自相关性

自相关性

观测样本 $x_1, \dots, x_n$ 的阶数为 $h \in \mathbb{N}$ 的样本自相关系数定义为：

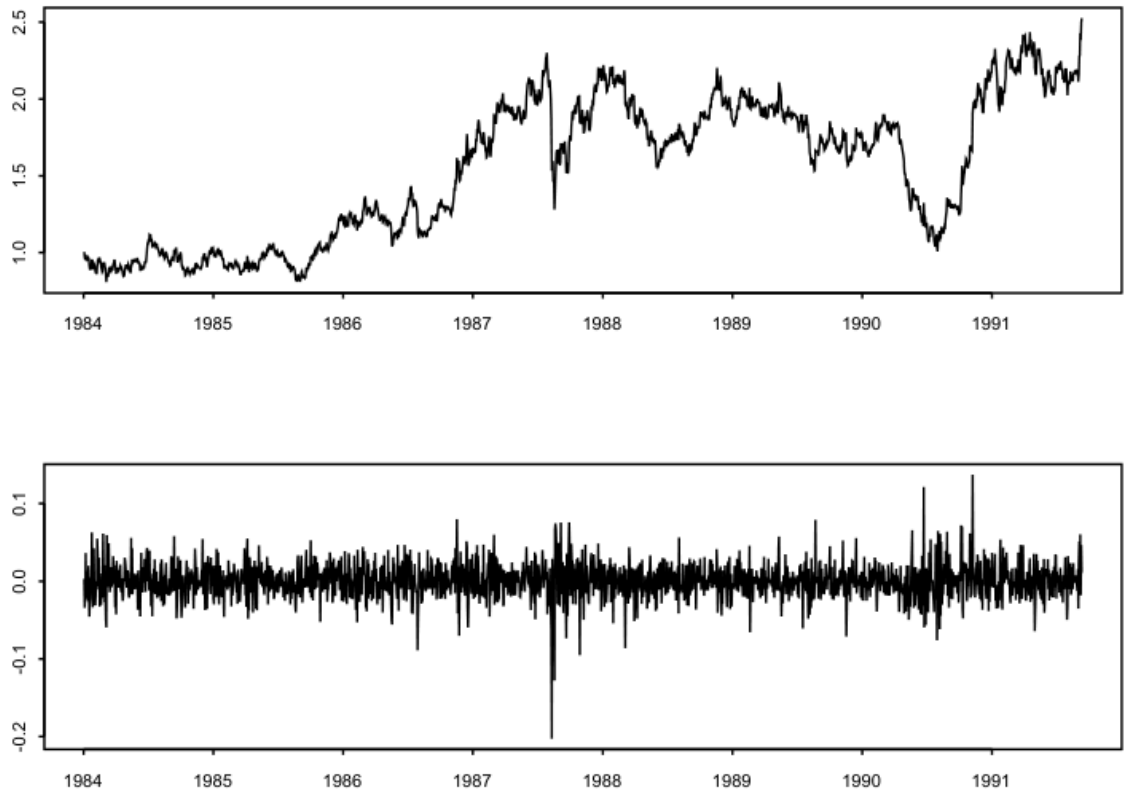
$$r_x(h) = \frac{\sum_{i=1}^{n-h} (x_{i+h} - \bar{x})(x_i - \bar{x})}{(n-h)s_x^2}$$

这里注意 $\sum_{i=1}^{n-h}$ 上面为 $n-h$ 的原因是希望第一组是 x_1, x_{1+h}，最后一组是 x_{i-h}, x_i.
对于点 (x_i, x_{i+1})（$i = 1, \dots, n-1$）的样本相关系数，实际上就是阶数为 1 的样本自相关系数。

当数据点 x_i 的索引 i 对应时间参数时，这些系数特别有趣，并且数据可能表现出时间依赖性，也就是我们常常说的时序分析。此时，我们测量的是变量 X_i 和 X_{i-h} 之间相隔 h 时间点的相关性。

例 股票价格

下图顶部图像显示了 1984 年至 1991 年期间纽约证券交易所惠普公司股票的价格变化。绘制了连续交易日的收盘价 a_i（$i = 1, 2, \dots, 2000$）；在图中，这些数值通过线性插值绘出。



由于股票价格通常形成指数增长（或下降）的序列，通常分析的是定义为

$$x_i = \log \frac{a_i}{a_{i-1}}$$

的“对数回报率”，而不是直接分析股票价格。这些值在上图的底部图像中绘出。由于 x_i 的索引 i 对应于第 i 个交易日，因此根据 $x_1, \ldots, x_{2000}$ 不太能很好地为独立随机变量 $X_1, \ldots, X_{2000}$ 建立一个模型。毕竟，第 i 日的显著变化会极大地影响第 $i+1$ 日的变化。

然而，经济计量学长期以来接受了相反的假设——股票并不是“随机游走”。

验证这一假设的第一步是计算序列 $x_1, \ldots, x_{2000}$ 的样本自相关性。下图左侧显示了这些样本自相关性系数，其中横轴为 $h = 0, 1, 2, \ldots, 30$，线段的高度表示对应的阶数 h 的样本自相关性系数（0 阶样本自相关性显然为 1）。几乎所有的样本自相关性系数都很小，这表明对数回报率显示出很少的线性相关性。

右图显示的是对数回报率平方 $x_1^2, \ldots, x_{2000}^2$ 的样本自相关性系数。虽然这些系数很小，但就这么说对数回报率的平方几乎没有相关性，那还是有相当多争议的：明显有太多系数明显偏离 0。因此，不能将 $x_1, \ldots, x_{2000}$ 简单地建模为独立变量的实现：时间效应应被考虑在内。股票价格并不形成随机游走。

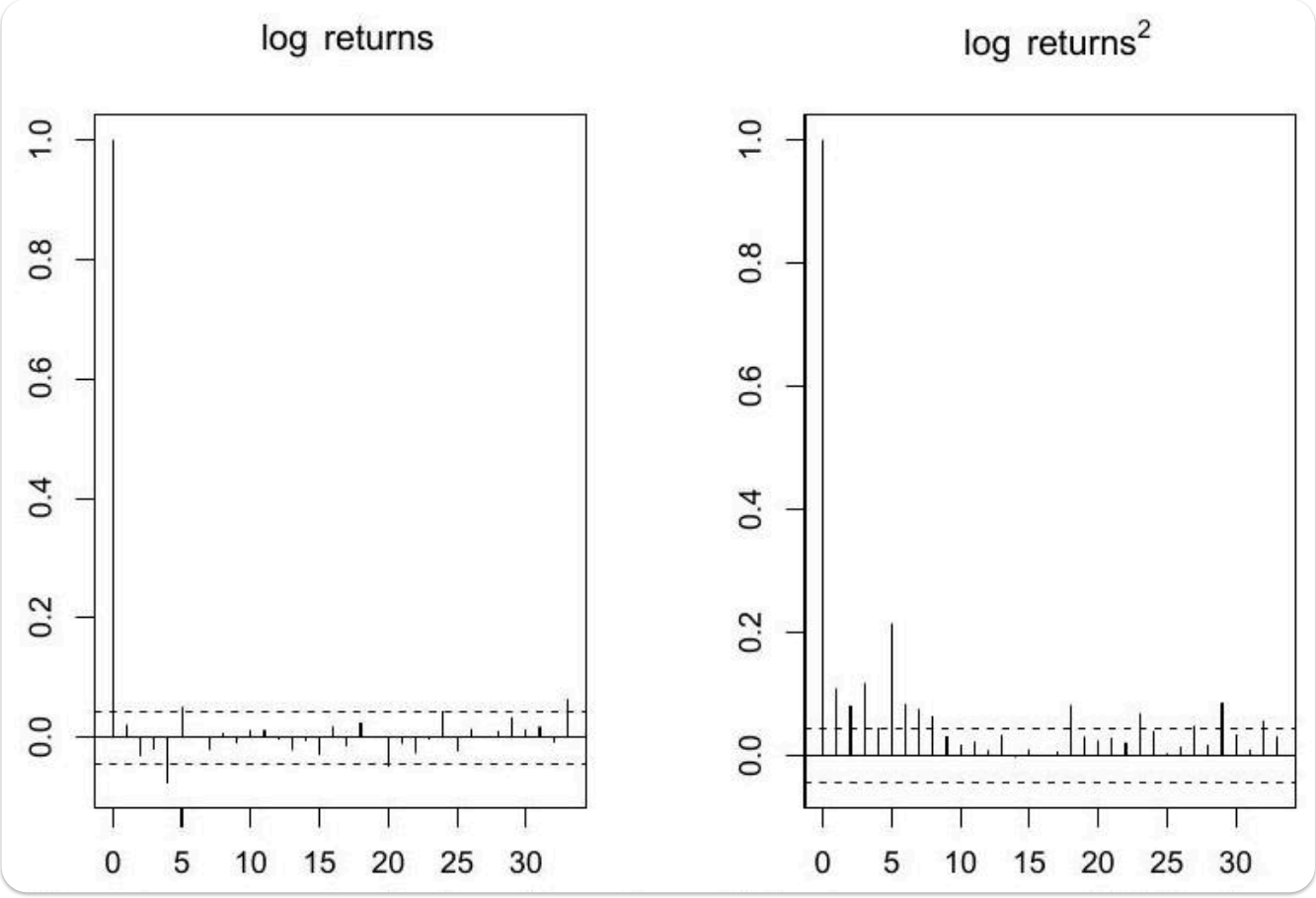


图: 1984 年至 1991 年期间惠普公司股票对数回报率的样本自相关性函数（左图）及其平方的样本自相关性（右图）。虚线表示高度 $\pm 1.96/\sqrt{2000}$。

上文中，我们提到左侧图中的系数“较小”，而右侧图中的系数“偏离 0”较多。我们可以通过假设检验方法客观地支持这些结论。图中的虚线给出了样本自相关系数作为样本变量时的临界值，以零假设为 $x_1, \ldots, x_{2000}$ 可视为独立变量的样本（误差范围为 5%）。落在虚线外的系数意味着需要拒绝零假设。需要注意的是，当我们假设零假设时，由于 5% 的误差范围，大约 $1/20$ 的系数可能因为“随机变化”而落在虚线外。在右图中，有太多值落在虚线之外，那么显然我们假设被否决了。