# Ch6相关分析与回归分析

本章概要

**♢相关分析**

**♢线性回归分析**

**♢logistic回归**

# 6.1相关分析（教材229页9.1）

## 6.1.1相关性概念

变量间的关系有两种类型：

* 确定性的函数关系y=f(x)
* 不完全确定的依存关系——相关关系

例1：（教材P230例9.1.1）

某医生测定了10名孕妇的15-17周分娩时脐带血及母血TSH的水平值

x<-c(1.21,1.30,1.39,1.42,1.47,1.56,1.68,1.72,1.98,2.10)

y<-c(3.90,4.50,4.20,4.83,4.16,4.93,4.32,4.99,4.70,5.20)

level<-data.frame(x,y)

plot(level) #plot(x,y)

## 6.1.2相关分析

### 1.原理简介

设有两个随机变量X与Y的观测值为,则它们之间的样本相关系数为



当n充分大时，样本相关系数作为总体相关系数



的估计。当|r|->1时，表明两者间有较强线性关系，当|r|->0时，表明两者间几乎无线性关系，r>0为正相关，r<0为负相关。

若（X，Y）服从二元正态分布，则

  
相关检验：HO:ρ=0，若T>t1-α(n-2)，则认为两者间存在显著的线性相关性。还可以根据spearman秩相关系数和 kendall相关系数进行相应的spearman相关检验和kendall相关检验。

### 2.R软件的实现

* 函数cor.test的调用格式1：

cor.test(x, y,

alternative = c("two.sided", "less", "greater"),

method = c("pearson", "kendall", "spearman"),

exact = NULL, conf.level = 0.95, continuity = FALSE, ...)

说明： x,y是长度相同的向量，alternative是备择假设，默认值为“two.sided”,method为选择检验方法，默认pearson检验，conf.level为置信水平，默认0.95。

* 函数cor.test的调用格式2：

cor.test(formula, data, subset, na.action, ...)

说明：formula是公式，形如“u+v”,“u”， “v” ,它们必须是相同长度的数值向量，data是数据框，subset是可选择向量，表示观测子集。

**例1续：进行相关分析**

attach(level)

cor.test(x,y)

输出： Pearson's product-moment correlation

data: x and y

t = 2.628, df = 8, p-value = 0.03025

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.08943 0.91723

sample estimates:

cor

0.6807

结果分析：相关检验的t值为2.628，检验的p值为0.03025,取显著性水平***α***=0.05，则p<***α***,所以拒绝原假设，从而认为脐带血及母血TSH的水平值间有显著线性关系。Pearson相关系数r2=0.6807,说明两者线性相关程度较强。

# 6.2回归分析（教材233页9.3）

相关分析只能得出两个变量间是否相关，但不能回答两个变量间存在何种具体的因果关系的函数关系。回归分析可以解决这一问题。

回归分析是研究变量之间的依存关系。如果因变量Y和自变量（或称为解释变量）X呈直线关系时，称直线回归，直线回归要求变量Y服从正态分布且方差相等。当变量间不是线性关系时，通常需要进行数据变换，再进行线性回归分析，或直接用原数据进行非线性回归。根据资料类型，各类回归分析：

|  |  |  |  |
| --- | --- | --- | --- |
| 实现函数 | 回归类型 | 资料类型 | |
|  |  | 因变量 | 自变量 |
| Reg | 线性回归 | 数值变量 | 数值变量 |
| glm | 协方差模型、一般线性模型 | 数值变量 | 数值变量、分类变量 |
| glm | Logistic回归 | 分类变量 | 数值变量、分类变量 |
| glm | Logistic\Poisson回归 | 分类变量 | 数值变量、分类变量 |
| nls | 非线性回归(不可线性化) | 数值变量 | 数值变量 |

### 1背景概述

**多元线性回归分析简介**

假定因变量y与k个解释变量x1,x2,…,xk具有线性关系，即

**总体回归模型**： ， 

或 

**样本回归模型**：

残差：

**最小二乘法**： ，矩阵形式

**总离差平方和的分解**式：

总平方和=解释平方和+残差平方和

TSS=ESS + RSS 即

**判定系数**： 

**回归模型的显著性检验（F检验）**：Ho：b1=b2=…=bk=0

H1：b1，b2，…，bk至少有一个不为0。

检验统计量

**变量（如xi）显著性检验（t检验）**：H0：bi=0，H1：bi≠0

检验统计量

### 2.R软件的实现

函数lm（）的调用格式：

lm(formula, data, subset, weights, na.action,

method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,

singular.ok = TRUE, contrasts = NULL, offset, ...)

说明：formula是设置回归模型，data是数据框, subset是样本观察的子集，weight是用于拟合的加权向量,no.action用于显示数据是否包含缺失值，method用于选择拟合方法，model,x,y，qr是逻辑值，用于是否返回相应值。

注：除了formula 是必选项，其它可选。

相关函数有：

summary():汇总回归计算结果

confint()：求参数的置信区间

anova():返回方差分析表

fitted():返回因变量的拟合值

predict():返回因变量的预测值

step():逐步回归优选回归模型

例2：教材P245例9.3.1

某公司在各地区销售一种特殊的化妆品，该公司观测了15个城市在某月对该化妆品的销售量y，使用该化妆品的人数x1和人均收入，数据如表。试建立y与x1,x2的线性回归方程，并作相应的检验。

解：1）先输入数据

y<-c(162,120,223,131,67,169,81,192,116,56,252,232,144,103,212)

x1<-c(274,180,375,205,86,265,98,330,195,53,430,372,236,157,370)

x2<-c(2450,3250,3802,2838,2347,3782,3008,2450,2137,2560,4020,4427,2660,2088,2605)

sales<-data.frame(y,x1,x2)

2）再观察变量间大致线性关系

par(mfrow=c(2,1))

plot(y,x1);plot(y,x2)

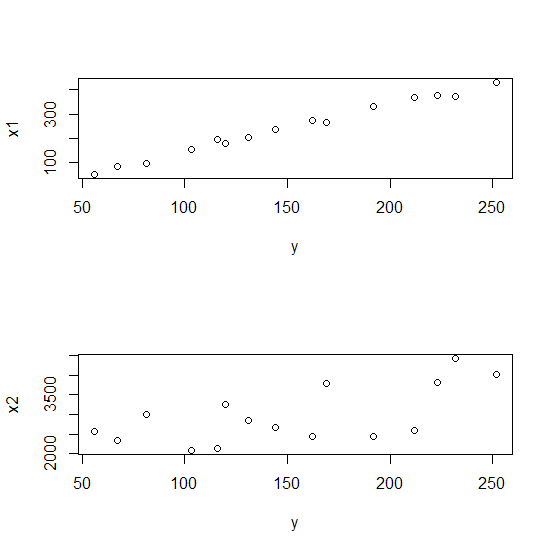
cor.test(y,x1);cor.test(y,x2)

3）建立回归模型

lm.reg<-lm(y~1+x1+x2,data=sales)

lm.reg

summary(lm.reg)

输出： > cor.test(y,x1)

Pearson's product-moment correlation

data: y and x1

t = 37.42, df = 13, p-value = 1.266e-14

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.9858 0.9985

sample estimates:

cor

0.9954

> cor.test(y,x2)

Pearson's product-moment correlation

data: y and x2

t = 3.004, df = 13, p-value = 0.01016

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.1902 0.8678

sample estimates:

cor

0.6401

> lm.reg

Call:

lm(formula = y ~ 1 + x1 + x2, data = sales)

Coefficients:

(Intercept) x1 x2

3.63897 0.49479 0.00926

> summary(lm.reg)

Call:

lm(formula = y ~ 1 + x1 + x2, data = sales)

Residuals:

Min 1Q Median 3Q Max

-3.909 -1.212 -0.347 1.758 3.311

Coefficients:

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

(Intercept) 3.63897 2.51372 1.45 0.17

x1 0.49479 0.00626 79.01 < 2e-16 \*\*\*

x2 0.00926 0.00100 9.24 8.3e-07 \*\*\*

---

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

Residual standard error: 2.25 on 12 degrees of freedom

Multiple R-squared: 0.999, Adjusted R-squared: 0.999

F-statistic: 5.29e+03 on 2 and 12 DF, p-value: <2e-16

结果分析：

由散点图和相关分析可知y和x1，x2间有较强的线性关系。

（1）回归方程为：

判定系数： ，说明因变量Y的变异中由模型能解释的部分占到99.99%，模型拟合效果很好。

（2）回归模型的显著性检验

检验的零假设Ho：b1=b2=0，由输出结果的方差分析表中F统计量为5290，检验的p值小于2\*10-16=0，在α =0.05的显著性水平下，p<α ,应拒绝H0,说明模型是显著成立的。

（3）变量的显著性检验

对变量x1而言，检验的零假设Ho：b1=0，由输出结果的参数估计部分的t统计量值为79.01，双边检验p=2\*10-16，所以在α =0.05的显著性水平下，应拒绝H0,说明x1变量对y变量有显著性影响。

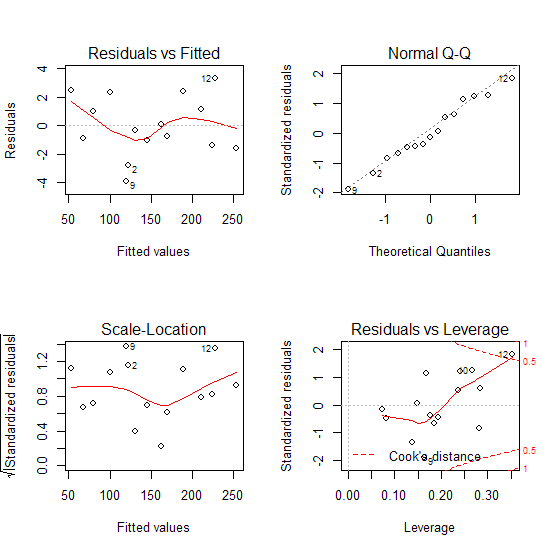
类似可得对变量x2，t统计量值为9.24，双边检验p值为p=8.3\*10-7，在α =0.05的显著性水平下，应拒绝H0,说明变量x2对变量y也有显著影响。

4）模型诊断图

plot(lm.reg) #四幅回归诊断图

par(mfrow=c(2,2))

plot(lm.reg)



5）模型优选：逐步回归

lm.step<-step(lm.reg)

输出：

> lm.step<-step(lm.reg)

Start: AIC=27

y ~ 1 + x1 + x2

Df Sum of Sq RSS AIC

<none> 61 27.000

- x2 1 433 494 56.417

- x1 1 31644 31705 118.843

分析：开始时包含全部变量x1和x2的模型，AIC统计量为27，去掉变量x2后AIC统计量升高为56.4，去年x1后AIC统计量升高为118.8。所以最优模型仍然为全部变量x1和x2的模型。

6）模型预测

预测：当x0=(x1=200,x2=300)时y的值，及其95%预测区间

exa<-data.frame(x1=200,x2=3000)

lm.pred<-predict(lm.reg,exa,interval="prediction",level=0.95)

lm.pred

输出：

fit lwr upr

1 130.4 125.3 135.5

结果分析：

=[125.3,135.5]

即点估计,95%的区间预测为 [125.3，135.5]

# 6.3Logistic回归

### 1.原理简介

线性回归中的响应变量要求是连续型的，但在实际研究中经常遇到非连续型（即分类型）的响应变量，常会采用Logisticl回归。

设响应变量Y有两个结果：成功（Y=1）和失败（Y=0），对响应变量Y有影响的p个自变量（或称为解释变量），记为X1,X2,…,Xp，在p个自变量的作用下出现“成功”的条件概率记为P=P(Y=1|X1,X2,…,Xp)，那么logistic回归模型为

（式1）

称为logistic回归模型的回归系数。

注：logistic 回归模型是一个非线性回归模型，自变量Xj(j=1,2,…,p)可以是连续型变量，也可以是分类变量或哑变量。在（－∞，＋∞）变化时，上面（式1）的比值总在（0，1）间变化，这正是概率P的取值范围。

对（式1）作logit变换，logistic回归模型可以变成下列线性形式：

(式2)

### 2.广义线性模型

Logistic回归模型属于广义线性模型（Generalized Linear Model）的一种，它是通常的正态线性模型的推广，它要求响应变量只能通过线性形式依赖于解释变量。下表是常见的广义线性模型的连接函数和误差函数类型：

|  |  |  |  |
| --- | --- | --- | --- |
| 变换 | 连接函数 | 回归模型 | 典型误差函数 |
| 恒等 |  | E(y)=x’B | 正态分布 |
| 对数 |  | Ln(E(y))=x’B | 泊松分布 |
| logit |  | logit(E(y))=x’B | 二项分布 |
| 逆（倒数） |  | 1/E(y)=x’B | 伽玛分布 |

### 3.R软件实现

* 拟合与计算广义线性模型的函数glm(),格式为：

glm(**formula, family = gaussian, data=data.frame**, weights, subset,

na.action, start = NULL, etastart, mustart, offset,

control = list(...), model = TRUE, method = "glm.fit",

x = FALSE, y = TRUE, contrasts = NULL, ...)

说明1：**formula为拟合公式， family 为分布族，可取gaussian（**正态分布），binomial(二项分布)，poission(泊松分布)，gama(伽玛分布)，分布族还可以通过link=来指定连接函数;**data为数据框。**

说明2：基于正态分布的广义线性模型，即为一般的线性模型，所以实现方式为

>glm(formula,family=Gaussian,data=data.frame)

等价于

>lm(formula,data=data.frame)

说明3：基于二项分布的广义线性模型，即为logistic回归模型，所以实现方式为

>glm(formula,family=binomial(link=logit),data=data.frame)

表达式中link=logit也可以不写，因为logit是二项分布族的连接函数，是默认状态。

**例3：教材P267 例9.5.1**

表9.10中是对45名驾驶员的调查结果，其中四个变量的含义为：

X1:表示视力状况，它是一个分类变量，1表示好，0表示有问题

X2:年龄（age），数值型

X3:驾车教育（drive），它也是一个分类变量，1表示参加过驾车教育，0表示没有参加过

Y:一个分类变量，表示去年是否出过事故，1表示出过，0表示没有出过

考察前三个变量X1,X2,X3与发生事故的关系。

解：1)先用数据框输入数据

#先用数据框输入数据

x1<-rep(c(1,0,1,0,1),c(5,10,10,10,10))

x2<-c(17,44,48,55,75,35,42,57,28,20,

38,45,47,52,55,68,18,68,48,17,

70,72,35,19,62,39,40,55,68,25,

17,45,44,67,55,61,19,69,23,19,

72,74,31,16,61)

x3<-c(1,0,1,0,1,0,1,0,0,0,1,0,1,0,0,

1,1,0,1,0,1,1,0,1,1,1,1,0,0,1,

0,0,0,0,0,1,1,0,1,0,1,1,0,1,1)

y<-c(1,0,0,0,1,1,1,0,1,1,0,1,1,0,1,

0,0,0,1,0,1,0,1,0,0,1,1,0,1,0,

0,1,1,0,1,0,0,0,1,0,1,0,1,0,0)

accident<-data.frame(x1,x2,x3,y)

2)再作logistic回归

#再作logistic回归

log.glm<-glm(y~x1+x2+x3,family=binomial,data=accident)

summary(log.glm)

输出：

Call:

glm(formula = y ~ x1 + x2 + x3, family = binomial, data = accident)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5636 -0.9131 -0.7892 0.9637 1.6000

Coefficients:

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

(Intercept) 0.597610 0.894831 0.668 0.5042

x1 -1.496084 0.704861 -2.123 0.0338 \*

x2 -0.001595 0.016758 -0.095 0.9242

x3 0.315865 0.701093 0.451 0.6523

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 62.183 on 44 degrees of freedom

Residual deviance: 57.026 on 41 degrees of freedom

AIC: 65.026

Number of Fisher Scoring iterations: 4

结果分析：初步的logistic回归模型：

3)再作模型诊断与更新：由于x2和x3没有通过检验，可用step（）作模型优选

log.step<-step(log.glm)

summary(log.step)

输出为：

> log.step<-step(log.glm)

Start: AIC=65.03

y ~ x1 + x2 + x3

Df Deviance AIC

- x2 1 57.035 63.035

- x3 1 57.232 63.232

<none> 57.026 65.026

- x1 1 61.936 67.936

Step: AIC=63.03

y ~ x1 + x3

Df Deviance AIC

- x3 1 57.241 61.241

<none> 57.035 63.035

- x1 1 61.991 65.991

Step: AIC=61.24

y ~ x1

Df Deviance AIC

<none> 57.241 61.241

- x1 1 62.183 64.183

> summary(log.step)

Call:

glm(formula = y ~ x1, family = binomial, data = accident)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.4490 -0.8782 -0.8782 0.9282 1.5096

Coefficients:

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

(Intercept) 0.6190 0.4688 1.320 0.1867

x1 -1.3728 0.6353 -2.161 0.0307 \*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 62.183 on 44 degrees of freedom

Residual deviance: 57.241 on 43 degrees of freedom

AIC: 61.241

Number of Fisher Scoring iterations: 4

结果分析：最后最优模型为

4)预测分析：计算视力好和视力差的人发生事故的概率

log.pre<-predict(log.step,data.frame(x1=1))

p1<-exp(log.pre)/(1+exp(log.pre));p1

log.pre<-predict(log.step,data.frame(x1=0))

p2<-exp(log.pre)/(1+exp(log.pre));p2

分析：p1=0.32;p2=0.65,说明视力差的司机发生交通事故的概率是视力好的人的两倍以上。

本节学习的R函数小结

|  |  |  |
| --- | --- | --- |
| 第5章 统计分析的主要函数 | | |
| 函数 | 功能 | 示例 |
|  |  |  |
|  |  |  |

**作业：练习本节所有命令**

* **9.1 9.6 9.9**