

第7章 回归分析

1. 简述回归分析研究的主要问题。

【参考答案】回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。

统计学中，回归分析(regression analysis)指的是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。回归分析按照涉及的变量的多少，可分为一元回归分析和多元回归分析；按照因变量的多少，可分为简单回归分析和多重回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。

回归分析研究的主要问题是：

- (1) 确定 Y 与 X 间的定量关系表达式，这种表达式称为回归方程；
- (2) 对求得的回归方程的可信度进行检验；
- (3) 判断自变量 X 对因变量 Y 有无影响；
- (4) 利用所求得的回归方程进行预测和控制。

2. 简述 pearson、spearman、kendall's tau 相关系数。

【参考答案】pearson 为积差相关系数，衡量两个定量/变量之间的线性相关程度。

spearman 为等级相关系数，衡量分级定序变量之间的相关程度。

kendall's tau 为相关系数等级相关度量。

3. 对 3 岁小朋友的身高和体重进行测试，测得 10 个数据见表 7-1。

表 7-1 身高和体重表

身高/cm	体重/kg	性别	身高/cm	体重/kg	性别
97	24	女	96	27	男
98	26	女	98	28	男
102	28	女	99	30	男
103	28	女	103	32	男
103	29	女	105	33	男

【参考答案】

(1) 建立身高和体重的回归方程式。

```
> 身高<-c(97,98,102,103,103,96,98,99,103,105)
> 体重<-c(24,26,28,28,29,27,28,30,32,33)
> hg<-data.frame(身高,体重)
> x<-lm(体重~身高,data=hg)
> x
Call:
lm(formula = 体重 ~ 身高, data = hg)
```

```
Coefficients:
(Intercept)      身高
-35.1018      0.6335
```

通过输出结果显示，可以得到回归方程式：

体重 = $-35.1018 + 0.6335 \times \text{身高}$

(2) 对回归方程式进行显著性检验。

```
> x<-lm(体重~身高,data=hg)
> summary(x)
```

```
Call:
lm(formula = 体重 ~ 身高, data = hg)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-2.34615 -1.42195  0.02036  1.51131  2.38688
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -35.1018    20.3488  -1.725   0.1228
身高         0.6335     0.2026   3.127   0.0141 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.905 on 8 degrees of freedom
Multiple R-squared:  0.55,    Adjusted R-squared:  0.4938
F-statistic: 9.778 on 1 and 8 DF,  p-value: 0.01408
```

(3) 预测小朋友身高在 104 cm 的体重及预测区间(置信水平 0.95)。

```
> z<-data.frame(身高=104)
> y<-predict(x,z,interval = "prediction",level = 0.95)
> y
      fit      lwr      upr
1 30.78054 25.87638 35.68471
```

4. R 语言数据集 women 提供了 15 个年龄在 30 至 39 岁间女性的身高和体重信息，通过身高来预测体重。

【参考答案】

(1) 做简单线性回归分析。

```
> x<-lm(weight~height,data=women)
> summary(x)
```

```
Call:
lm(formula = weight ~ height, data = women)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-1.7333 -1.1333 -0.3833  0.7417  3.1167
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height       3.45000    0.09114   37.85 1.09e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-squared:  0.991,    Adjusted R-squared:  0.9903
F-statistic: 1433 on 1 and 13 DF,  p-value: 1.091e-14
```

(2) 画出身高预测体重的散点图及回归线。

```
> plot(women$height,women$weight,xlab = "Height",ylab="Weight")
> abline(x)
```

用身高预测体重的散点图以及回归线见图 7-1。

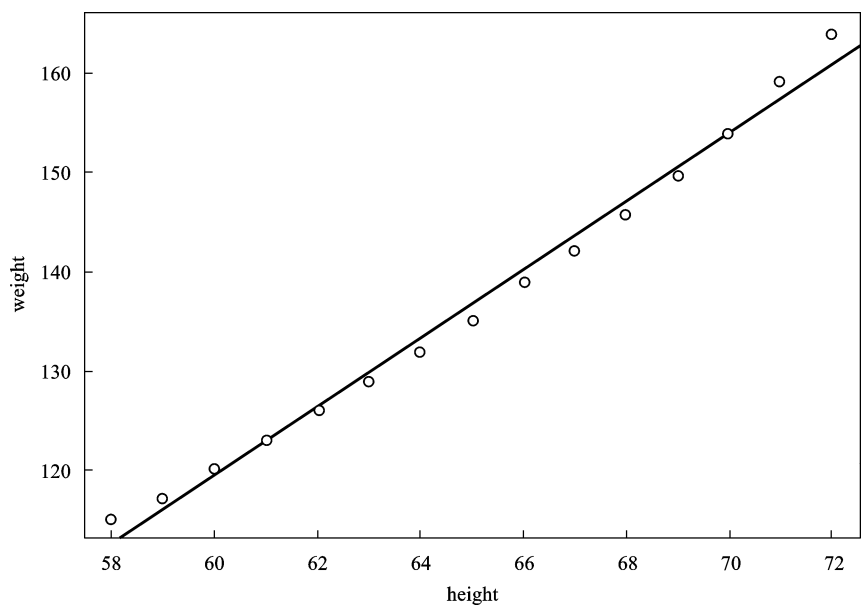


图 7-1 散点图以及回归线