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：黄金总体呈现价格上升趋势，但波动较大

比特币价格变化趋势：比特币总体呈现价格上升趋势，有两个比较大的波峰

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，结果如下：

黄金：Dickey-Fuller = -2.4368, Lag order = 10, p-value = 0.3934

比特币：Dickey-Fuller = -1.4395, Lag order = 12, p-value = 0.8156

结论：原始数据，图像均值随时间变化且单位根检验p>0.05，因此是不稳定的时间序列

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，结果如下：

黄金：Dickey-Fuller = -11.357, Lag order = 10, p-value = <0.01

比特币：Dickey-Fuller = -11.633, Lag order = 12, p-value = <0.01

结论：一阶差分数据，图像均值基本为0且单位根检验p<0.05，是稳定的时间序列

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，结果如下:

黄金：Dickey-Fuller = -20.351, Lag order = 10, p-value = <0.01

比特币：Dickey-Fuller = -18.999, Lag order = 12, p-value = <0.01

结论：二阶差分数据, 图像均值基本为0且单位根检验p<0.05，是稳定的时间序列

最终结论：我们不能直接使用原始数据进行时间序列建模，因为它是不稳定的，需要使用它的一阶差分或二阶差分数据进行时间序列建模。

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，结果如下：

黄金：X-squared = 7495.2, df = 6, p-value < 2.2e-16

X-squared = 14834, df = 12, p-value < 2.2e-16

X-squared = 22018, df = 18, p-value < 2.2e-16

比特币：X-squared = 10716, df = 6, p-value < 2.2e-16

X-squared = 20966, df = 12, p-value < 2.2e-16

X-squared = 30765, df = 18, p-value < 2.2e-16

结论：原始数据p<0.05,不是白噪声

再对一阶差分数据检验，结果如下：

黄金：X-squared = 35.268, df = 6, p-value = 3.824e-06

X-squared = 47.324, df = 12, p-value = 4.097e-06

X-squared = 56.106, df = 18, p-value = 8.576e-06

比特币：X-squared = 21.896, df = 6, p-value = 0.001265

X-squared = 63.942, df = 12, p-value = 4.275e-09

X-squared = 71.685, df = 18, p-value = 2.339e-08

结论：一阶差分数据p<0.05,不是白噪声

最后对二阶差分数据检验，结果如下：

黄金：X-squared = 361.9, df = 6, p-value < 2.2e-16

X-squared = 383.89, df = 12, p-value < 2.2e-16

X-squared = 405.63, df = 18, p-value < 2.2e-16

比特币：X-squared = 604.6, df = 6, p-value < 2.2e-16

X-squared = 654.46, df = 12, p-value < 2.2e-16

X-squared = 684.06, df = 18, p-value < 2.2e-16

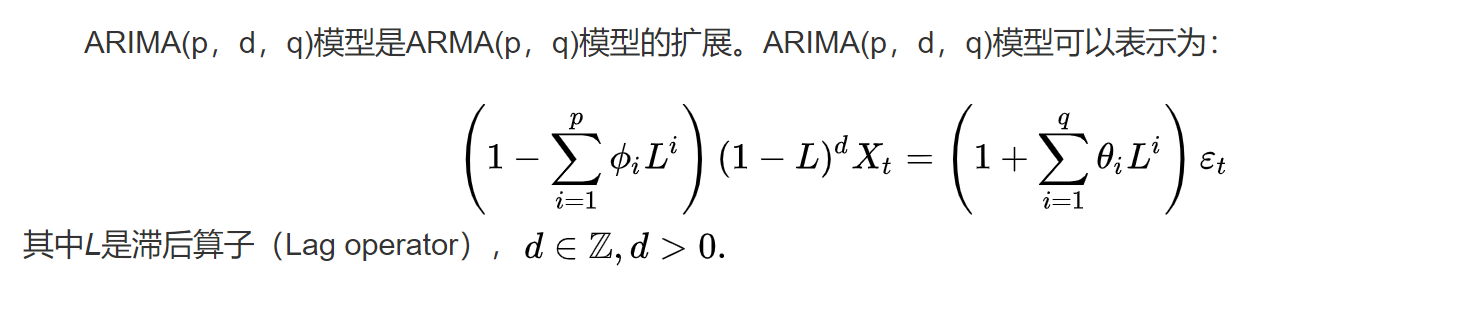
结论：二阶差分数据p<0.05,不是白噪声

最终结论：数据来自专业的数据统计中心，应当不是白噪声，检验结果也证明了这一点。它们都不是白噪声。

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：由于时间序列的相关性与之前的相同系列的值进行了计算，这被称为序列相关或自相关。一个时间序列的自相关系数被称为自相关函数，或简称ACF。这个图被称为相关图或自相关图。

偏自相关图：偏自相关是剔除干扰后时间序列观察与先前时间步长时间序列观察之间关系的总结。在滞后k处的偏自相关是在消除由于较短滞后条件导致的任何相关性的影响之后产生的相关性。一项观察的自相关和在先验时间步上的观测包括直接相关和间接相关。这些间接相关是线性函数观察（这个观察在两个时间步长之间）的相关。偏自相关函数试图移除这些间接相关。相应的图被称为偏自相关图

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

二阶差分数据自相关图与偏自相关图如下:

图

根据拖尾与截尾情况可以确定p，q参数，但是考虑到每天都要重新确定参数，通过观察图像定参是不实际的。

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：

黄金：ARIMA(4,1,5)

Coefficients:

ar1 ar2 ar3 ar4 ma1 ma2 ma3 ma4 ma5

-0.8116 0.5011 -0.1238 -0.4883 0.8421 -0.4472 0.1576 0.3524 -0.1531

s.e. 0.1132 0.2086 0.1975 0.0993 0.1128 0.2181 0.2155 0.1148 0.0300

sigma^2 = 183.8: log likelihood = -5084.4

AIC=10188.8 AICc=10188.98 BIC=10240.22

比特币：ARIMA(2,1,1)

Coefficients:

ar1 ar2 ma1

0.4319 0.1023 -0.5057

s.e. 0.2190 0.0246 0.2198

sigma^2 = 646287: log likelihood = -14796.41

AIC=29600.81 AICc=29600.84 BIC=29622.85

其中对黄金p=4，d=1，q=5

对比特币p=2，d=1，q=1

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：

黄金：X-squared = 4.9084, df = 7.1428, p-value = 0.6862

比特币：X-squared = 1.3484, df = 7.5099, p-value = 0.9919

结论：模型残差p>0.05,是白噪声，模型有效。

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：由图像可以清晰的看到，模型拟合程度很好。

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。