**成都信息工程大学计算机学院**

**数据统计与分析实验报告**

“基于R语言的线性相关分析和

一元线性回归模型”实验

实验三 “基于R语言的线性相关分析和

一元线性回归模型”实验

**一、实验目的**

1. 掌握R语言进行线性相关分析的方法。

2. 掌握R语言建立一元线性回归模型的方法。

**二、实验原理**

1. 线性相关分析

线性相关分析是用相关系数来表示两个变量间相互的线性关系，并判断其密切程度的统计方法。Pearson相关系数用来反映两个变量的线性相关关系，其定义公式是：



Pearson相关系数的取值范围是[-1, 1]，当-1<r<0时，表示具有负线性相关，越接近-1，负相关性越强。0<r<1，表示具有正线性相关，越接近1，正相关性越强。r=-1表示完全负线性相关，r=1表示具有完全正线性相关，r=0表示两个变量不具有线性相关性。

相关系数的显著性可用t检验来进行。

2. 一元线性回归模型

给定一组数据点(x1, y1)、(x2, y2)、...、(xn, yn)，如果通过散点图可以观察出变量间大致存在线性函数关系，则可以建立如下模型：



其中a,b称为一元线性回归的回归系数；ε表示回归值与测量值之间的误差。已知(x1, y1),(x2 ,y2),...,(xn, yn)，代入回归模型得到：



利用偏导数，可得



即：

1. **实验内容**

1. 线性相关分析

教材152页，习题1(1)、习题1(2)，习题2(1)、习题2(2)、习题2(3)、习题2(4)。

答：习题1(1)判断speed与dist的大致线性相关性。

|  |  |
| --- | --- |
| **题目要求** | **作speed与dist的散点图。并判断线性关系。** |
| 代码 | *speed = c(4,4,7,7,8,9)*  *dist = c(2,10,4,22,16,10)*  *plot(dist,speed)*  *plot(speed,dist)* |
| 散点图 |  |
| 结论 | speed与dist之间不大致呈线性关系 |

答：习题1(2) 计算speed与dist的相关系数并做假设检验。

|  |  |
| --- | --- |
| **题目要求** | **计算speed与dist的相关系数并做假设检验。** |
| 代码 | *cor(dist,speed)*  *cor(speed,dist)*  *cor.test(speed,dist)* |
| 运行结果 | [1] 0.4144095  [1] 0.4144095  Pearson's product-moment correlation  data: speed and dist  t = 0.9107, df = 4, p-value = 0.414  alternative hypothesis: true correlation is not equal to 0  95 percent confidence interval:  -0.5984074 0.9174239  sample estimates:  cor  0.4144095 |
| 结论 | 相关度较强 |

答：习题2 (1) 作x与y的散点图。

|  |  |
| --- | --- |
| **题目要求** | **作x与y的散点图。并判断线性关系。** |
| 代码 | *x = c(0.10, 0.11, 0.12, 0.13, 0.14 ,0.15 ,0.16, 0.17, 0.18, 0.20, 0.21, 0.23)*  *y = c(42, 43.5, 45, 45.5, 45, 47.5, 49, 53, 50, 55, 55, 60)*  *plot(x,y)* |
| 散点图 |  |
| 结论 | 大致呈线性关系 |

答：习题2 (2)计算x与y的相关系数并做假设检验。

|  |  |
| --- | --- |
| **题目要求** | **计算x与y的相关系数并做假设检验。** |
| 代码 | *cor(x,y)*  *cor(y,x)*  *cor.test(x,y)* |
| 运行结果 | [1] 0.9736872  [1] 0.9736872  Pearson's product-moment correlation  data: x and y  t = 13.511, df = 10, p-value = 9.505e-08  alternative hypothesis: true correlation is not equal to 0  95 percent confidence interval:  0.9061336 0.9928074  sample estimates:  cor  0.9736872 |
| 结论 | 相关程度较强 |

答：习题2（3）做y对x的最小二乘回归，并给出常用统计量。

|  |  |
| --- | --- |
| **题目要求** | **做y对x的最小二乘回归。** |
| 代码 | *model = lm(x~y)*  *plot(x~y)*  *abline(model)* |
| 运行结果 | Call:  lm(formula = y ~ x)  Coefficients:  (Intercept) x  28.49 130.83 |
| y对x的最小二乘回归方程 |  |
| **题目要求** | **给出常用统计量。** |
| 代码 | mean(x)  mean(y)  sd(x)  sd(y)  var(x)  var(y) |
| 运行结果 | [1] 0.1583333  [1] 49.20833  [1] 0.0410838  [1] 5.520451  [1] 0.001687879  [1] 30.47538 |

答：习题2（4）根据回归方程，进行预测。

|  |  |
| --- | --- |
| **题目要求** | **预测x=0.22时，y的值。** |
| 代码 | *predict(model,data.frame(x=0.22))* |
| 运行结果 | 57.27648 |
| **题目要求** | **预测x=0.25时，y的值。** |
| 代码 | *predict(model,data.frame(x=0.25))* |
| 运行结果 | 61.20153 |

2. 一元线性模型建立

表中是某地搜集到的新房屋的销售价格Y和房屋面积X的数据：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 房屋面积 X | 115 | 110 | 80 | 135 | 105 |
| 销售价格(万元) Y | 24.8 | 21.6 | 18.4 | 29.2 | 22 |

试求：Y对X的经验回归直线方程，并求这两个变量之间的线性相关系数。

答：求Y对X的经验回归直线方程，并求这两个变量之间的线性相关系数。

|  |  |
| --- | --- |
| **题目要求** | **求Y对X的经验回归直线方程。** |
| 代码 | *x = c(115,110,80,135,105)*  *y = c(24.8,21.6,18.4,29.2,22)*  *lm(y~x)* |
| 运行结果 | Call:  lm(formula = y ~ x)  Coefficients:  (Intercept) x  1.8166 0.1962 |
| Y对X的经验回归直线方程 |  |
| **题目要求** | **求这两个变量Y与X之间的线性相关系数。** |
| 代码 | cor(y,x) |
| 运行结果 | [1] 0.9597298 |
| 线性相关系数 | 0.9597298 |

**四、实验组织运行要求**

以学生自主训练为主的开放模式组织教学。

**五、思考题**

如何自己编程，用R语言实现一元线性模型的参数检验？

Example <- function(x, y) {

plot(x,y)

abline(x,y)

lm(y~x)

cor(y~x)

Cor.test(y~x)

}

**六、实验总结**

R 语言是为数学研究工作者设计的一种数学编程语言，主要用于统计分析、绘图、数据挖掘。如果你是一个计算机程序的初学者并且急切地想了解计算机的通用编程，R 语言不是一个很理想的选择，可以选择 Python、C 或 Java。R 语言与 C 语言都是贝尔实验室的研究成果，但两者有不同的侧重领域，R 语言是一种解释型的面向数学理论研究工作者的语言，而 C 语言是为计算机软件工程师设计的。R 语言是解释运行的语言（与 C 语言的编译运行不同），它的执行速度比 C 语言慢得多，不利于优化。但它在语法层面提供了更加丰富的数据结构操作并且能够十分方便地输出文字和图形信息，所以它广泛应用于数学尤其是统计学领域。虽然 R 主要用于统计分析或者开发统计相关的软件，但也有人用作矩阵计算。其分析速度可媲美专用于矩阵计算的自由软件 GNU Octave 和商业软件 MATLAB。