时间序列分析 自协方差/自相关系数/偏自相关系数

随风
大数据、人工智能

关注他

105 人赞同了该文章

一组数据需要观察的话，我们需要了解一下他们的组成结构，正如我们要了解原子、分子、电子等的结构一个道理。

以 Z_t 表示一组数据，或一个时间序列。

（一）通用的几个基本概念：均值、方差、标准差、协方差、相关系数

1、均值

均值（期望）是统计学中最常用的统计量，用来表明数据集中相对集中较多的中心位置。

数学表示： $u_t = E(Z_t)$

2、方差

方差是用来度量一组数据的离散程度。概率论中方差用来度量随机变量和其期望（即均值）之间的偏离程度。统计中的方差（样本方差）是每个样本值与全体样本值的平均数之差的平方值的平均数。

数学表示： $\delta_t^2 = E(Z_t - u_t)^2$

3、标准差

标准差（均方差），是离均差平方的算术平均数的平方根，用σ表示。标准差是方差的算术平方根。标准差能反映一个数据集的离散程度。

数学表示： $\delta_t = \sqrt{\delta_t^2}$

4、协方差

协方差用来度量两个变量各个维度偏离其均值的程度，这与只表示一个变量误差的方差不同。协方差的值如果为正值，则说明两者是正相关的(从协方差可以引出“相关系数”的定义)，结果为负值就说明负相关的，如果为0，也是就是统计上说的“相互独立”。

数学表示： $cov(Z_{t1}, Z_{t2}) = E(Z_{t1} - u_{t1})(Z_{t2} - u_{t2})$

假设有两个随机变量X、Y，大致上有：

(1)若协方差为正数：X增大，Y增大；X减小，Y减小，即变化趋势相同。

(2)若协方差为负数：X增大，Y减小；X减小，Y增大，即变化趋势相反。

(3)若协方差为零：X与Y的变化没有任何关系。

5、相关系数

相关系数是研究变量之间线性相关程度的量。求出协方差之后，我们考虑一个问题就是协方差对应这每一个“协”关系，他们对应得比值是多少，所谓对应的比值可以理解为每一个“协”距离整体的距离比值是百分之几？两个的“协”对应他们的整体距离的比值是百分之几就能够表示他们之间有多相关，这个相关系数越大，表示这两个数值越有关系。可以理解为，如果两个序列，一个是3000多这个基数去变动，一个是10000多这个基数去变动，他们的绝对数据肯定是不一样的，但是他们的变动比率是一样的，所谓相关性也可以理解为把两个值统一化，在同一个维度来评价这两个值的协方差关系，因此在同一个维度来衡量这两个值的协方差关系就叫做相关系数。

数学表示：$$r(Z_{t_1}, Z_{t_2}) = \frac{cov(Z_{t_1}, Z_{t_2})}{\sqrt{\delta_{t1}^2}\sqrt{\delta_{t2}^2}}$$

相关系数的绝对值越大，相关性越强：相关系数越接近于1或-1，相关度越强，相关系数越接近于0，相关度越弱。通常情况下通过以下取值范围判断变量的相关强度：

(1) 0.8-1.0 极强相关

(2) 0.6-0.8 强相关

(3) 0.4-0.6 中等程度相关

(4) 0.2-0.4 弱相关

(5) 0.0-0.2 极弱相关或无相关

时间序列的特点是一维，因此如果借用统计学上面的指标衡量，有些不太适宜。根据时间序列的特点，形成了自协方差、自相关函数、偏自相关函数。看到前面都加了一个“自”，原因是时间序列没法在找到一个别的数据和自己来进行比较；只能自己和自己来比较，自己和自己慢几拍（滞后期）的这些数据进行比较，所以加入了一个“自”。

（二）时间序列自有的几个基本概念：自协方差、自相关系数、偏自相关系数

1、自协方差

在统计学中，特定时间序列或者连续信号 Z_t 的自协方差是信号与其经过时间平移的信号之间的协方差。

数学表示：

$$r(k) = \frac{1}{n}\Sigma_{t=k+1}^{n}(Z_t - \bar{Z})(Z_{t-k} - \bar{Z})$$

可以认为自协方差是某个信号与其自身经过一定时间平移之后的相似性，自协方差 $r(k)$ 就表示了在那个时延的相关性。

2、自相关系数（ACF）

自相关系数度量的是同一事件在两个不同时期之间的相关程度，形象的讲就是度量自己过去的行为对自己现在的影响。

数学表示： $$ACF(k) = \Sigma_{t=k+1}^{n}\frac{(Z_t - \bar{Z})(Z_{t-k} - \bar{Z})}{\Sigma_{t=1}^{n}(Z_t - \bar{Z})^2}$$

自相关（autocorrelation），也叫序列相关，是一个信号于其自身在不同时间点的相关度。非正式地来说，它就是两次观察之间的相似度对它们之间的时间差的函数。它是找出重复模式（如被噪声掩盖的周期信号），或识别隐含在信号谐波频率中消失的基频的数学工具。它常用于信号处理中，用来分析函数或一系列值，如时域信号。

3、偏自相关系数（PACF）

根据ACF求出滞后k自相关系数 $ACF(k)$ 时，实际上得到并不是Z(t)与Z(t-k)之间单纯的相关关系。

因为Z(t)同时还会受到中间k-1个随机变量Z(t-1)、Z(t-2)、……、Z(t-k+1)的影响，而这k-1个随机变量又都和z(t-k)具有相关关系，所以自相关系数里面实际掺杂了其他变量对Z(t)与Z(t-k)的影响。

为了能单纯测度Z(t-k)对Z(t)的影响，引进偏自相关系数（PACF）的概念。对于平稳时间序列{Z(t)}，所谓滞后k偏自相关系数指在给定中间k-1个随机变量Z(t-1)、Z(t-2)、……、Z(t-k+1)的条件下，或者说，在剔除了中间k-1个随机变量Z(t-1)、Z(t-2)、……、Z(t-k+1)的干扰之后，Z(t-k)对Z(t)影响的相关程度。

数学表达：

$$PACF(k)=\frac{E(Z_t-EZ_t)(Z_{t-k}-EZ_{t-k})}{\sqrt{E(Z_t-EZ_t)^2}\sqrt{E(Z_{t-k}-EZ_{t-k})^2}}=\frac{cov[(Z_t-\bar{Z}_t),(Z_{t-k}-\bar{Z_{t-k}})]}{\sqrt{var(Z_t-\bar{Z}_t)}\sqrt{var(Z_{t-k}-\bar{Z_{t-k}})}}$$

计算某一个要素对另一个要素的影响或相关程度时，把其他要素的影响视为常数，即暂不考虑其他要素的影响，而单独研究那两个要素之间的相互关系的密切程度时，称为偏相关。

总结：时间序列借用统计学的数据结构分析公式

（1）期望还是等与期望

（2）自协方差 = 协方差（期望用整个时间序列的期望，一个期望）

（3）自相关系数 = 相关系数（期望用整个时间序列的期望，一个期望）

（4）偏自相关系数 = 相关系数（期望用各自序列的期望，两个期望）