴承测其直径, 结果如下表所示. 设σ12= σ22,问两台机床的加工精度有无显著差异?(取α=0.05)
* 要求：（1）判断使用哪种分析方法；（2）给出相关R语言实现；（3）给出R代码执行输出并解释分析结果。



1. 我们应该使用(two-sample)t-test，并且

H0: μX =μY

HA: μX≠μY

1. 代码：
2. #We may use t test to test if two samples are statistically different with each other
3. X = c(20.5, 19.8, 19.7, 20.4, 20.1, 20, 19.0, 19.9)
4. Y = c(20.7, 19.8, 19.5, 20.8, 20.4, 19.6, 20.2)
5. t\_test = t.test(X, Y, pair = FALSE)

3. 结果

|  |
| --- |
| Welch Two Sample t-test  data: X and Y  t = -0.84779, df = 12.187, p-value = 0.4129  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -0.7767939 0.3410796  sample estimates:  mean of x mean of y  19.92500 20.14286 |

根据以上结果，p值远大于我们所设定的alpha值0.05，统计不显着。故我们没有足够证据拒绝H0假设。所以X(机床甲)的平均直径和Y(机床乙)的平均直径，在统计上并没有明显不同。

***任务2***

* R自带的iris数据集中包含了150朵鸢尾花的数据，包含的字段有Sepa.Length(花萼长度)、Sepal.Width(花萼宽度)、Petal.Length(花瓣长度)、Petal.Width(花瓣宽度)以及Species(品种)。尝试分析Sepal.Length(花萼长度)、Sepal.Width(花萼宽度)、Petal.Length(花瓣长度)、Petal.Width(花瓣宽度)之间的相关性。
* 要求使用R语言完成以下分析任务：
* （1）计算相关系数矩阵；
* （2）绘制相关矩阵图；
* （3）对相关系数进行显著性检验
* （4）解释变量之间的相关性结论。

(1)

###Task 2 Correlation

library(igraph)

library(psych)

library(corrplot)

#See structure of data frame

str(iris)

head(iris, n=10)

#We only need to analyze column 1 to column 4

iris2 = iris[ ,1:4]

correlation = cor(iris2)

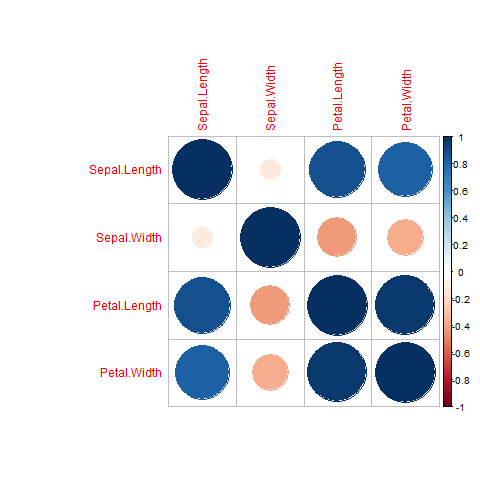
corrplot(correlation)

(2)

a.

|  |
| --- |
| Sepal.Length Sepal.Width Petal.Length Petal.Width  Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411  Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259  Petal.Length 0.8717538 -0.4284401 1.0000000 0.9628654  Petal.Width 0.8179411 -0.3661259 0.9628654 1.0000000 |

b. 以色彩表示



(3) The result of correlation test is shown below:

|  |
| --- |
| Probability values (Entries above the diagonal are adjusted for multiple tests.)  Sepal.Length Sepal.Width Petal.Length Petal.Width  Sepal.Length 0.00 0.15 0 0  Sepal.Width 0.15 0.00 0 0  Petal.Length 0.00 0.00 0 0  Petal.Width 0.00 0.00 0 0 |

(4)

Sepal length和Petal Length以及Petal width高度正相关，相关係数分别为0.872和0.817(经进位)，统计显着。表示sepal lenth的值越大，相应的petal length和petal width的值就会越大。

Sepal width和Petal length以及Petal width之间存在低度负相关，相关係数分别为-0.428与-0.366(经进位)，统计显着。表示sepal width的值越大，相应的petal length和petal width的值就会越大。

Petal length和Petal width之间存在高度正相关，相关係数为0.963(经进位)，统计显着。表示petal lenth的值越大，petal width的值就会越大。

***任务3***

* 在“lm.csv”文件中保存了145家某时期某国电力企业的总成本（TC）、产量（Q）、工资率（PL）、燃料价格（PF）及资本租赁价格（PK）的数据。尝试以总成本为因变量，其他变量为自变量，使用多重线性回归方法研究总成本与其他变量之间的关系。使用R语言完成以下任务：
* （1）构建使用所有自变量的多重线性回归模型，根据模型信息解释这个模型是否合适；
* （2）尝试以Akaike信息统计量为准则，使用逐步回归法选择合适的变量构建多重线性回归模型，解释选择的依据，并说明这个模型是否合适。

1. 在R语言中建立多元线性模型，代码如下:

###Task 3 Regression model

data = read.csv("D:\\lm.csv")

str(data)

head(data, n=10) #See the structure of data

y = data$TC

x1 = data$Q

x2 = data$PL

x3 = data$PF

x4 = data$PK

reg = lm(y~x1+x2+x3+x4, data = data)

summary(reg)

根据此代码，我们得到结果:

|  |
| --- |
| Call:  lm(formula = y ~ x1 + x2 + x3 + x4, data = data)  Residuals:  Min 1Q Median 3Q Max  -17.814 -1.609 -0.092 2.231 43.761  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -2.222e+01 6.587e+00 -3.373 0.000961 \*\*\*  x1 6.395e-03 1.629e-04 39.258 < 2e-16 \*\*\*  x2 5.655e+00 2.176e+00 2.598 0.010366 \*  x3 2.078e-01 6.410e-02 3.242 0.001482 \*\*  x4 2.844e-02 2.650e-02 1.073 0.285088  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 5.579 on 140 degrees of freedom  Multiple R-squared: 0.9228, Adjusted R-squared: 0.9206  F-statistic: 418.1 on 4 and 140 DF, p-value: < 2.2e-16 |

报表中，我们发现变量x4(PK)没有通过显着性检验，因此统计上它并不能够成为我们多元迴归的其中解释变量。此模型并不合适。

(2)

我们想透过逐步迴归，了解哪些是真正能够适当解释应变量的变数，使用代码如下:

AIC = step(reg)

结果如下:

|  |
| --- |
| Start: AIC=502.63  y ~ x1 + x2 + x3  Df Sum of Sq RSS AIC  <none> 4394 502.63  - x2 1 181 4575 506.49  - x3 1 388 4782 512.90  - x1 1 48338 52732 860.96  > AIC = step(reg2)  Start: AIC=502.63  y ~ x1 + x2 + x3  Df Sum of Sq RSS AIC  <none> 4394 502.63  - x2 1 181 4575 506.49  - x3 1 388 4782 512.90  - x1 1 48338 52732 860.96 |

我们得知，移除x4可以使得AIC指标变小，x4如我们预期，是可以被移除的。同时我们也进一步知道剩下x1, x2和x3三个变量，能够作为此多元迴归的解释变数(删掉任何一个都将使AIC值上升)。

新的迴归测量代码如下:

#We think x4 caanot explain y properly, so we remove it

y = data$TC

x1 = data$Q

x2 = data$PL

x3 = data$PF

reg2 = lm(y~x1+x2+x3, data = data)

summary(reg2)

结果如下:

|  |
| --- |
| Call:  lm(formula = y ~ x1 + x2 + x3, data = data)  Residuals:  Min 1Q Median 3Q Max  -17.290 -1.503 -0.385 2.179 44.779  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -1.654e+01 3.928e+00 -4.212 4.48e-05 \*\*\*  x1 6.406e-03 1.627e-04 39.384 < 2e-16 \*\*\*  x2 5.098e+00 2.115e+00 2.411 0.017208 \*  x3 2.217e-01 6.283e-02 3.528 0.000565 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 5.582 on 141 degrees of freedom  Multiple R-squared: 0.9221, Adjusted R-squared: 0.9205  F-statistic: 556.5 on 3 and 141 DF, p-value: < 2.2e-16 |

所有的变数都通过显着性检验，调整后R平方也高达0.9205，经过调整后我们得到一个适配的模型。