《线性回归》
—检查数据

杨　　瑛

清华大学　数学科学系
Email: yangying@mail.tsinghua.edu.cn
Tel: 62796887

2019.03.07

主要内容：检查数据

1. 检查数据
   - 检查数据
   - 一元数据
   - QQ图
   - 二元数据
   - 多元数据
   - 说明
   - R作业
2. 产生模拟数据
   - 产生模拟数据
   - R作业

检查数据

- 数据的图形化检查在数据分析的所有阶段都很重要(参见前面的例子)。
- 检查数据是开始学习和使用R的好方法。

一元数据

- 基本的单变量显示:
  - ✓ 直方图- hist()
    适用于较大的数据集。
  - ✓ 密度估计- plot(density())
    直方图的平滑版本。

用频率估计概率

- 总结主要特征:
  - ✓ 箱线图- boxplot()
    适用于离群值、非对称性以及比较各种分布。

R:

学习和掌握：hist(), plot(),density(), boxplot(). 要搞清楚每一个函数的原理和用法，以及'()'中可以添加什么内容。下同。

QQ图

- 图形工具用来确定样本是否与某个理论分布一致（通常是正态分布）。
- 分布的第p个分位数: 点x使得$P(\mathbf{X} \leq x) = p$ (画图)。
- 样本的第p个分位数: 点x使得$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \leq x) \approx p$
- QQ图中的每个点对应于概率p:
  x坐标：理论分布的第p个分位数
  y坐标：样本的第p个分位数
- 如果样本来自理论分布，那么样本和理论分位数近似相等。因此x和y坐标近似相等。QQ-图看起来像直线$y = x$.

R:

学习R命令：qqnorm(), qqline(), qqplot()

二元数据

- 散点图– plot(x,y)
  - ✓ 为说明趋势：
    添加非参数回归曲线
    lines(loess.smooth(x,y))
  - ✓ 如果有多个点事重叠的话：
    如果点重叠则抖动点
    jitter() 或者手动添加随机噪音

R:

请利用R中的help，进一步了解loess.smooth, loess, jitter等的原理和用法。

多元数据：

- 在有三个变量的情况下:
  - ✓ 三维散点图
    如果您可以交互式地转换图形，这将非常有用.
- 在更多变量的情况下：散点图矩阵- pairs()

R:

pairs().

说明：

- 添加标题和坐标轴标签
  main, xlab, ylab
- 注意坐标轴的范围
  xlim=c(a,b), ylim=c(a,b)
- 添加图例
  legend
- 尝试优化视觉比例
- 多幅图形
  par()

R作业.
要求：简要说明每个命令的含义和用法，在此基础之上做出图形。R-code和作业内容分为两个文件打包提交。

1. 掌握前面slide中R部分中提到的R。自己产生模拟数据，并作图。
2. 设x=runif(30,0,1); y=rnorm(30,0,1); x1=runif(30,0,1); y1=rnorm(30,0,1); x2=runif(30,0,1); y2=rnorm(30,0,1); 试利用这些数据画一张图并用到下面所有的命令：plot(), par(), main, sub, xlab, ylab, xlim, ylim, asp, col, cex, type, lwd, legend, points,lines

产生模拟数据：

- 常见分布随机数的产生。R中有很多这样的函数，例如, rnorm(), runif(), rexp(), rt(), rpois(), rgeom(), rlnorm(), rhyper(), rgamma(), rbeta(), rcauchy(), rchisq(),.....
- 产生[a,b]内n个随机数: 利用sample(a:b,n,replace), 当replace=T时是在[a,b]内有放回的随机抽出n个整数.
- 利用统计推断中学习过的随机数的产生方法, 例如，取舍法，Box-Müller方法，分布函数的反函数方法。
- 多元正态数据和其它多元数据的产生。请查阅文献。

R作业：
要求：与前面一个R作业的要求相同.

- 学习和掌握上一页slide中的方法(可以结合前面的R函数，例如，plot, hist, qqnorm, boxplot 等，以不同的形式展现数据的特征。)