应用回归分析
引言

马学俊(主讲) 杜悦(助教)

苏州大学

数学科学学院

https://xuejunma.github.io/

概要

关于RA的发展情况

- 高斯(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提出，他发现:

$$\hat{y} = 33.73 + 0.516x$$

  - 子代平均身高介于其父代的身高和种群的平均身高之间
  - 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的趋势
  - 子代身高有向族群身高“回归”的趋势
- 现代的回归：
  - 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

```
> rm(list=ls())
> dat_ryr <- read.table("NewYorkRivers.txt", head=TRUE)
> head(dat_ryr)
        River Agr Forest Rsdntial ComIndl Nitrogen
1       Olean  26     63      1.2    0.29     1.10
2   Cassadaga  29     57      0.7    0.09     1.01
3       Oatka  54     26      1.8    0.58     1.90
4   Neversink   2     84      1.9    1.98     1.00
5 Hackensack   3     27     29.4    3.11     1.99
6   Wappinger  19     61      3.4    0.56     1.42
```

Egyptian Skulls

- 数据名称: EgyptianSkulls.txt
- 变量说明
  - Y: Year 大改的年份(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

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

Financial Ratios.txt

- 数据名称: `FinancialRatios.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: 未分配利润/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

FinancialRatios.txt

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

NewDrugs.txt

- 数据名称: `New Drugs.txt`
- 1992-1995年16中疾病引入的新药
  - D: 新药个数
  - P：每十万人的的病人数
  - M：1994年研究经费(百万美元)
- 分析目的：破产因素

FinancialRatios.txt

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

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$(x1)$、单位产量消耗$(x2)$、原材料价格$(x3)$之间的关系

$$y = x_1 x_2 x_3$$

函数关系

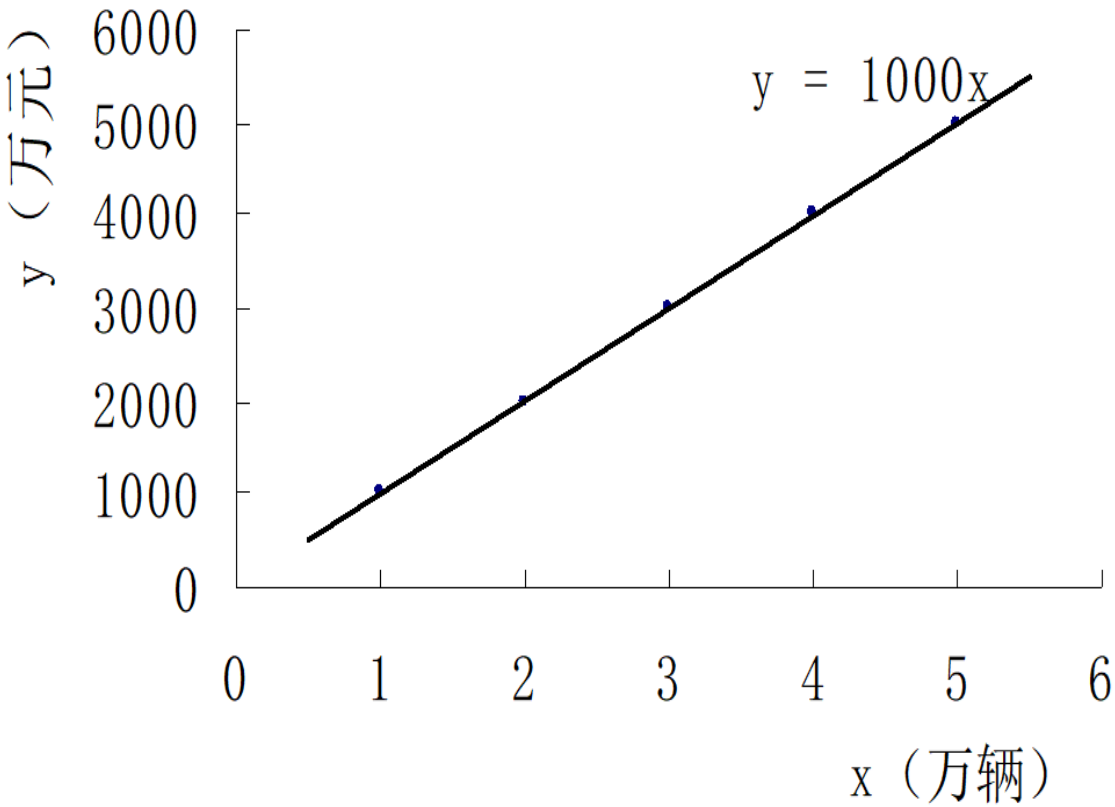


图1.1 函数关系图

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$(x1)$、降雨量$(x2)$、温度$(x3)$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

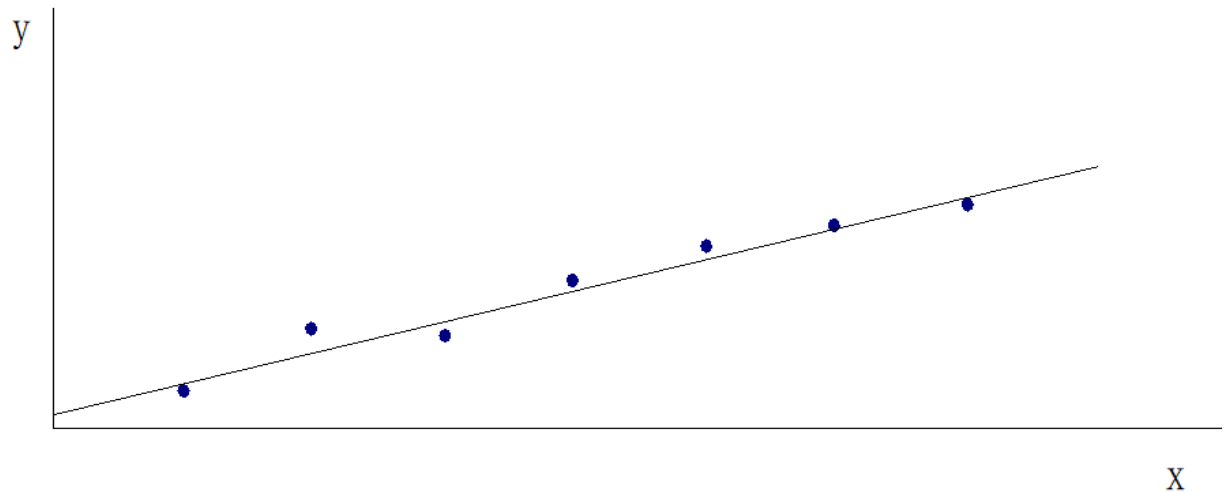
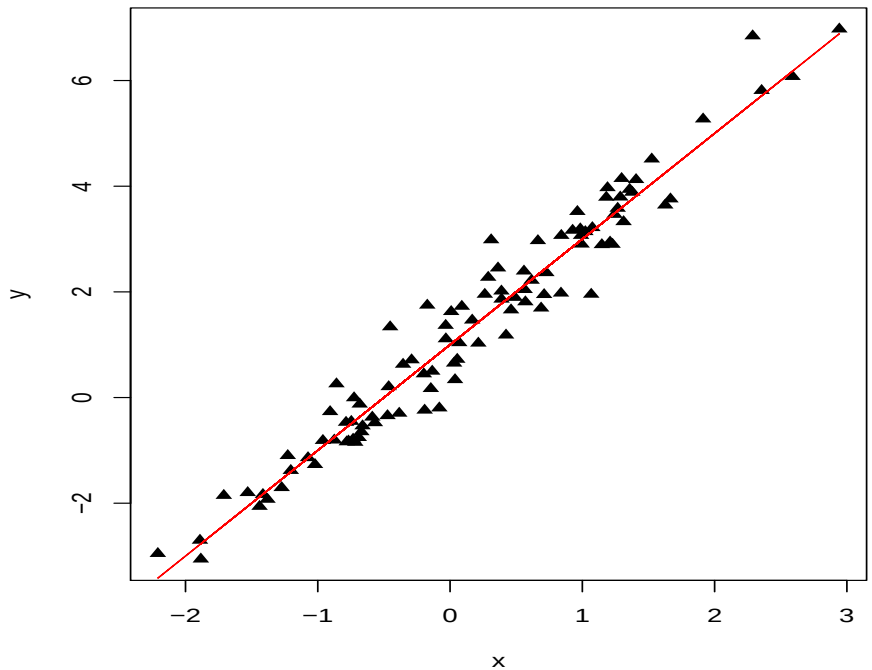


图1.2　y与x非确定性关系图

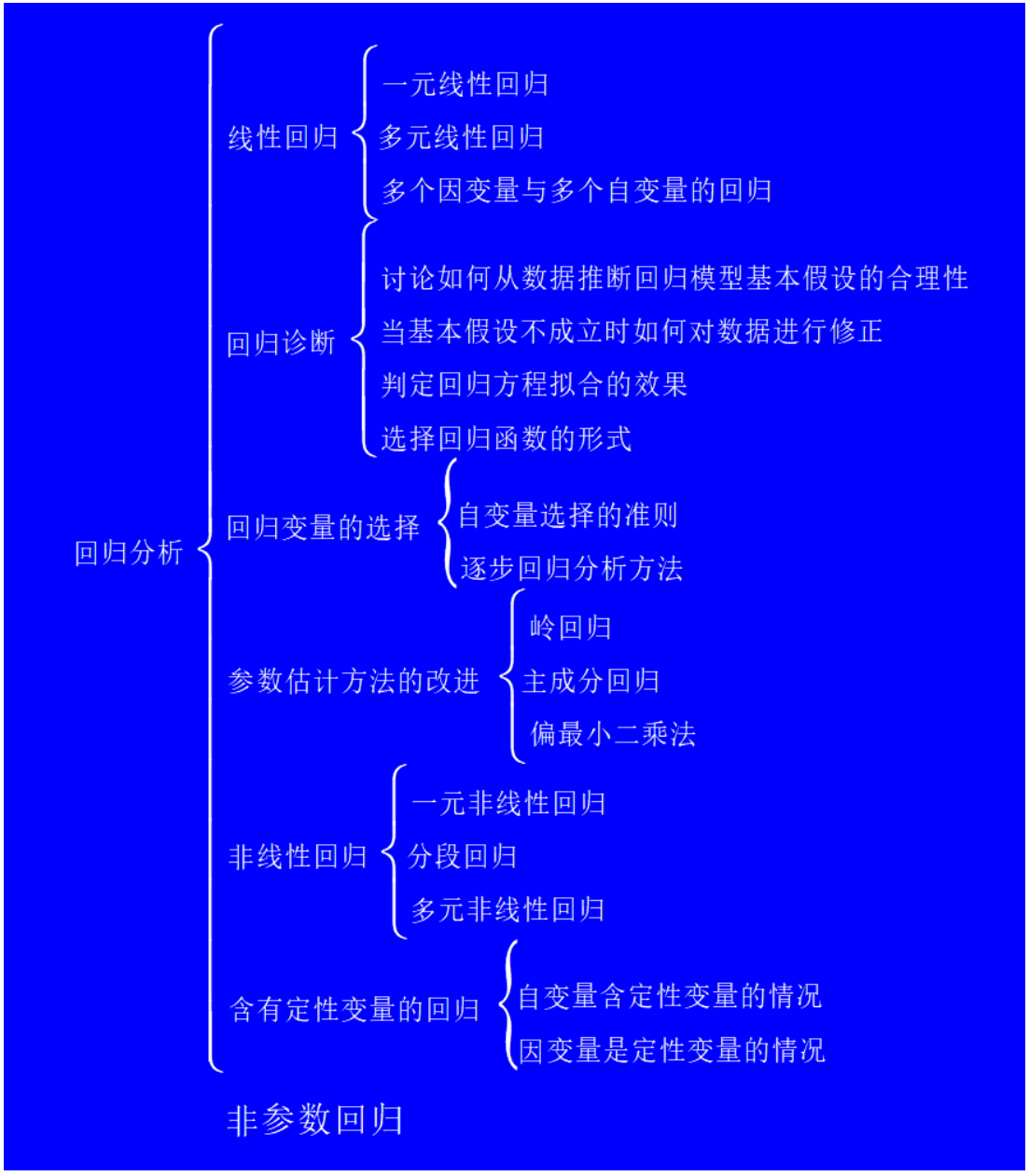
模拟例子



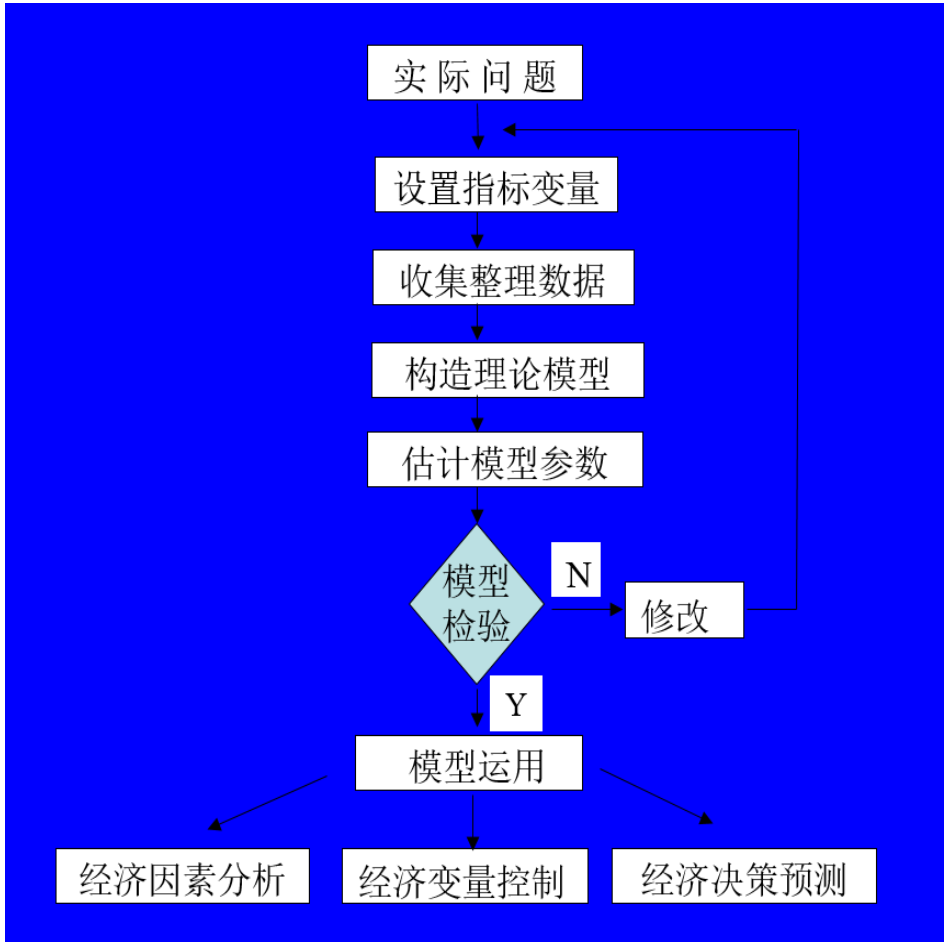
R代码

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

主要内容


回归分析
线性回归
一元线性回归
多元线性回归
多个因变量与多个自变量的回归
回归诊断
讨论如何从数据推断回归模型基本假设的合理性
当基本假设不成立时如何对数据进行修正
判定回归方程拟合的效果
选择回归函数的形式
回归变量的选择
自变量选择的准则
逐步回归分析方法
参数估计方法的改进
岭回归
主成分回归
偏最小二乘法
非线性回归
一元非线性回归
分段回归
多元非线性回归
含有定性变量的回归
自变量含定性变量的情况
因变量是定性变量的情况
非参数回归

建立实际问题回归模型的过程



这里要说明的是，当变量及样本较多时，参数估计的计算量很大，只有依靠计算机才能得到可靠的结果。现在这方面的现成计算机软件很多，如SPSS、R、SAS、Minitab等都是参数估计的基本软件。

回归分析应用与发展述评

从Gauss提出最小二乘法算起,回归分析已经有200年的历史。回归分析的应用非常广泛,我们大概很难找到不用它的领域,这也正是一百多年来经久不衰,生命力强大的根本原因。这里简述回归分析在经济领域的广泛应用。我们知道计量经济学是现代经济学中影响最大的一门独立学科。诺贝尔经济学奖获得者萨缪尔森曾经说过:第二次世界大战后的经济学是计量经济学的时代。然而,计量经济学中的基本计量方法就是回归分析,计量经济学的一个重要理论支柱是回归分析理论。

自从1969年设立诺贝尔经济学奖以来,已有60多位学者获奖,其中绝大部分获奖者是统计学家、计量经济学家、数学家。从大多数获奖者的著作看,他们对统计学及回归分析方法的应用都有娴熟的技巧。这足以说明统计学方法在现代经济研究中的重要作用。

矩阵理论和计算机技术的发展为回归分析模型在经济研究中的应用提供了极大的方便。国民经济是一个错综复杂的系统,对于宏观经济问题常需要涉及几十个甚至几千个变量和方程,如果没有先进的计算机和求解线性方程组的矩阵计算理论,要研究复杂的经济问题是不可想象的。

一个20阶的线性方程组要用克莱姆法则去求解,就需要计算1022次乘法运算,这可是一个天文数字。然而用矩阵变换的方法只需6000次乘法运算。也正是由于计算方法的改进和现代计算机的发展,使得过去不可想象的事情变成了现实。

计量经济学研究中涉及的变量和方程也越来越多。例如英国剑桥大学的多部门动态模型,多达2 759个方程,7 484个变量;由诺贝尔经济学奖获得者克莱因发起的国际连接系统,使用了7 447个方程和3 368个外生变量。

模型技术在经济问题研究中的应用在我国也已盛行起来。从20世纪80年代初期以来,每年都有许多国家级和省级鉴定的计量经济应用成果。特别是在一些省级以上的重点经济课题中,经济学硕士学位的论文中,如果没有模型技术的应用,给人的印象总感分量不足。这些足以说明模型技术的应用在我国也倍受重视。这里要强调说明的是,回归分析方法是模型技术中最基本的内容。

回归分析的理论和方法研究200年来也得到不断发展。统计学中的许多重要方法都与回归分析有着密切的联系。如时间序列分析、判别分析、主成分分析、因子分析、典型相关分析等。这些都极大地丰富了统计学方法的宝库。

回归分析方法自身的完善和发展至今是统计学家研究的热点课题。例如自变量的选择、稳健回归、回归诊断、投影寻踪、非参数回归模型等近年仍有大量研究文献出现。在回归模型中,当自变量代表时间,因变量不独立并且构成平稳序列时,这种回归模型的研究就是统计学中的另一个重要分支——时间序列分析。它提供了一系列动态数据的处理方法,帮助人们科学地研究分析所获得的动态数据,从而建立描述动态数据的数学模型,以达到预测、控制的目的。

因变量y和自变量x都是一维时,称它为一元回归模型;当x是多维,y是一维时,则它为多元回归模型;若x是多维,y也是多维的,则称它为多重回归模型。特别是当因变量观察矩阵Y的诸行向量假定是独立的,而列向量假定是相关的,就称为半相依回归方程系统。对于满足基本假设的回归模型,它的理论已经成熟,但对于违背基本假设的回归模型的参数估计问题近些年仍有较多研究。

在实际问题的研究应用中,人们发现经典的最小二乘估计的结果并不总是令人满意,统计学家们从多方面进行努力试图克服经典方法的不足。例如,为了克服设计矩阵的病态性,提出了以岭估计为代表的多种有偏估计。Stein于1955年证明了当维数p大于2时,正态均值向量最小二乘估计的不可容许性,即能够找到另一个估计在某种意义上一致优于最小二乘估计.

从此之后人们提出了许多新的估计,其中主要有岭估计,主成分估计,Stein估计,以及特征根估计,偏最小二乘法。这些估计的共同点是有偏的,即它们的均值并不等于待估参数。于是人们把这些估计称为有偏估计。当设计矩阵X 呈病态时,这些估计都改进了最小二乘估计。


regression
variables
analysis
linear
Jump
Regression
dependent
independent
model
parameters
assumptions
beta
data
values
variable
squares
statistical
least
also
extrapolation
edit
may
methods
hat
number
used
Statistical
sum
function
unknown
different
sometimes
example
error
widehat
form
varepsilon
set
models
normal
equations
Statistics
Galton
sample
term
estimates
observations
method
population
Gauss
errors
logit
known
information
observed
bar
response
three
fit
conditional
distribution
variance
using
one
estimated
given
value
include
probit
estimation
relationship
Bayesian
Linear
assumption
case
measurements
Fisher
parameter
predicted
Nonlinear
packages