3、数据预处理

3-1、数据清洗

对LBMA-GOLD.csv和BCHAIN-MKPRU.csv文件中的数据进行分析，数据基本信息如下：

*表格或图片*

对LBMA-GOLD.csv中的缺失值进行填充处理，我们采取前后两天的均值进行填充的策略。

3-2、数据可视化

为了更好的观察黄金和比特币的价格变化趋势，我们对LBMA-GOLD.csv和BCHAIN-MKPRU.csv文件中的数据进行可视化得到图1和图2

图1

图2

黄金价格变化趋势：

比特币价格变化趋势：

3-3挖掘时间序列

为了后续采用时间序列模型ARIMA进行数据预测，我们通过对LBMA-GOLD.csv和BCHAIN-MKPRU.csv文件中的数据及其转化后得到的数据进行平稳性检验与白噪声检验，以此来挖掘有意义的时间序列。

3-3-1 平稳性检验

首先对原始数据进行平稳性检验，使用观察图像与单位根检验两种方法进行相互对照。

可视化原始数据得到图3

图3

进行单位根检验，结果1如下：

结果1

结论：

再对原始数据进行一阶差分得到一阶差分数据，同样使用两种方法进行检验，

可视化一阶差分数据得到图4

图4

进行单位根检验，结果2如下：

结果2

结论：

最后对原始数据进行二阶差分得到二阶差分数据，同样使用两种方法进行检验，

可视化二阶差分数据的到图5

图5

进行单位根检验，结果3如下:

结果3

结论：

最终结论：

3-3-2 白噪声检验

我们需要判断数据是否是白噪声，白噪声没有研究的意义。

我们采用Ljung-Box test

首先对原始数据进行检验，结果4如下：

结果4

结论：

再对一阶差分数据检验，结果5如下：

结果5

结论：

最后对二阶差分数据检验，结果6如下：

结果6

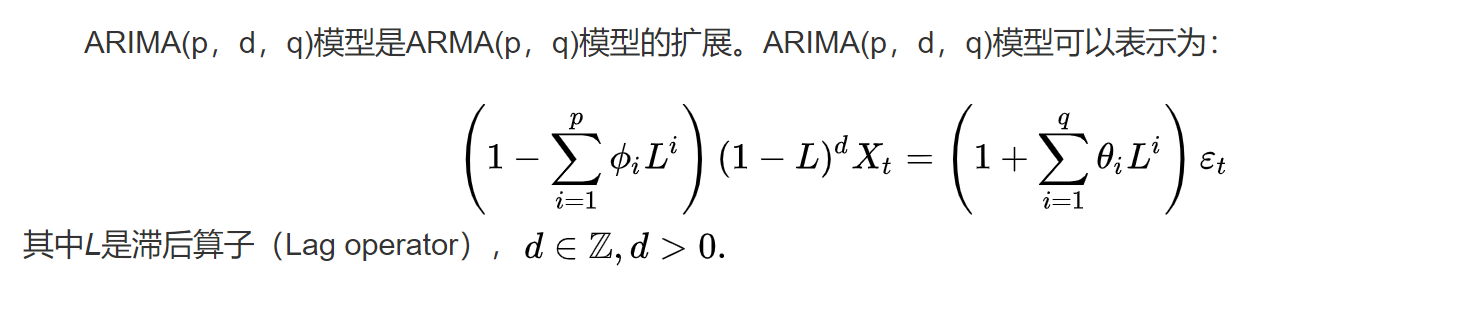
结论：

最终结论：

4-1时间序列模型ARIMA—数据预测

4-1-1、ARIMA模型理论介绍

ARIMA模型（英语：Autoregressive Integrated Moving Average model），差分整合移动平均自回归模型，又称整合移动平均自回归模型（移动也可称作滑动），是[时间序列](https://baike.baidu.com/item/%E6%97%B6%E9%97%B4%E5%BA%8F%E5%88%97)预测分析方法之一。ARIMA(p，d，q)中，AR是“自回归”，p为自回归项数；MA为“滑动平均”，q为滑动平均项数，d为使之成为平稳序列所做的差分次数（阶数）。“差分”一词虽未出现在ARIMA的英文名称中，却是关键步骤。



4-1-2 利用自相关图和偏自相关图确定参数p，q

自相关图：

偏自相关图：

拖尾：始终有非零取值，不会在k大于某个常数后就恒等于零(或在0附近随机波动)

截尾：在大于某个常数k后快速趋于0为k阶截尾当自相关图和偏自相关图都拖尾

通过3-3可知，一阶差分数据与二阶差分数据是有意义的时间序列

一阶差分数据自相关图与偏自相关图如下：

图

结论：p,q

二阶差分数据自相关图与偏自相关图如下:

图

结论：p,q

4-1-3 使用R语言auto.arima函数确定最佳参数p，d，q

由于每天只能使用截止当天的价格数据，也即每天使用的训练数据都不一致，故通过自相关图和偏自相关图为模型确定最佳参数不现实，故我们采用R语言auto.arima函数实现自动化定参。

采用auto.arima函数利用全部已知数据训练得到最佳模型信息如下：

模型信息\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

其中p=，d=，q=

4-1-4 对模型残差进行白噪声检验

合理的模型的模型残差应当为白噪声，我们对所得模型的残差进行白噪声检验，结果如下：

结果\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

结论：

4-1-5 模型预测及可视化

我们利用模型对历史数据进行预测，得到结果并可视化得到图6

图6

由于数据样本较多，图6线条重合较大，我们取其中100个样本得到图7

图7

结论：

4-1-6 批量化预测未来7天数据

根据截止当天的价格数据预测未来7天的价格数据，同样使用auto.arima函数进行自动化arima建模，得到未来7天价格数据。

通过分析预测数据，发现大致呈线性变化，我们利用线性回归拟合这些数据，得到未来的价格变化趋势，并用拟合斜率量化此趋势，用于投资决策。