

笔试答卷

侯怡枫

2017 年 2 月 10 日

答卷中用到的源代码都将在 Email 附件中附上。所有程序都在 Linux Ubuntu LTS 16.04 操作系统上编译。

第 1 题

这一题的源代码在 ans1.R 文件中。首先用 R 将数据绘图获得一个直观的感受。

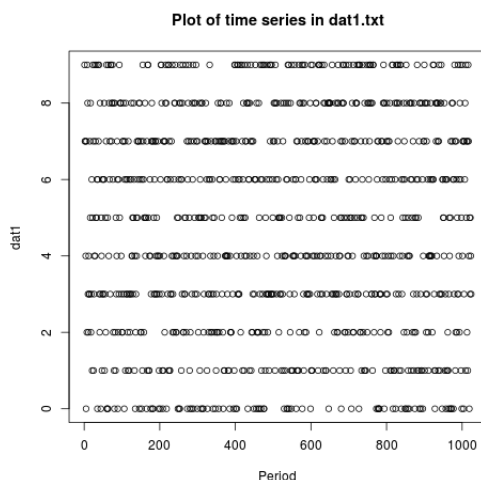


图 1: dat1.txt 中数据的时间序列图

根据图 1，改时间序列并不呈现确定性时间趋势（deterministic time trend）。然后，图 2 展示改题数据的样本 ACF。图 3展示改题数据的样本 PACF。

从图 2中可以看到，该序列和 $k = 1$ 的滞后项的样本相关系数接近 1。而大于 1 的滞后项的相关系数则呈指数化递减。从图 3中可以看到，偏自相关函数在 0 附近浮动。这显示此序列可能由一个自回归（AR）过程产生。

答题者选择“存在单位根”作为原假设。相应的备择假设是“产生改序列的随机过程是平稳的”。答题者计划使用 ADF（augmented Dickey-Fuller）检验。这里假设该序列是由一个自回归（AR）模型产生。为使用 ADF 检验，首先要选择一个信息标准（Information Criterion）来确定自

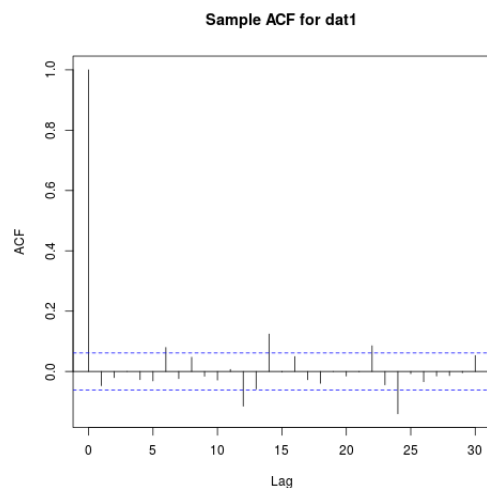


图 2: dat1.txt 中数据的样本 ACF plot

回归的滞后步长 (lag)。这里答题者选择 AIC 进行模型选择。对本题时间序列运行 AIC 后确定的滞后步长是 24。

在确定滞后步长后，运行 ADF 检验后得到的 Dickey-Fuller 统计值是-7.6675，对应的 p 值小于 0.01。所以结论是拒绝“存在单位根的原假设”。单位根检验在金融时间序列分析和量化投资中非常重要。**统计套利**这一类策略成功的前提假设是金融资产的价格倾向于回归到均值。但是对一个存在单位根的价格序列来说这个前提是不成立的。

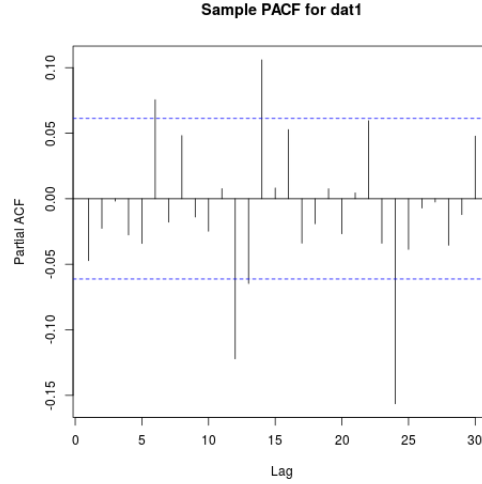


图 3: dat1.txt 中数据的样本 PACF plot

第四题

本题涉及的代码在“ans4.r”文件中。

Hurst 是用来度量时间序列的长期记忆性。Hurst 指数的值在 0.5 到 1 之间意味着长期的正相关性，简单来说意味着序列中一个较高的值往往会跟随者另一个较高的值。Hurst 指数的值在 0 到 0.5 之间意味着长期的负相关性。Hurst 指数的值等于 0.5 则意味着一个完全不相关的序列。Hurst 指数的定义如下：

$$E\left[\frac{R(n)}{S(n)}\right] = Cn^H, n \rightarrow \infty \quad (1)$$

其中 $R(n)$ 是前 n 个值的值域， $S(n)$ 是标准差， $E[x]$ 是期望值， n 是观察的期数， C 是一个常数。对本题上证指数序列，答题者使用 R/S 方法进行参数估计和假设检验。首先假设整个序列的观察数量是 N ，将其等分为每个长度为 n 的子区间。定义 A 为 $\frac{N}{n}$ 的整数部分，也就是子区间的数量。具体方法如下，（参考 [?]）：

1. 计算均值：

$$m = \frac{1}{n} \sum_{i=1}^n X_i \quad (2)$$

2. 计算均值调整后的序列：

$$Y_t = X_t - m, \quad t = 1, 2, \dots, n \quad (3)$$

3. 计算累计偏离序列：

$$Z_t = \sum_{i=1}^t Y_i, \quad t = 1, 2, \dots, n \quad (4)$$

4. 计算值域 R ：

$$R(n) = \max(Z_1, Z_2, \dots, Z_n) - \min(Z_1, Z_2, \dots, Z_n) \quad (5)$$

5. 计算标准差 S :

$$S(n) = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - m)^2} \quad (6)$$

6. 计算每一个子区间的 $R(n)/S(n)$ 值并求平均。

7. 最后把 $\log(n)$ 作为控制变量, $\log[R(n)/S(n)]$ 作为应变量进行线性回归, 估计斜率并进行假设检验。这里的 \log 函数定义为以 2 为底。

首先, 仅仅是为了简化计算, 这里采用最近的 4096 (2^{12}) 天的收盘价数据作为观察样本。将 $N = 4096$ 个数据划等分分成 2, 4, ..., 2048 个区间。对每一种划分使用如上算法计算出 $R(n)/S(n)$ (作为 n 的函数) 的值并绘图, 见图 4

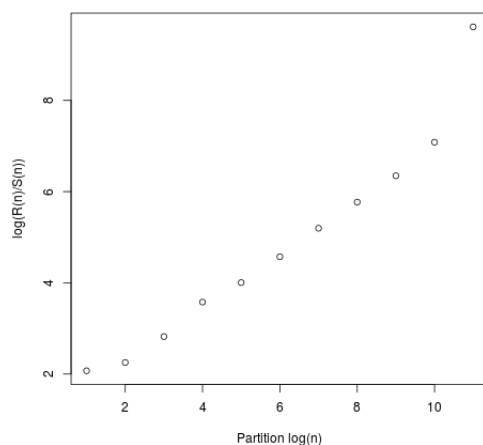


图 4: $\log \frac{R(n)}{S(n)}$

用 OLS 估计出的斜率 (H 值) 为 0.71771 (长期正相关)。对于统计检验我们选择 “ $H = 0.5$ (没有长期相关性)” 作为原假设。计算的 p 值为 1.633015e-05。误差项的有限分布的推到是一件非常困难的工作, 具体可参考文献。所以这里的假设检验的前提是回归的误差项是正态分布的。