

内容: 简单线性回归模型.

任务: 阅读下面的材料, 重复代码命令 (手工输入!), 并做练习 1-3, 提交结果。

线性回归分析的 R 函数主要包括:

```
lm, summary, plot, coefficient, residual, fitted
```

它们分别被用来完成如下任务:

- `lm`: 利用最小二乘法分析数据, 推断线性回归模组中的参数。模型结构主要以 $y \sim x_1 + x_2 \dots$ 的形式指定 (R 中称为 formula), 左侧为响应, 右侧 x_1, x_2, \dots 为自变量。例如

$$myfit = lm(y \sim x, data = mydata)$$

其中 y 是响应, x 是自变量, 数据集为 `mydata`. 分析结果存放在 `myfit` 中。

- `summary`, `plot`: 回归分析 `lm` 输出结果的汇总和图示, 调用方式为

$$summary(myfit), plot(myfit)$$

- `coefficient`, `residual`, `fitted`: 调取 `lm` 输出结果中的回归系数估计值, 残差, 拟合值。

首先从 github 安装 `alr4` 程序包:

```
> install.packages("devtools")
> library(devtools)
> devtools::install_github("cran/alr4")
```

例 1. R 程序包 `alr4` 中有一个数据集 `Heights`, 是 Karl Pearson 收集的 1100 多个家庭的母女身高数据, 其中每个家庭至多 2 个女儿, 女儿的年龄在 18 岁以上, 母亲年龄小于 65 岁. 我们感兴趣的是母亲身高 `Mheight` 对女儿身高 `Dheight` 的影响, 因此假设如下线性模型

$$dheight = a + b \times mheight + \epsilon, \epsilon \sim (0, \sigma^2) \quad (*)$$

R 命令如下:

```
> install.packages("alr4")
> library(alr4)
> Heights
      mheight dheight
      59.7      55.1
      58.2      56.5
      60.6      56.0
      ...
> myfit = lm(dheight~ mheight, data=Heights)
> myfit
Call:
lm(formula = dheight ~ mheight, data = Heights)
```

```
Coefficients:
(Intercept)      mheight
29.9174      0.5417
```

回归系数的 LS 估计 $\hat{a} = 29.9174$, $\hat{b} = 0.5417$, 更多的细节通过 `summary` 函数得到:

```
> summary(myfit) ##details
Call:
lm(formula = dheight ~ mheight, data = Heights)

Residuals:
Min       1Q   Median       3Q      Max
-7.397 -1.529  0.036  1.492  9.053

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.91744    1.62247   18.44  <2e-16 ***
mheight      0.54175    0.02596   20.87  <2e-16 ***
---
Residual standard error: 2.266 on 1373 degrees of freedom
Multiple R-squared:  0.2408,    Adjusted R-squared:  0.2402
F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16

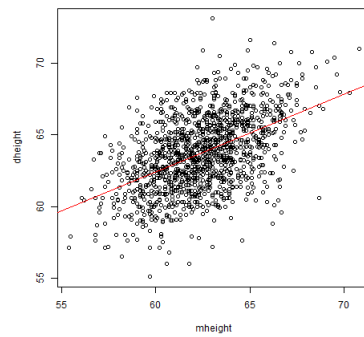
> coefficients(myfit)
      (Intercept)      mheight 
      29.917437      0.541747
```

汇总 (`summary`) 主要分三个部分

- Residuals 部分用 5 个数进行了描述 (Min 最小值, 1Q 即 25% 分位数, Median 中位数, 3Q 即 75% 分位数和 Max).
- Coefficients 部分列出了回归系数的 LS 估计 (Estimate)、标准差 (Std. error)、t 检验 (t value)、p 值 ($\text{Pr}(>|t|)$), 这里有两行, 第一行为截距项 (Intercept), 第二行为自变量 mheight 对应的回归系数.
- 最后三行描述了模型整体上的一些性质。Residual standard error: 2.266 为误差的标准差 ($\hat{\sigma}$), 自由度为 $n - 2 = 1373$ 。Mutiple R-squared: 0.2408 是决定系数 R^2 。最后一行为回归方程的显著性 F 检验用于同时检验所有自变量的显著性, 因为这里只有一个自变量, 所以它等价于 mheight 的 t 检验, 即 $F = t^2 = 20.87^2$ 。

下面通过二维散点图查看拟合效果: 首先用 `plot` 函数画出散点图, 然后在其上通过使用 `abline` 函数添加拟合得到的回归直线:

```
> plot(Heights)
> abline(myfit,col="red") #col=2
```



为了考察模型是否很好地拟合了数据，我们通常考察残差是否服从正态分布，以及考察残差是否与自变量存在某种非线性关系。这称为残差分析。

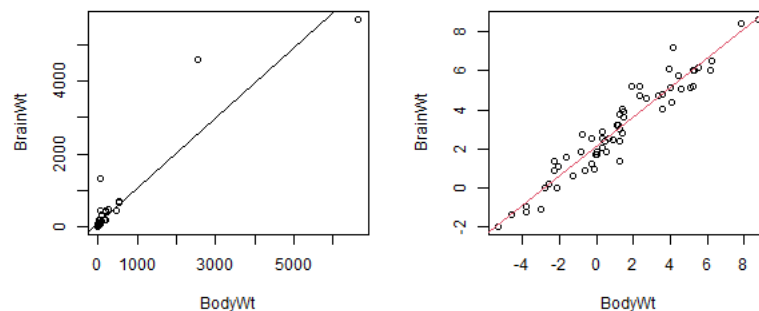
```
# 残差分析
res=residuals(myfit) # 残差
qqnorm(res) # 残差是否正态?
hist(res)# 残差是否正态?
plot(Heights[, "mheight"],res)
# 残差与自变量 mheight 是否存在某种非线性关系?
plot(myfit) # 包括了以上几个图
```

这里，你应该能看到残差基本服从正态，残差与自变量 mheight 没有明显的非线性关系，即模型拟合良好 (残差分析的细节以后会讲到)。

例 2. alr4 包中的数据集 brains 给出了 62 种哺乳动物的脑重量 (g) 和体重数据 (kg), 我们希望了解脑重 (响应) 与体重 (自变量) 的关系。

```
# 画出散点图，并拟合简单线性模型
> plot(brains[,2:1]) # 没有明显的线性关系
> a=lm(BrainWt~BodyWt, data= brains )
> abline(a) # 拟合效果不好

# 对两个变量取对数，
> plot(log(brains[,2:1])) # 在对数尺度上线性关系比较明显
> b=lm(BrainWt~BodyWt, data=log(brains))
> abline(b,col=2) # 效果很好
```



原始数据拟合得到的回归直线为

$$BrainWt = 91 + 0.97 \times BodyWt$$

除了拟合效果不好之外, 该模型还有其它不合理之处, 比如当 $BodyWt$ 趋于 0 时, $BrainWt$ 不会趋近 0。考虑对数变换, 对数数据拟合得到的回归直线为

$$\log(BrainWt) = 2.14 + 0.75 \times \log(BodyWt)$$

即 $3/4$ 幂次律:

$$BrainWt = e^{2.14} \times BodyWt^{3/4}$$

现在我们预测某种体重为 4kg 的动物的脑重, 第一个模型得到预测

$$91 + 0.97 \times 4 = 94.88(g)$$

第二个模型得到预测

$$e^{2.14} \times 4^{3/4} = 24.04(g)$$

查看体重在 4kg 附近的几种动物数据, 显然第二个模型预测更准确一些。

> brains[brains[,2]<4.5 &brains[,2]>3.5 ,]			
BrainWt	BodyWt		
Vervet	57.998	4.190	
Yellow-bellied marmot	17.000	4.050	
Rock hyrax2	21.000	3.600	
Raccoon	39.201	4.288	
Red fox	50.400	4.235	

练习题

练习 1. 福布斯 2019 财富榜前 100 名数据:

<http://staff.ustc.edu.cn/~ynyang/2025/lab/forbes2019.txt>

或

<http://staff.ustc.edu.cn/~ynyang/2025/lab/forbes2019.xlsx>

第 1 列是排名 (Rank), 第 4 列为财富值 (Wealth), 试研究财富与排名的关系。

练习 2. 因为身高 (H :height) 越高的人体重 (W :weight) 也偏大, 所以判断体重是否超标不能只看体重, 还要考虑其身高。为了消除掉体重中与身高有关的成分, 可考虑简单线性回归模型:

$$\log(W) = a + b \times \log(H) + \epsilon, \epsilon \sim N(0, \sigma^2) \quad (1)$$

该模型中 $\xi = \frac{\epsilon}{\sigma} = \frac{\log(W) - a - b \times \log(H)}{\sigma} \sim N(0, 1)$ 与体重有关但与身高 H 无关, 所以可认为是一个判断体重是否超标的指标, 如果参数 a, b, σ 已知, 那么我们可以用 $\xi > 1.645$ 作为体重超标的判别标准 ($1.645 = \text{qnorm}(0.95)$)。若 $\hat{a}, \hat{b}, \hat{\sigma}$ 是参数 a, b, σ 的估计, 那么判别标准为

$$\frac{\log(W) - \hat{a} - \hat{b} \times \log(H)}{\hat{\sigma}} > 1.645$$

等价地, 若

$$\xi = \frac{W}{H^b} > \exp(\hat{a} + 1.645\hat{\sigma})$$

则判别为偏胖 (我们称 ξ 为体重指数)。 下载成年人身高-体重数据

<http://staff.ustc.edu.cn/~ynyang/2025/lab/height-weight.txt>

该数据的三个变量为: sex (1: M, 0: F), weight (kg), height (m).

- (a) 应用模型 (1), 使用所有数据 (不考虑性别) 求出 a, b, σ 的 LS 估计, 计算你自己的体重指数 ξ , 判断自己体重是否超标, 并计算群体中超过你的体重指数 ξ 值的人的比例。
- (b) 我们可以不取对数, 直接建立线性模型 $W = a + bH + \epsilon, \epsilon \sim N(0, \sigma^2)$, 则

$$\epsilon/\sigma = (W - a - bH)/\sigma \sim N(0, 1)$$

拟合该模型求出 a, b, σ^2 的 LS 估计, 得到消除了身高影响的体重指数

$$\eta = (W - \hat{a} - \hat{b}H)/\hat{\sigma}$$

计算自己的 η 值, 并计算群体中超过你的 η 值的人的比例。与 (a) 的结果是否一致? 哪个更合理?

- (c) 显然性别与 W, H 都有关, 因此我们应该在对数尺度简单模型中添加一项控制性别: $\log(W) = a + b \times \log(H) + c \times Sex + \epsilon$, 相应地, R 命令为 `lm(logW ~ logH + Sex)`。此时, 你能否推断出 $b = 2$?

练习 3. 有人声称如下论断: 如果一个正随机变量 x 服从对数正态分布, 即 $\log(x) \sim N(0, 1)$, 则 x 的首位非 0 数字 d 服从 Benford 定律, 即

$$P(d = i) = \log_{10}(1 + 1/i), i = 1, 2, \dots, 9.$$

试通过模拟实验验证上述论断是否成立。

提示: 你需要产生大量的 $N(0, 1)$ 随机数, 并写一个函数取出任一正实数的首位非 0 数字 (注意 1.22, 0.00123, 123 的首位非 0 数字都是 1), 例如:

```
n=100000 #n 个随机数
z=rnorm(n) # 标准正态随机样本, 样本量 =n
x=exp(z)
d= firstdigit(x) # 写一个函数计算正数的首位数字
table(d) /n # 统计各个数字的频率
```