回归分析与方差分析

李伟

回归分析的统计用途

- 挑选解释变量(explanatory variable)；
- 描述解释变量与响应变量(response variable)的关系;
- 生成两者之间的等式;
- 在y = kx + b中，y称为响应变量，x称为解释变量。

回归分析的应用情形

- 一个人在跑步机上时，预期消耗的卡路里数与时间、平均速度、年龄、性别、身体质量指数等的关系；
- 一个用户的哪些经历会导致他沉溺于大型多人在线角色扮演游戏；
- 教育环境中的哪些因素最能影响学生成绩得分；
- 运动场馆和职业运动对大都市的发展有何影响。

线性回归模型

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip}, i = 1,2,\cdots,n$$

假定：

1. $Y_1, Y_2, Y_3, \cdots, Y_n$独立，服从正态分布，并且具有相同的方差；
2. $Y_1, Y_2, Y_3, \cdots, Y_n$与$X_1, X_2, X_3, \cdots, X_p$之间存在线性关系。

R语言中的线性回归函数

lm():

lm(formula = , data =)

其中：

formula给出的是你的回归方程形式；

data是一个数据框，给出的是你的数据。

回归案例一:体重与身高的关系

我们采用的是R自带的数据集women.其给出了15个年龄在30~39周岁间女性的身高和体重信息。

我们希望通过身高来预测体重。通过一个等式来帮助我们分辨出那些过重或者过瘦的个体。

回归案例一:体重与身高的关系

```
# 以下试图建立 women 数据集 weight 和 height 之间关系

fit1 <- lm(weight ~ height, data = women)
# 建立回归方程

summary(fit1)
# 给出相应的回归信息
```

回归案例一:体重与身高的关系

women$weight #原始的体重值

fitted(fit1) #估计后的体重值

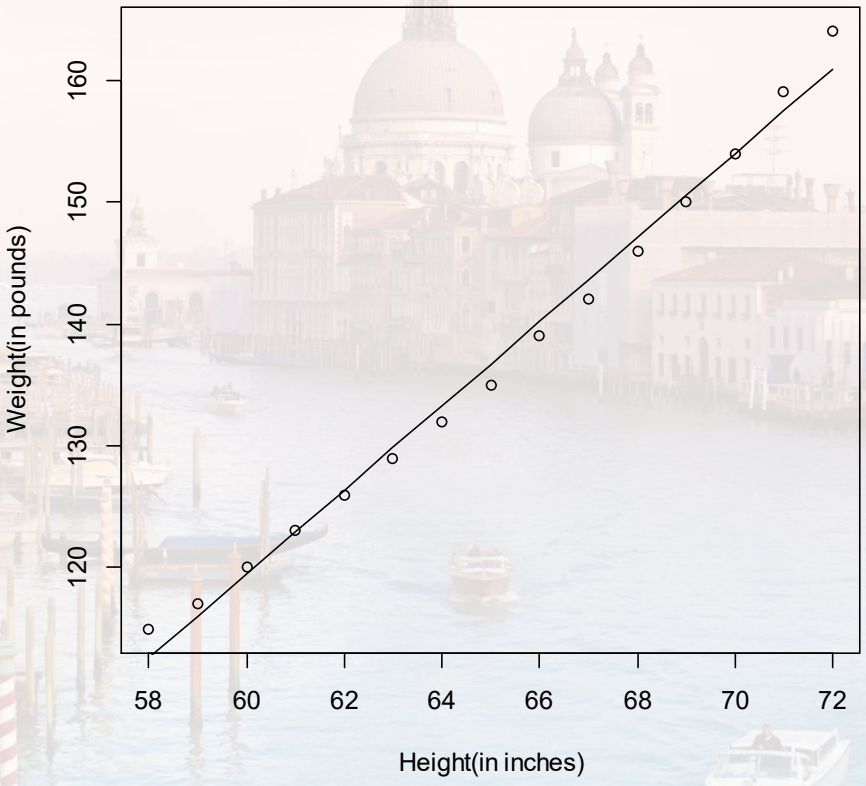
residuals(fit1) #原始值-估计值

回归案例一:体重与身高的关系

```
plot(women$height, women$weight,
     xlab = "Height(in inches)", ylab = "Weight(in pounds)",
     type = "p")
#画出原始的点

lines(women$height, fitted(fit1))
#画出估计的直线
```

回归案例一:体重与身高的关系



回归案例一:体重与身高的关系

#从图形上，可以看出，我们可以尝试用二次多项式去拟合

```
fit2 <- lm(weight ~ height + I(height ^ 2), data = women)
#建立回归方程

summary(fit2)
#给出相应的回归信息
```

回归案例一:体重与身高的关系

women$weight # 原始的体重值

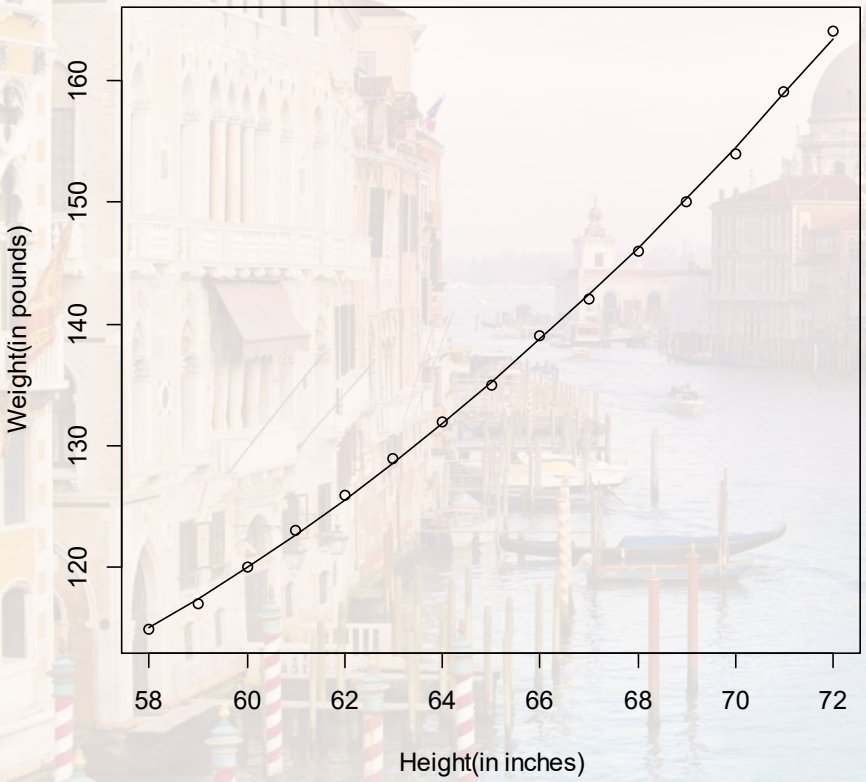
fitted(fit2) # 估计后的体重值

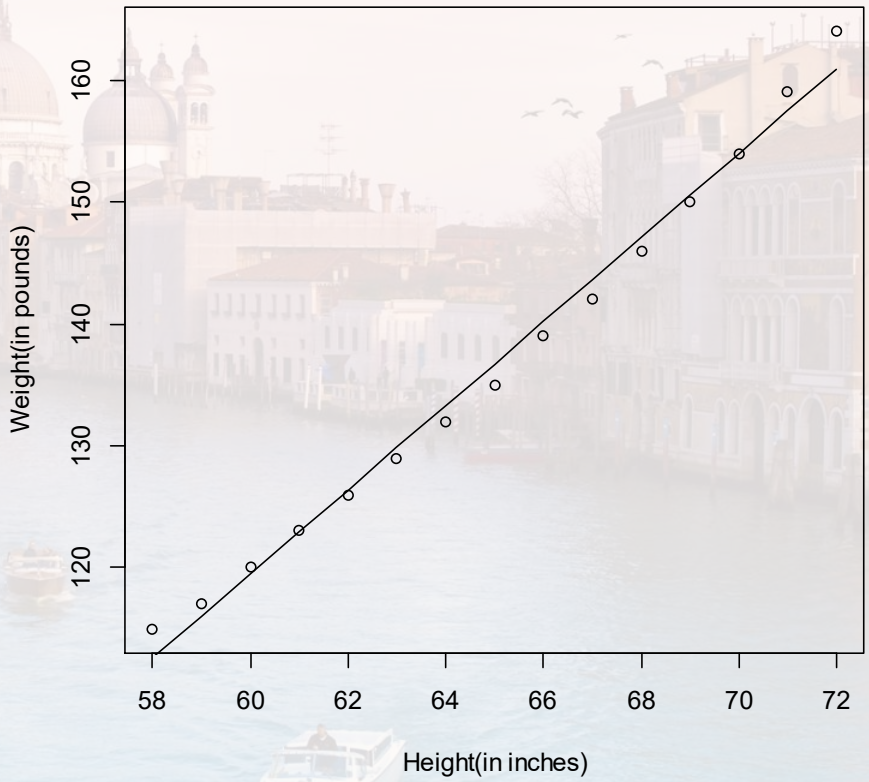
residuals(fit2) # 原始值-估计值

回归案例一:体重与身高的关系

```
plot(women$height, women$weight,
     xlab = "Height(in inches)", ylab = "Weight (in pounds)",
     type = "p")
#画出原始的点
lines(women$height, fitted(fit2))
#画出估计的直线
```

回归案例一:体重与身高的关系





作业1

建立mtcars数据集中的变量mpg(miles per gallon)与wt(weight)之间的一次线性关系与二次关系，比较两者的结果。

单因素方差分析

统计模型：

在单因素实验中，记因子为A，设其有r个水平，记为$A_1, A_2, \cdots, A_r$,

在每个水平下考察的指标视为一个整体，现有r个水平,故有r个整体。

假设：

(1)每个整体都服从正态分布，并且方差相等；

(2)从每个整体中抽出的样本相互独立。

方差分析案例-降低胆固醇的五种治疗方法对比

- 我们用的是multcomp包中的cholesterol数据集；
- 其中三种治疗方法是某种药物20mg/次/天，10mg/2次/天，5mg/4次/天；剩下的两种方法是drugD与drugE。
- 目的是为了比较不同治疗方法的效果。

方差分析案例-降低胆固醇的五种治疗方法对比

```
library(multcomp)
table(cholesterol$trt)
#不同治疗方法的实验次数
fit <- aov(response ~ trt, data = cholesterol)
#作方差分析
```

方差分析案例-降低胆固醇的五种治疗方法对比

summary(fit)

#方差分析结果

TukeyHSD(fit)

#对不同治疗方法进行两两比较

作业

✡ 比较ToothGrowth数据集中，采用不同的饲养方式，对小鼠某细胞长度的影响是否相同。

参考文献

R语言实战。By Robert I. Kabacoff. 高涛等译。人民邮电出版社。