国信证券厦门营业部

R语言学习手册

版本：1.0

国信证券

方莲

2014年7月

国信证券厦门营业部

R 语言学习手册

版本：1.0

方 莲

2014 年 7 月

GUOSEN SECURITY COMPANY

厦 门 营 业 部

保 密 承 诺

本商业计划书内容涉及商业秘密,仅对有投资意向的投资者公开。本公司要求投资公司及相关投资者收到本商业计划书时做出以下承诺:

1. 未经允许不能复印／复制或向第三方详细复述本商业计划书。
2. 未经允许不能自行使用或协助任何第三方使用借鉴本计划书的内容方法从事与本公司有竞争关系的相应领域的商业行为。
3. 未经允许不能自行使用或协助任何第三方使用借鉴本计划书的内容方法从事与本公司有竞争关系的相应领域的商业行为。

目录

插图目录

表格目录

摘　　要

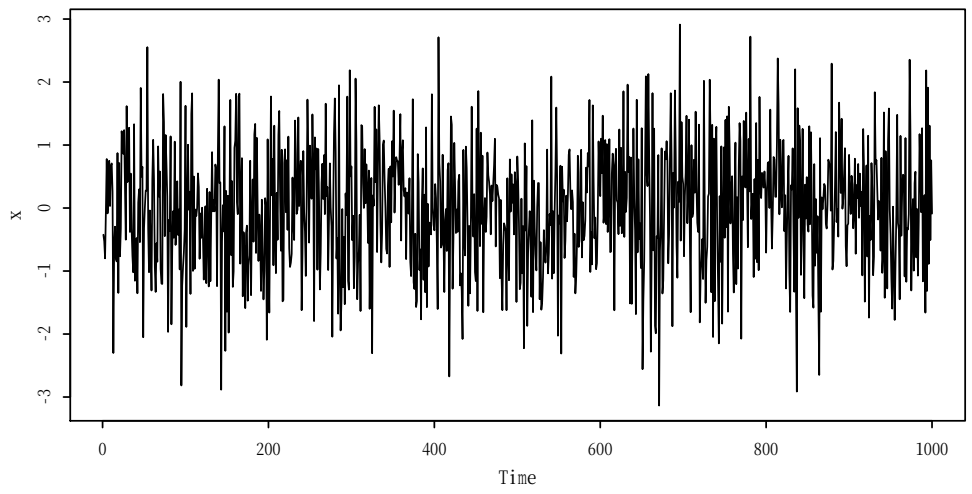
目前 R 在数量金融领域大受欢迎,被广泛的使用在诸如资产定价、收益率曲线估计与预测、

第一章 引言

接触 R 也有一段时间了，大概是从研究生一年级的时候，有授课教师便开始鼓捣我们多多学习编程语言。可是以前也只是为了应付课程作业的需要，「幼稚」地把 R 当作一种简单的统计分析与画图之用，并没有深入编程语言的精髓部分。比如，对于 R 中对象特征的详细讨论，我并没有一个系统化的训练与学习过程。这不得不说是一种巨大的遗憾，因为现在 R 统计语言正发展的如火如荼，大有一统整个数据分析界的势头。而在这样的形势下，如果想要真正掌握这门语言的精髓与要领，也就不得不继续学习，希望通过对底层编程思想的透彻理解，为今后的软件使用、函数编程与数据分析提供扎实的基础。

```
x <- rnorm(1000)

plot.ts(x)
```



1.1 R 是什么

1.2 R 能做什么

1.3 为什么使用 R

1.4 怎么安装

1.5 怎么用 R

第二章 统计与金融

现代金融的发展主要是借力统计技术的提高。从对风险的概率特征的描述开始，整个金融理论体系都是使用概率理论与统计方法来阐述的。Markowitz 的资产组合理论成功运用方差（variance）来刻画一项资产在时间范围内不确定性的量化指标，到最成功的运用统计随机思想的 Black-Scholes 期权定价公式，在整个金融领域都随处可见统计思想。因此，在深入理解整个金融市场之前，我们需要对描述其动态特征的概率统计理论有一定的掌握。

这里我特指滥觞于 Markowitz (1952)、大量使用数学尤其是概率统计理论与随机过程方法的金融理论与实证金融研究

Risk Measure

■ [P1.] Risk means uncertainty in future returns from an investment, in particular, that the investment could earn less than the expted return and even results in a loss, that is, a negative return. Risk is often measured by the standart deviation of the return, which we also call the *volatility*. Recently there has been a trend toward measuring risk by value-at-risk (VaR) and expected shortfall (ES). These focus on large losses and are more direct indications of financial risk than the standard deviatin of the return.

对于为何我们在处理金融实证研究时主要是针对资产（股票、债券、衍生品等）回报率，可以参考经典的 Compbell,Lo and Makinley 的"The Econometrics of Financial Markets"已经解释的十分到位。这是因为：

- 收益率衡量的是在单位投资数量上所获得的投资回报，是一个相对的概念，与投资规模大小没有关系。因而这有利于比较不同投资规模、投资期限的资产给予投资者的风险补偿。
- 收益率与股票价格相比在统计特征上更加方便。首先是收益率反映了股票价格微小的变动情况，比较服从随机过程对数据平稳性的要求；其次，收益率还是一个

2.1 需要注意的问题

Ruppert 提到的几个在处理量化分析时需要注意的问题，我个人觉得都十分的在理。

模型是对现实的简要、精辟而不失去其关键信息的数学描述。有关如何将统计模型运用到数量金融分析当中，我们应该多加思考，而不是一味的盲目追随模型。

- 永远都要把数据放在首要位置，只有对数据做出合理的解释才能进一步去拟合各种理论。这里我们可以使用 R 中非常多的优秀做图软件包，如 ggplot2()。Ruppert 提

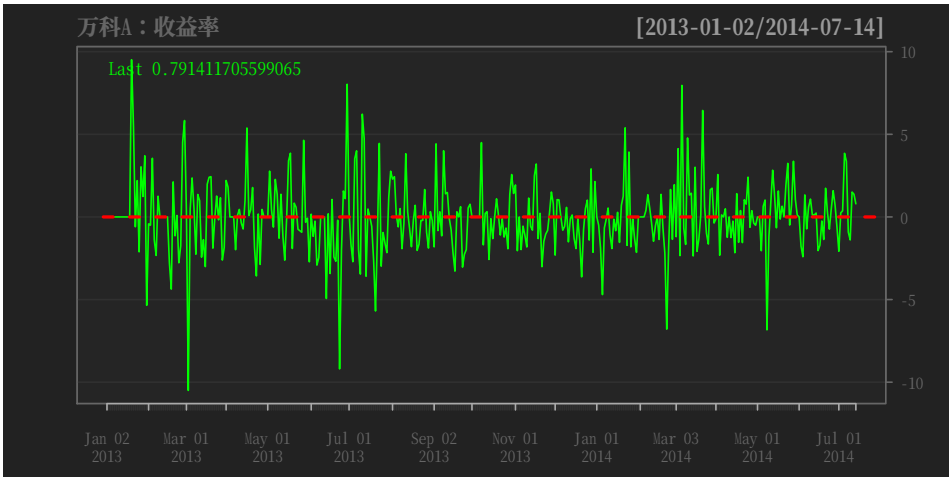
到说我们可以使用绘图来检测出「好的数据」和「坏的数据」以及「异常值」(outlier)。对于「bad data」, 要么是尽量去修复使其符合需要, 要么就是果断的将其从样本中删除, 因为这样的「坏数据」不仅无益于结果分析, 反而会提供干扰信息。而对那些看起来是「异常」的数据, 我们需要格外提高警惕。比如, 从金融危机当中产生的「outlier」, 反映了由于市场的结构调整与重新构建而出现的与先前模式完全不同的特征, 这些新的数据集要求我们为此提供具有针对性的统计模型。比如, 现在比较前沿的统计极值理论, 可以运用于对极端风险的建模, 这些在金融理论文献当中通常被归类为「Rare Disaster」。从一定意义上说, 正是由于我们通过图形发现的「异常特征」数据, 才启发学术界与实务界不要止步于现有的数学模型, 而是应该激流勇进、迎刃而上, 提出更加可信的理论模型。

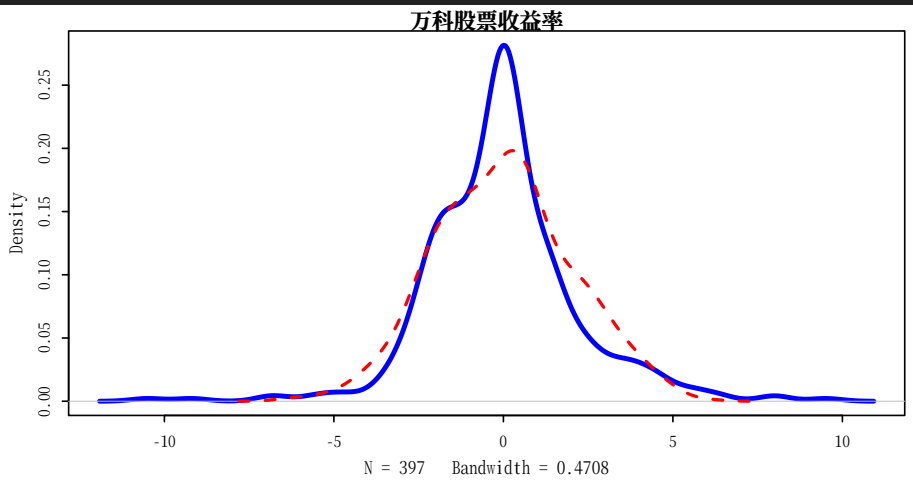
- 时刻警惕模型的适用性。Box 有句很出名的话:『All models are false but some are usefull』。意思是说, 没有任何一个模型能够保证其完全正确性, 但是一个科学「有用」的模型必须能够为解释某个现象、某个问题而提出针锋相对的思路, 为我们理解问题找到合理而可行的切入点。
- 偏差 -方差的权衡。我们知道, 在统计理论里讲究对变量估计的无偏性、相合性。所谓的无偏, 指的是变量的估计值等于其期望值, 是对真是数值的完美拟合。可是真正做到无偏估计十分的困难, 有时是难以找到这样的参数估计, 有时是因为纯粹意义上的无偏估计并没有合理的解释。因此, 我们面临这相合性的束缚, 即估计值对期望值收敛的逼近程度。许多的情况是, 无偏性与相合性是一对矛盾的产物, 必须在两者之间做出一定的取舍。一个总的原则是采用「简约」(principle of parsimony)原则, 即用尽量少的参数来取得尽可能小的误差范围。
- 模型的不确定性。我们知道, 任何一个模型对只是对问题的无限逼近的选择, 由经济体自发生成系统均衡时的客观数据无法道出其产生的内在机制。我们经济计量学家也只是通过这些显现的测量数据来「推断」其内在的随机过程。因此, 任何一个模型都具有不完美的内在缺陷。模型的这样不确定性分析要求我们通过比较不同模型的分析结果来尽可能的选择一个合理的模型。现在, 我们已经可以借助 Baysian 方法来比较多个复杂模型的分析结果。
- 金融数据并不一定(甚至完全不是)**正态分布**。我们原来在课本里学习到的大多数金融理论都把资产回报率假设为一个服从正态高斯分布的随机过程, 从而可以利用高斯分布计算风险概率及相应的方差。可是真实的金融数据并非是一个服从常态的高斯分布。比如, 股票的回报率往往具有「偏峰」(skewness)、「后尾」(fat tail)的特点, 左右两边的分布是不对称的, 而且表现出波动率聚集(volatility clustering)。很明显的, 这些特征有别于正态高斯分布。往往的, 一个可能的备选方案是 t-distribution。

```
rtn <- function(stock.code, from = "2013-01-01") {
    y <- getSymbols(stock.code, from = from, auto.assign = FALSE)
    rtn <- diff(log(Cl(y)), 1) * 100
    rtn <- rtn[-1]
}

chartSeries(rtn("000002.sz"), name = "万科A: 收益率")
abline(h = 0, col = "red", lwd = 2, lty = "dashed")

plot(density(rtn("000002.sz")), main = "万科股票收益率", col = "blue", lwd = 3)
lines(density(rnorm(1000, 0, 2)), col = "red", lwd = 2, lty = "dashed")
```





- 金融数据的方差往往不是固定的，而是时变的。我们知道，正太高斯分布有两个参数，$\mathcal{N}(\mu, \sigma^2)$。如果假定收益率服从高斯分布，则暗含着假定其方差是固定的。可是我们往往看到金融数据的方差是随着时间变化的，尤其在一些特定的时点上，股票市场的波动往往高的惊人。比如，1987 年的黑色星期四，整个 NSE 股票下跌了近 30%。为了处理时变的方差，我们需要引入如 ARCH、GARCH、Voatility Model 等。

```
chartSeries(rtn("^GSPC", from = "1980-01-01"), name = "Change in S&P500")  ## 画图
abline(h = 0, col = "red", lwd = 2, lty = "dashed")
```