时间序列 Arima 模型理论及在 Data Studio 中的应用

目 录

1 Arima 模型理论

Arima 模型是以加权的方法对白噪声的组合来建立模型的，并以模型和实际数据的残差服从均值较小的正态分布为目标。

1.1 平稳序列建模

所谓的平稳性是指某一时间序列是由同一个随机过程生成的，即时间序列 x(t)(t=1,2,3,…..)的每一个数值都服从同一个随机分布，该随机分布生成的事件序列满足以下条件：

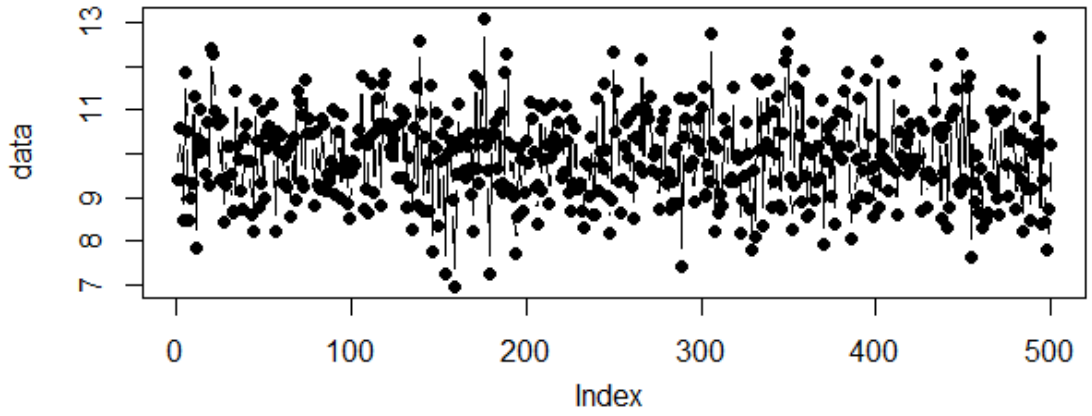
均值与时间 t 无关，任何 k 阶滞后序列的均值都相同

方差与时间 t 无关，任何 k 阶滞后序列的方差都相同

X(t)与 X(t-k)的协方差只与滞后阶数 k 有关，与时间 t 无关。

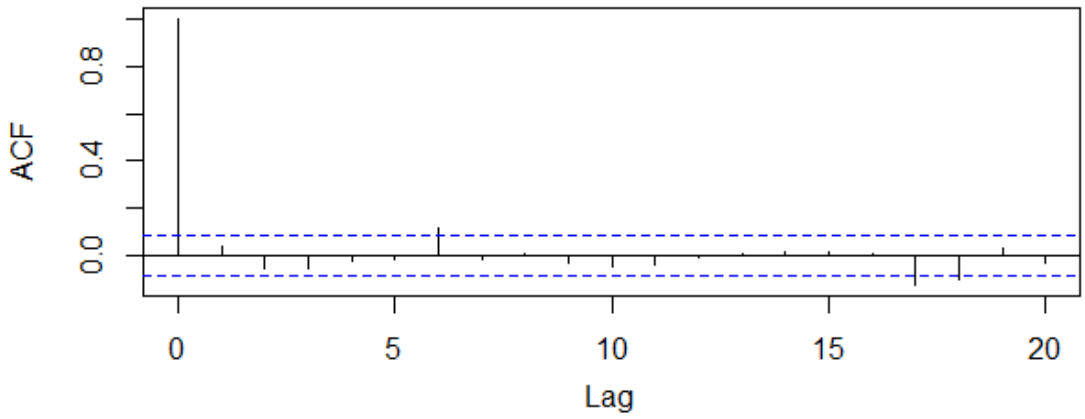
自相关系数只与滞后阶数 k 有关，与 t 无关。

从直观上看，所谓的平稳过程可以通过查看 acf 自相关图来确定，如果 acf 成指数级衰减，则表示平稳（如下图）。



平稳序列

上图是一个有均值为 10，标准差为 1 的正态分布生成的时间序列。acf 自相关图用来展示 X(t)时间序列与 X(t-1)，X(t-2),..，X(t-k)各阶之后时间序列之间的相关系数。如果 k 阶滞后序列 X(t-k)与原始序列 X(t)的相关系数不在[-0.2,0.2]之间，则称 X(t-k)与 X(t)不具有相关性。如果一个时间序列的任何一阶滞后序列与原始序列都不具有相关性，则该时间序列不具有自相关性。



对于一个平稳性的时间序列，如果其均值为 0，且不具有自相关性，这样的平稳性时间序列为白噪声序列。

MA 滑动平均过程与 AR 自回归过程

q 阶滑动平均过程（MA(q)）是把若干白噪声 e 做加权得到的，其公式如下：

$$Y_t = e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}$$

p 阶自回归过程（AR(p)）是使用序列本身做为变量，并对其加权得到，公式如下：

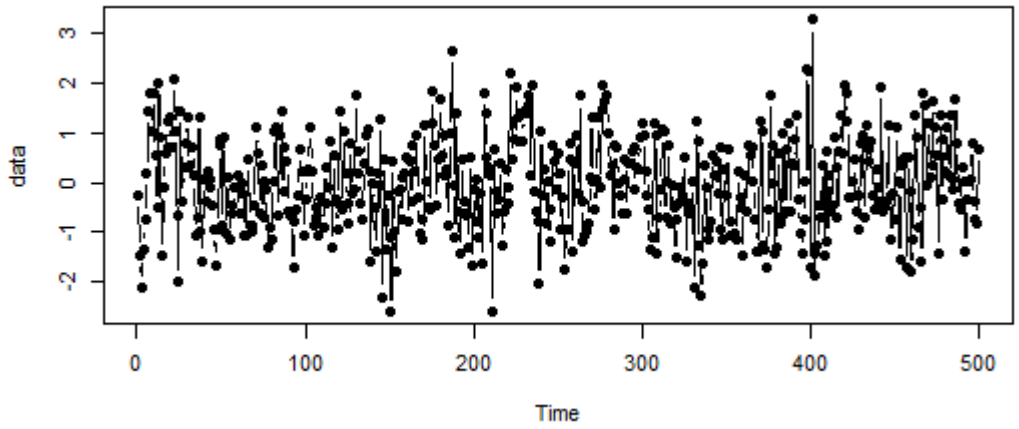
$$Y_t = \varphi_1 Y_{t-1} + \varphi_1 Y_{t-1} + \cdots + \varphi_1 Y_{t-1} + e_t$$

ARMA 自回归滑动平均混合模型

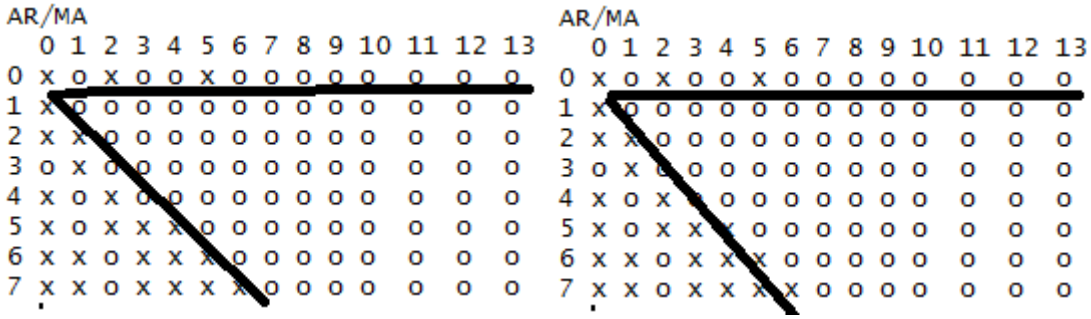
在实际数据分析中很少有指标数据可以单纯用 AR 或者 MA 来建立模型，通过把平均混合模型 MA 和自回归模型 AR 进行叠加，就组成了自回归滑动平均混合模型 ARMA（p，q），其公式如下：

$$Y_t = e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + \varphi_1 Y_{t-1} + \varphi_1 Y_{t-1} + \cdots + \varphi_1 Y_{t-1} + e_t$$

下图为模拟一个 ARMA（1,1）生成的序列



在 AR 和 MA 中可以通过观察 acf 与 pacf 的拖尾和截尾来得到 AR 和 MA 的阶数，而对于 ARMA 的 acf 与 pacf 图同时表现出拖尾的特征，为了确定参数 p、q 可以使用混合自相关图



观察得到 p、q 的两种可能的取值，具体取值通过比较分别建模后的质量确定。

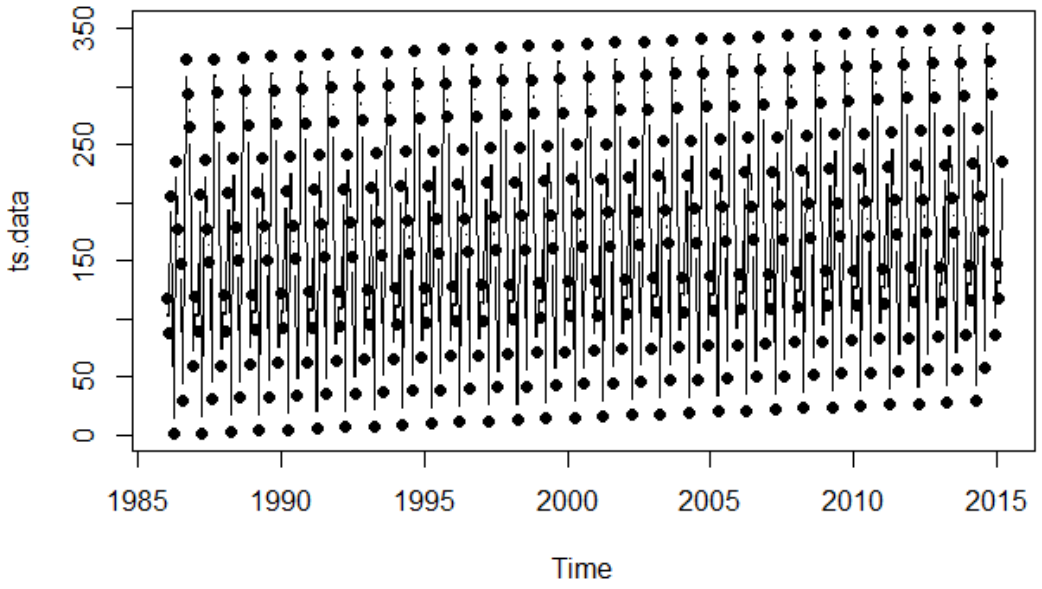
1.2 非平稳时间序列的 ARIMA 过程

对于不能用 ARMA 等模型建模的非平稳时间序列，要进行平稳化处理，在 ARMA（p，q）的基础上引入了差分阶数 d 的概念。建立 ARIRMA（p,d,q）模型。

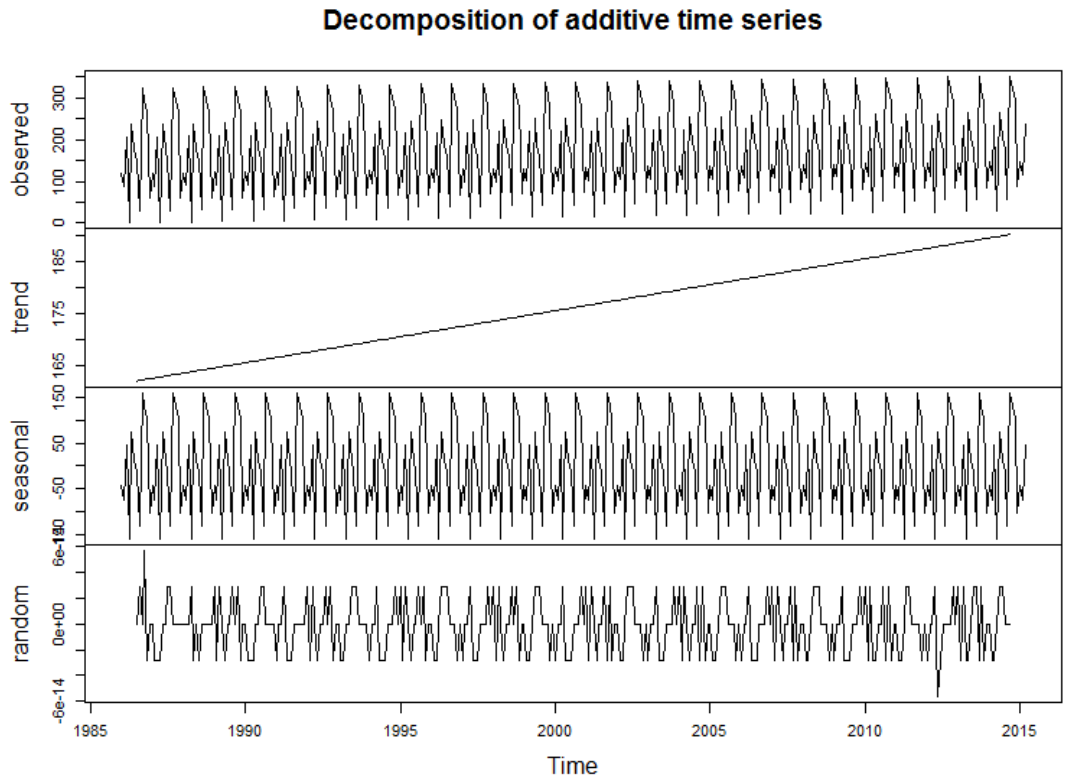
常用的平稳化处理手段有取对数和差分。对非平稳的时间序列 data 如果 d

阶差分 diff（data，d）为平稳序列，则可以建立 ARIMA（p，d，q）模型，一般情况下 d 最多可以取到 3，避免过差分。如果仅仅差分处理不能平稳化，则可以先取对数，再差分 diff（log（data），d），大部分非平稳序列都能经过取对数和差分处理转化为平稳序列。再根据 ARMA 建模方法确定其他参数。

下图是 1986 年 1 月至 2015 年 3 月至今每桶原油的月度价格。可以见看出按每年有较强周期性，并且逐年上升，平稳模型对于这样的时间序列并不适用，选择使用 ARIMA 模型建模是可行的。



通过 R 语言的 decompose 函数吧时间序列分解为长期趋势和周期性变化。从 trend 可以看出，确实存在一个向上的趋势。



使用 shapiro.test 检验 diff（trend，d）直到 diff（trend，d）或者 diff（log（trend，d））为正态分布。确定 d 后使用混合自相关图 eacf 确定可能的 p 和 q，建立模型。

2 ARIMA 在 Data Studio 中的应用

Data Sutdio 中使用了 ARIMA 模型，需要的参数定义如下

```
<parameter-string key="column.date" description="日期列"
optional="false" expert="false" name="日期列" />
      <parameter-string key="column.target" description="目标列"
optional="false" expert="false" name="目标列" />
      <parameter-int key="column.targetindex" description="目标列序号"
optional="false" expert="false" name="目标列序号" />
      <parameter-int key="trainpercent" description="训练数据集百分比"
optional="false" expert="false" name="训练数据集百分比"/>
      <parameter-category key="frequency" description="时间周期，比如 年，
季度，月" default="1" name="时间周期">
          <value>1</value>
          <value>12</value>
          <value>4</value>
```

```
            <value>7</value>
        </parameter-category>
        <parameter-int key="startyear" description="起始年"
optional="false" expert="false" name="起始年" />
        <parameter-int key="startquarter" description="起始季度"
optional="false" expert="false" name="起始季度" />
        <parameter-int key="startmonth" description="起始月"
optional="false" expert="false"  name="起始月" />
        <parameter-int key="startindex" description="起始星期"
optional="false" expert="false" name="起始星期"/>
        <parameter-int key="startday" description="起始天"
optional="false" expert="false" name="起始天"/>

        <parameter-boolean key="auto" description="使用默认的参数"
default="false" optional="true" expert="false" name="使用默认的参数"/>
        <parameter-int key="p" description="自回归阶数" optional="true"
expert="false" name="移动平均阶数(p)"/>
        <parameter-int key="d" description="差分阶数" optional="true"
expert="false" name="差分阶数(d)"/>
        <parameter-int key="q" description="移动平均阶数" optional="true"
expert="false" name="移动平均阶数(q)"/>
        <parameter-int key="P" description="季节自回归阶数"
optional="true" expert="false" name="季节自回归阶数P)"/>
        <parameter-int key="D" description="季节差分阶数目"
optional="true" expert="false" name="季节差分阶数(D)"/>
        <parameter-int  key="Q"  description=" 季 节 移 动 平 均 阶 数 "
optional="true" expert="false" name="季节移动平均阶数((Q)"/>
```

- Column.data 参数选择数据中的一列，用来指示数据的时间信息
- Column.target 指定用于分析的数据列
- Trainpercent 设置数据集中作为训练数据的百分比
- Frequency 周期，供选择的有 1,4,7,12；分别代表年，季度，周，月
- 参数 p ，d ，q 分别代表 ARIMA（p,d,q）de 自回归阶数，差分阶数和滑动平均阶数。
- 参数 P，D，Q 代表季节性趋势的 p，d，q
- XML 中的起始年月星期等信息是通过数据集中选择的日期列中最小的日期计算得到的。

在 Data Studio 中对应设置如下