

引言

马学俊 (主讲) 晁越 (助教)

苏州大学

数学科学学院

<https://xuejunma.github.io/>



① 回归的来源

② 实际问题

③ 变量间关系

④ 主要内容

⑤ 发展述评

关于 RA 的发展情况

- Galton(1822-1911) 在 1886 年发表了关于回归的开山论文《遗传结构中向中心的回归 (Regression towards mediocrity in heredity structure)》到现在是 130 多年
- 研究父代身高与子代身高之间的关系，发现：

$$\hat{y} = 33.73 + 0.156x$$

- 子代平均身高介于其父代的身高和种群平均身高之间
- 高个子父亲的儿子的身高低于其父亲身高的趋势
- 矮个子父亲的儿子的身高则高于其父亲身高的趋势
- 子代身高有向种群身高“回归”的趋势
- 现代回归：
 - 探索和检验自变量 (X) 和因变量 (Y) 之间的关系：因果关
系，数量关系
 - 基于自变量的取值变化预测因变量的取值
 - 描述自变量与因变量之间的关系

New York Rivers

- 数据名称: NewYorkRivers.txt
- 变量说明: 纽约州 20 条河流
 - Y:Nitrogen 平均氮浓度 (毫升/升)
 - X_1 :Agr 农业用地面积百分比
 - X_2 :Forest 森林占地面积百分比
 - X_3 :Rsdntial 住宅用地面积百分比
 - X_4 :ComIndl 工业用地面积百分比
- 分析目的: 河流流域土地利用状况对水质污染的影响

_____ NewYorkRivers.txt _____

```
1 > rm(list=ls())
2 > dat_ryr <- read.table("NewYorkRivers.txt", head=TRUE)
3 > head(dat_ryr)
```

	River	Agr	Forest	Rsdntial	ComIndl	Nitrogen
5 1	Olean	26	63	1.2	0.29	1.10
6 2	Cassadaga	29	57	0.7	0.09	1.01
7 3	Oatka	54	26	1.8	0.58	1.90
8 4	Neversink	2	84	1.9	1.98	1.00
9 5	Hackensack	3	27	29.4	3.11	1.99
10 6	Wappinger	19	61	3.4	0.56	1.42

EgyptianSkulls.txt

- 数据名称: EgyptianSkulls.txt
- 变量说明:
 - Y: 大概的年份 (负值 = 公元前; 正值 = 公元后)
 - X_1 : MB 头盖骨的最大宽度
 - X_2 : BH 头盖骨的颅最高点的高度
 - X_3 : BL 头盖骨颅底牙槽的长度
 - X_4 : NH 头盖骨的鼻高度
- 分析目的: 推断埃及头骨的年代

_____ EgyptianSkulls.txt _____

```
1 > rm(list=ls())
2 > dat_es <- read.table("EgyptianSkulls.txt", head=TRUE)
3 > head(dat_es)
4      Year MB   BH   BL NH
5 1 -4000 131 138   89 49
6 2 -4000 125 131   92 48
7 3 -4000 131 132   99 50
8 4 -4000 119 132   96 44
9 5 -4000 136 143 100 54
10 6 -4000 138 137   89 56
```

Financial Ratios.txt

- 数据名称: Financial Ratios.txt
- 66 家企业的营运财务比, 其中 33 家破产、33 家企业经营稳定
 - Y: 是否破产
 - X_1 : 未支付利润/总资产
 - X_2 : 支付利息和税金前利润/总资产
 - X_3 : 销售额/总资产
- 分析目的: 破产因素

Financial Ratios.txt

```
1 > rm(list=ls())
2 > dat_fr <- read.table("FinancialRatios.txt", head=TRUE)
3 > head(dat_fr)
4   Y      X1      X2  X3
5 1 0   -62.8  -89.5 1.7
6 2 0    3.3   -3.5 1.1
7 3 0 -120.8 -103.2 2.5
8 4 0  -18.1  -28.8 1.1
9 5 0   -3.8  -50.6 0.9
10 6 0  -61.2  -56.2 1.7
```

NewDrugs.txt

- 数据名称: NewDrugs.txt
- 1992-1995 年 16 年中疾病引入的新药
 - D: 新药个数
 - P: 每十万人的病人数
 - M: 1994 年研究经费 (百万美元)
- 分析目的: 新药功效

_____ NewDrugs.txt _____

```
1 > rm(list=ls())
2 > dat_nd <- read.table("NewDrugs.txt", head=TRUE)
3 > head(dat_nd)
4
5 1      IschemicHeartDisease  6  8976  198.4
6 2      LungCancer          3   874   80.2
7 3      HIV/AIDS            21  1303 1049.6
8 4      AlcoholUse          2 18092  222.6
9 5 CerebrovascularDisease    2  9467  108.5
10 6      COPD                1  4271   48.9
```

11

变量间的关系

函数关系

- 商品的销售额与销售量之间的关系

$$y = px$$

- 圆的面积与半径之间的关系

$$S = \pi R^2$$

- 原材料消耗额与产量 (x_1)、单位产量消耗 (x_2)、原材料价格 (x_3) 之间的关系

$$y = x_1 x_2 x_3$$

函数关系

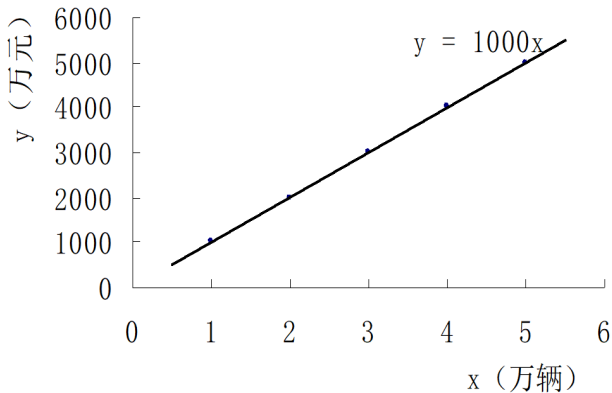
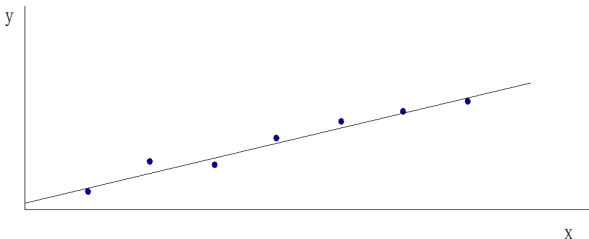


图1.1 函数关系图

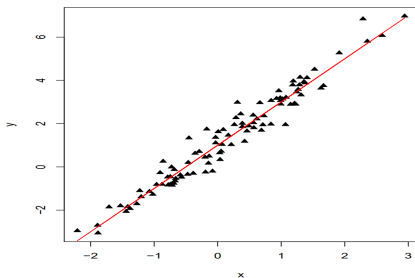
相关关系

相关关系

- 子女身高 (y) 与父亲身高 (x) 之间的关系
- 收入水平 (y) 与受教育程度 (x) 之间的关系
- 粮食亩产量 (y) 与施肥量 (x_1)、降雨量 (x_2)、温度 (x_3) 之间的关系
- 商品的消费量 (y) 与居民收入 (x) 之间的关系
- 商品销售额 (y) 与广告费支出 (x) 之间的关系



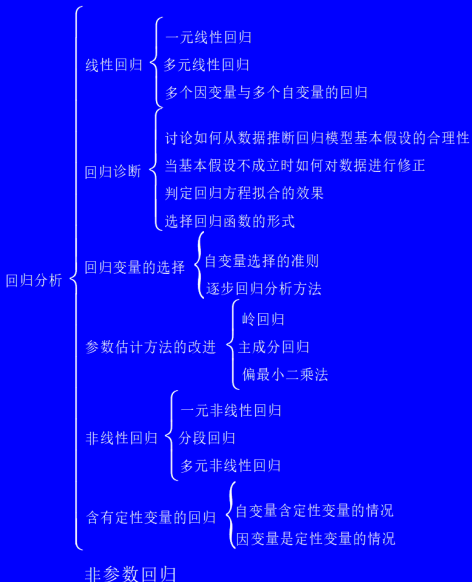
模拟例子



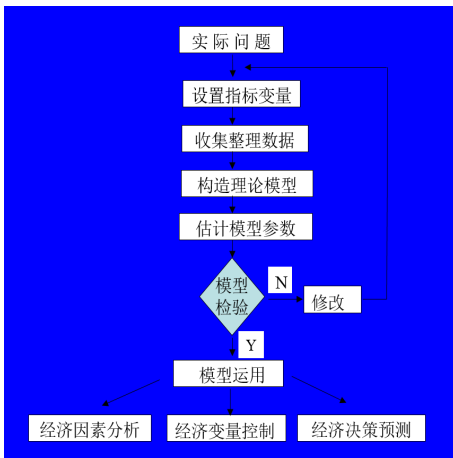
R 代码

```
1 rm(list=ls())  
2 n <- 100  
3 x <- rnorm(n)  
4 y <- 1+ 2 * x + 0.5 * rnorm(n)  
5 plot(x, y, pch=17)  
6 lines(x, 1+ 2*x, col="red")
```

主要内容



建立实际问题回归模型的过程



变量及样本较多，参数估计的计算量很大，只有依靠计算机。

现在这方面的现成计算机软件很多，如SPSS、R、SAS、Minitab。

回归分析应用与发展述评

- 线性回归
- 分位数回归 (Quantile regression)
- 众数回归 (Mode regression)
-
- LASSO

作业

设 y_1, \dots, y_n 是一组样本，其中 μ 和 σ 都是未知的。构建下面模型

$$y_i = \mu + \varepsilon_i, \quad i = 1, 2, \dots, n$$

我们可以采用

- 最小二乘法估计: 最小化 $\sum_{i=1}^2 (y_i - \mu)^2$
- 最小绝对值估计: 最小化 $\sum_{i=1}^2 |y_i - \mu|$

回答下面问题:

- ① 证明 μ 的最小二乘估计是样本均值
- ② 证明 μ 的最小绝对值估计是样本中位数
- ③ 列出样本均值的一个优点和一个缺点
- ④ 列出样本中位数的一个优点和一个缺点
- ⑤ 你会选择 μ 的两个估计量的哪一个？说出理由。