第一章：回归分析概述

黄磊

http://userweb.swjtu.edu.cn/Userweb/yellones/index.htm

数学学院

西南交通大学

August 27, 2019

Outline

教材参考书

1. 教材：何晓群，《应用回归分析（R语言版）》
   理论与应用相结合，作为主讲内容
2. 中文参考书：王松桂，陈敏，陈立萍，《线性统计模型：线性回归与方差分析》
   注重理论，作为补充
3. 英文参考书：Regression-Linear Models in Statistics, By N.H. Bingham and John M.Fry, Springer.
   拓展视野，培养学术英语

程序语言

1. 要求程序：R语言+R studio。SPSS自己看书，鼓励自学matlab 或者python.

   R下载地址
   https://mirrors.ustc.edu.cn/CRAN
   R studio下载地址
   https://www.rstudio.com/products/rstudio/download/#download

   News: COPSS Award 2019, Hadley Wickham, R studio 首席科学家

2. R学习资源：
   R官方网站：https://mirrors.ustc.edu.cn/CRAN
   或者老师提供的PPT教案

辅导安排

1. 定期习题课，每两周一次,每次一节课或者30分钟
2. R语言和程序辅导课，每个月一次，由老师安排，主讲R的数据操作和线性模型相关内容
3. 由班长建立QQ群，方便同学交流学习，方便老师上传教案材料+其他学习资料
4. 实际教学过程中，以上安排会根据进度稍作调整。

期末综合考核

1. 总成绩(占比%)=平时作业(10%)+考勤(5%)+小课题(10%)
   +期中考试(20%)+期末考试(55%)

2. 小课题要求,
   (1)文章阅读(5%, 10月底完成)，(2)数据建模(5%, 11月底完成)

小课题—文章阅读

文章阅读(5%, 10月底完成)

1. 首先，每三人组成一个团队，从老师推荐的期刊中下载一篇与回归分析或者其他统计方法、模型相关的文章,
2. 然后，将文章分为背景介绍(1.Introduction), 模型和方法（2.Methodology), 模拟计算与数据分析(3.Simulation and Real Data Analysis), 形式不仅限于此,
3. 最后，以小组为单位完成一篇1000字以内!的文章总结介绍，重点突出该文章研究的意义及创新之处（包括说清楚过往方法的局限性，该文章所提方法的优势等）并讲清楚统计模拟和数据分析的结果如何支撑其研究意义和创新之处。

小课题—推荐期刊

下载近五年的文章，2014—2019，文章最好来自以下推荐清单

- Journal of the American Statistical Association
- Statistica Sinica
- Journal of Multivariate Analysis
- Journal of Statistical Computation and Simulation
- Computational Statistics & Data Analysis

- Biometrika
- Statistics in Medicine
- Statistical Methods in Medical Research
- Lifetime Data Analysis
- Biometrics

- Journal of Business and Economic Statistics
- Journal of Econometrics
- Econometric Theory
- Journal of Applied Econometrics
- Journal of Empirical Finance

小课题—数据建模

数据建模(5%, 11月底完成)

1. 同样，每三个人一组，自行下载可以用线性模型分析的数据（经济金融、医学、社会科学）

2. 可将任务分为，
   (1) 数据背景意义介绍，
   (2) 数据收集、清理工作，
   (3) 建模、模型选择，
   (4) 统计分析与总结

3. 以小组为单位完成一篇1000字以上的数据建模报告

Outline

变量间的统计关系

变量间的关系因紧密程度（如何用概率刻画？）不同, 大致分为两种

1. 确定关系，例如，银行的定期存款和，保险的保费和保额，自由落体距离公式（理想），还有吗？

$$y = f(x_1, x_2, \ldots, x_p)$$

2. 非确定关系，例如，母婴用品消费与家庭收入，自由落体距离公式（现实世界），股票、基金的收益，还有吗？

$$y = f(x_1, x_2, \ldots, x_p) + \varepsilon$$

非确定性关系的成因：某些因素没考虑周全，试验误差，测量误差，偶然因素等等。

非确定性关系图例

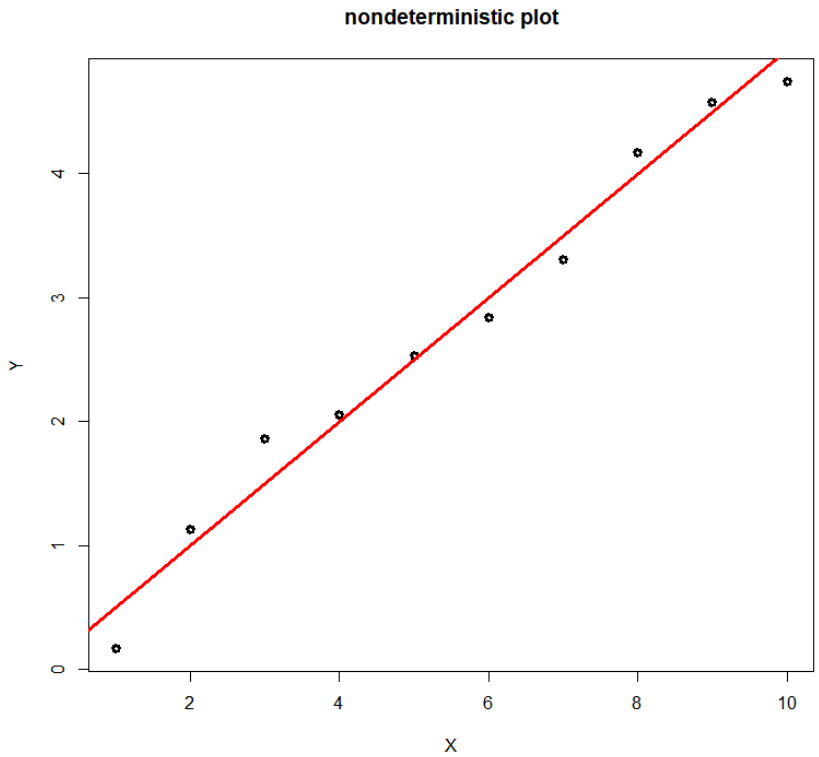

Figure 1: 一个$y = 0.5 * x + \varepsilon$的非确定关系图

$x < -seq(1, 10, 1); y < -0.5 * x + 0.2 * rnorm(10);$

$plot(x, y, lwd = 3); abline(a = 0, b = 0.5, col =' red', lwd = 3.5)$

变量间的统计关系

如此的y和x存在一定的关系，但又非确定关系，称之为变量间统计关系

1. 相关关系
   a. y与x地位平等
   b. y与x全是随机变量
   c. 主要目的是刻画相关程度（线性相关）和方式（copula函数）
2. 回归分析
   a. y被解释(响应变量，因变量)；x用来解释y, 称为自变量(回归变量，解释变量)
   b. y是随机变量，x可随机(上证综合指数日收益率)，也可为确定变量(例如，年龄，学历)
   c. 不仅解释x对y的影响方式(正负)和程度（显著），还可以进行预测和控制

回归方程

给定x时y的条件期望，称之为回归函数

$$f(x) = E(y|x) \tag{2.1}$$

它从平均意义上刻画了变量x与y之间的统计规律。实际问题中，x称为(自变量），y(因变量）。回归模型之目的，就是以样本数据

$$(x_1, y_1), \ldots, (x_n, y_n),$$

去估计回归函数。实践中，最经典的，应用范围最广的，就是考虑(2.1)中的$f(x)$为线性函数，

$$E(y|x) = \alpha + \beta x, \tag{2.2}$$

未知的参数α, β就需要用样本观测数据去估计(估计方法为本课程的重点)，假如我们从样本数据中得到了估计值，$\hat{\alpha}, \hat{\beta}$, 代入(2.2),得到

$$\hat{y} = \hat{\alpha} + \hat{\beta} x \tag{2.3}$$

回归方程名称和由来

(2.3)称为线性回归方程，或者经验回归方程。
(2.2)称为理论回归方程。
α称为截距项，回归常数；β称为回归系数，其对应的估计值$\hat{\alpha}, \hat{\beta}$称为经验回归常数和经验回归系数。
Regression的基本思想和方法归功于F.Galton 以及他的学生K.Pearson 研究父母身高和子女身高的遗传问题时，他们得到经验回归方程

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

标记y的均值为$\bar{y}$, x的均值为$\bar{x}$, 由于$\bar{y}, \bar{x}$也满足经验回归方程（下一章将证明）。因此

$$\hat{y} - \bar{y} = 0 + 0.516 * (x - \bar{x})$$

因此人类子代回归稳定身高水平得以解释。稳定就来自于回归系数$|\beta| < 1$。

一、主要内容

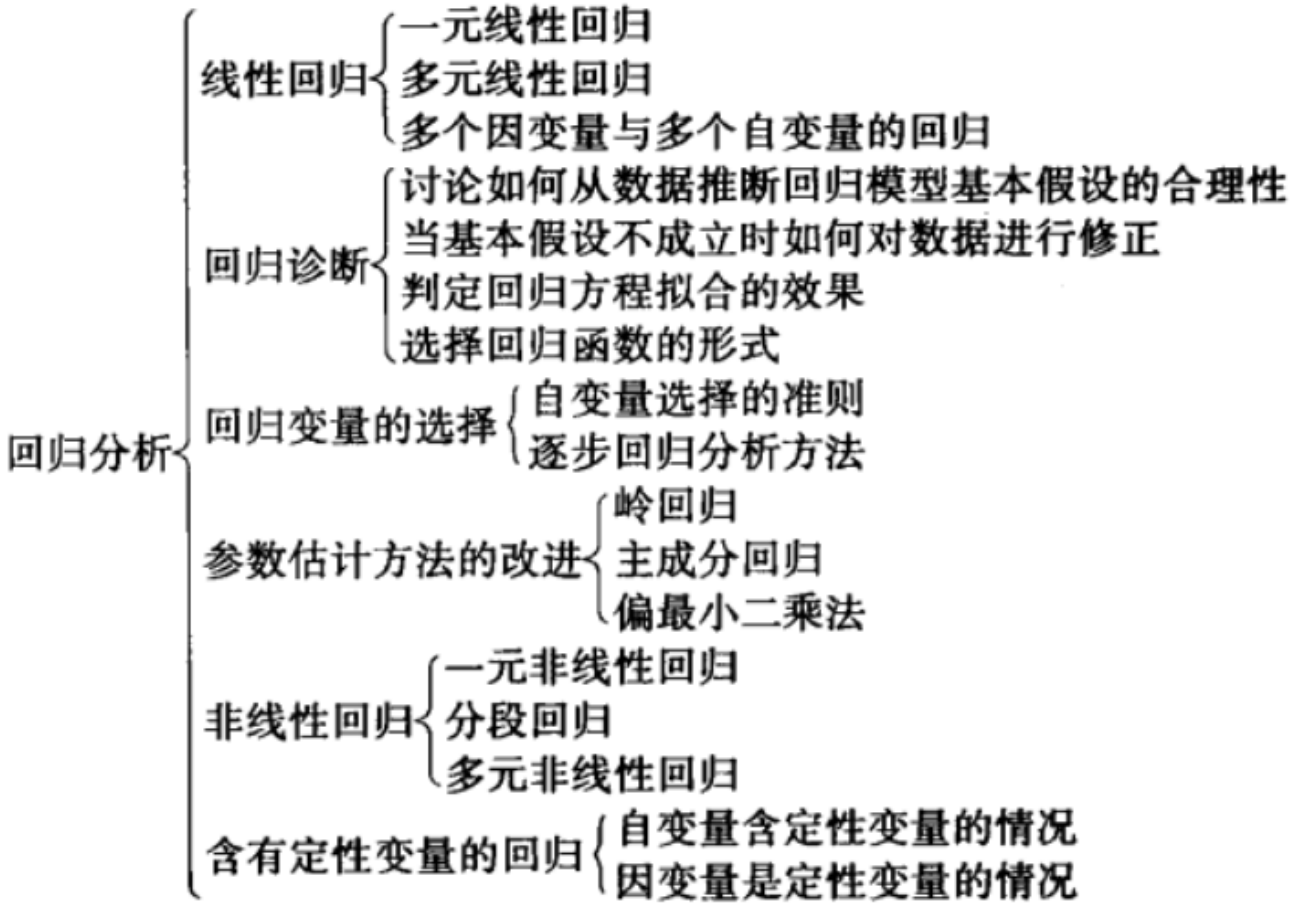

Figure 2: 回归分析的主要内容

二、回归模型的一般形式

随机变量y与相关变量$x_1, x_2, \ldots, x_p$之间的一般回归模型为

$$y = f(x_1, x_2, \ldots, x_p) + \varepsilon \tag{2.4}$$

1. y响应变量，$x_1, x_2, \ldots, x_p$为解释变量。经济学中，y内生变量，x_j外生变量。
2. $f(\cdot)$函数在本课程里是已知函数，这部分又称为确定性分量，ε又可成为随机性分量。当$f(\cdot)$未知时，就是非参数、半参数回归分析（研究生重要课程《近代回归分析》）。
3. ε 随机误差，因其引入，(2.4)才被描述为一个随机方程。
   a. 认识的局限，成本的制约，未引入的解释变量
   b. 采集过程中的观测误差
   c. 理论模型设定的误差，或者其他不可避免的随机因素

一般线性回归模型

当(2.4)中的函数$f(\cdot)$是关于$x_1, \ldots, x_p$的线性函数时，就得到一般线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon, \tag{2.5}$$

其中，$\{\beta_j, j = 0, 1 \ldots, p\}$ 是未知参数，回归系数。在获得样本观测值$\{(y_i; x_{1i}, \ldots, x_{pi}), i = 1, \ldots, n\}$后，线性模型可表示为

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i, i = 1, \ldots, n. \tag{2.6}$$

基本条件

为了估计模型参数的需要以及推到估计量的统计性质，古典线性回归模型通常应满足以下基本假设：

(1) 解释变量$x_1, \ldots, x_p$ 是非随机变量，观测值$\{x_{i1}, x_{i2}, \ldots, x_{ip}\}$是常数。

(2) 零均值，等方差及不相关假定条件(Gauss-Markov条件）

$$\begin{cases} E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2, i = 1, \ldots, n \\ Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j, andi, j = 1, \ldots, n \end{cases} \tag{2.7}$$

(3) 正态分布假设条件

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2) \\ \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \quad \text{相互独立} \end{cases} \tag{2.8}$$

(4) 维度与样本量限制条件，即$p < n$ 且 p有限

线性回归的重要性

为什么要研究线性回归模型？

1. 应用最为广泛
2. 得到比较深入人心的一般结果—解释度高
3. 很多非线性模型可以转化成线性模型

线性回归模型，通常研究的问题包括

a. 根据样本数据，求参数$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$和$\sigma^2$的估计

b. 对各种假设进行统计检验

c. 根据回归方程进行预测和控制, 以及实际问题的结构分析(即解释).

建立实际问题回归模型

通常分为四步

I 确定研究内容，收集、整理数据

II 确定回归模型形式，变量的选择

III 模型的估计（参数的估计）与检验

IV 模型的应用

这一节内容结合小课题(数据建模)自主学习。
同样的，下一节内容—回归分析应用于发展述评, 结合小课题(文章阅读)自主学习。