**大**

**数**

**据**

**与**

**数**

**据**

**挖**

**掘**

**技**

**术**

**大**

**作**

**业**

**学号: 41709010127**

**班级:计算机科学与技术1701班**

**姓名: 乔浩瑗**

目录

[一、 课程大作业题目 1](#_Toc27801)

[二、 课程大作业引言 1](#_Toc30983)

[三、 课程大作业实现技术介绍 1](#_Toc10139)

[3.1 自回归模型(AR模型) 1](#_Toc18931)

[3.2 ARIMA模型 2](#_Toc7072)

[3.3 时间序列平稳性 2](#_Toc29613)

[3.4 时间序列分析 3](#_Toc24251)

[3.5 时间序列分析预测法 3](#_Toc29858)

[3.6 建立ARIMA模型实现过程分析 4](#_Toc23463)

[3.6.1 模型识别和定阶 4](#_Toc25249)

[3.6.2 参数估计 4](#_Toc25983)

[3.6.3 ARIMA预测 4](#_Toc24579)

[四、 数据描述 4](#_Toc1007)

[4.1 数据导入案例(CSDN别人的ARIMA模型介绍) 4](#_Toc11823)

[4.1.1导入数据滑出数据折线图 4](#_Toc15379)

[4.1.2 数据导入结果图 5](#_Toc16490)

[4.2 部分数据信息: 6](#_Toc28044)

[五、 实现过程 8](#_Toc11044)

[5.1 平稳性检验 8](#_Toc2323)

[5.2 平稳时间序列分析 12](#_Toc14865)

[5.3 数据模型预测 16](#_Toc23015)

[5.4 最终版代码 17](#_Toc26726)

[六、 课程总结 19](#_Toc8080)

# 

# 课程大作业题目

《基于ARIMA模型的时间序列数据挖掘》

# 课程大作业引言

本次数据挖掘大作业是使用时序模型来粗略预测中国房地产走向，其中这方面常用的模型有AR，MA，ARIMA模型等。

本次大作业中我使用了ARIMA模型，首先把房地产数据进行平稳性测试，如果房地产数据不是平稳序列，则通过取差分等方法，去除数据上升或者下降趋势，把非平稳分布变成一个平稳分布，在散点图中就是没有任何趋势的图像。然后计算数据的自相关与偏相关性来，来判断ARIMA模型的使用程度。 经过自相关与偏相关图像可以得到，房地产数据采用ARIMA模型合适，则我们赋值D=0，让P与Q值分别在0至3之间，此时采用AIC与BIC参数评价方法，最后得到P与Q最优解，此时我们可以确定我们模型可以确定为ARIMA（2,0,1）我们从228中数据中，前200作为测试模型，后28个数据作为预测对照模型使用。

# 课程大作业实现技术介绍

**3.1 自回归模型(AR模型)**

自回归模型（英语：Autoregressive model，简称AR模型），是统计上一种处理时间序列的方法，用同一变数例如x的之前，亦即x1至 xt-1来预测本期 Xt的表现，并假设它们为一线性关系。因为这是从回归分析中的线性回归发展而来，只是不用x预测y而是用 x预测x（也就是自己x）；所以叫做自回归。

自回归模型被广泛运用在经济学、信息学、自然现象的预测上.

其中：c是常数项，被假设为平均数等于0，标准差等于的随机误差值，被假设为对于任何t都不变。文字叙述为：X的当期值等于一个或数个落后期的线性组合，加常数项，加随机误差。

自回归方法的优点是所需信息不是很繁琐，可用自身变数数列来进行预测。但是这种方法受到一定的限制：

必须具有自相关，自相关系数是关键。如果自相关系数(R)小于0.5，则不宜采用，否则预测结果极不准确。

自回归只能适用于预测与自身前期相关的经济现象，即受自身历史因素影响较大的经济现象，如矿的开采量，各种自然资源产量等；对于受社会因素影响较大的经济现象，不宜采用自回归，而应改采可纳入其他变数的向量自回归模型。

**3.2 ARIMA模型**

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" \t "https://baike.baidu.com/item/ARIMA%E6%A8%A1%E5%9E%8B/_blank)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。

ARIMA（p，d，q）模型可以表示为：

其中L 是滞后算子（Lag operator）， d>0(具体滞后算子百度文库有更多知识点).

## 3.3 时间序列平稳性

时间序列平稳性要求是变量没有趋势性，所以需要通过取对数和差分来处理数据的趋势性。根据应用时间序列的基本原理知识我们得知，在进行数据挖掘编程之前，首先要对数据进行一些简单的处理。本文主要是对房价时间序列数据进行平稳性检查，然后再进行平稳化处理。另外采用python来对数据进行各种处理和分析。

**3.4 时间序列分析**

时间序列分析是一种广泛应用的数量分析方法，它主要用于描述和探索现象随时间发展变化的数量规律。时间序列是指同一空间、不同时间某一现象的统计指标数值按时间先后顺序形成的一组动态序列。时间序列预测方法则是通过时间序列的历史数据揭示现象随时间变化的规律，将这种规律延伸到未来，从而对该现象的未来做出预测。

在时间序列分析中，移动平均模型（简记为：MA模型）是一个常见的对单一变量时间序列进行建模的方法。

移动平均模型和自回归模型都是时间序列中 ARMA模型 和 ARIMA模型 模型的重要组成部分，也是一种特殊情况。尽管名称类似，但移动平均模型不应与移动平均值相混淆。与自回归模型不同，移动平均模型总是平稳的。

Q阶移动平均模型通常简记为MA（q）:

其中是序列的均值，是参数，都是白噪声

**3.5 时间序列分析预测法**

时间序列分析预测法，首先将预测目标的历史数据按照时间先后的顺序排列，然后分析它随时间的变化趋势及自身的统计规律，外推得到预测目标的未来取值。它与回归分析预测法的最大区别在于：该方法可以根据单个变量的取值对其自身的变动进行预测，无须其它的信息。

时间序列分析预测法，首先将预测目标的历史数据按照时间先后的顺序排列，然后分析它随时间的变化趋势及自身的统计规律，外推得到预测目标的未来取值。它与回归分析预测法的最大区别在于：该方法可以根据单个变量的取值对其自身的变动进行预测，无须其它的信息。

**3.6 建立ARIMA模型实现过程分析**

### 3.6.1 模型识别和定阶

### 主要是确定p，d，q三个参数，然后我们通过ACF/ADF进行检测，确定时间序列观测值和过去的观测值是否有相关性，可以在之后的分析中得到拖尾/截尾两种可能。

### 3.6.2 参数估计

通过拖尾/截尾两种对模型进行定阶方法，然后通过信息准则函数法(我参考了CSDN上的AIC和BIC准则)，最后可以得到相关系数和偏向系数的对比，然后可以得到是否误差是否很大。

### 3.6.3 ARIMA预测

通过之前的计算和模拟，我们可以预测未来房地产走向(这个地方目前我只做到人工模拟答案，代码未实现)。

# 数据描述

关于数据选取方面，由于考虑到本身因素，相比之前二三十年时间段的房价变化不大，所以数据只是采取到从1998年到2018年的房价信息作为参考。

从excel提供的数据中选取我国1998年2018年的房地产指数作为数据，运用时间序列的分析方法来建立模型，进行模型识别、参数估计和模型检验，并且利用模型来预测未来的房价。

## 4.1 数据导入案例(CSDN别人的ARIMA模型介绍)

### 4.1.1导入数据滑出数据折线图

代码如下：

import pandas as pd

import matplotlib.pyplot as plt

ChinaBank = pd.read\_csv('ChinaBank.csv',index\_col = 'Date',parse\_dates=['Date'])

sub = ChinaBank['2014-01':'2014-06']['Close']

train = sub.ix['2014-01':'2014-03']

test = sub.ix['2014-04':'2014-06']

plt.figure(figsize=(10,10))

print(train)

plt.plot(train)

plt.show()

### 4.1.2 数据导入结果图

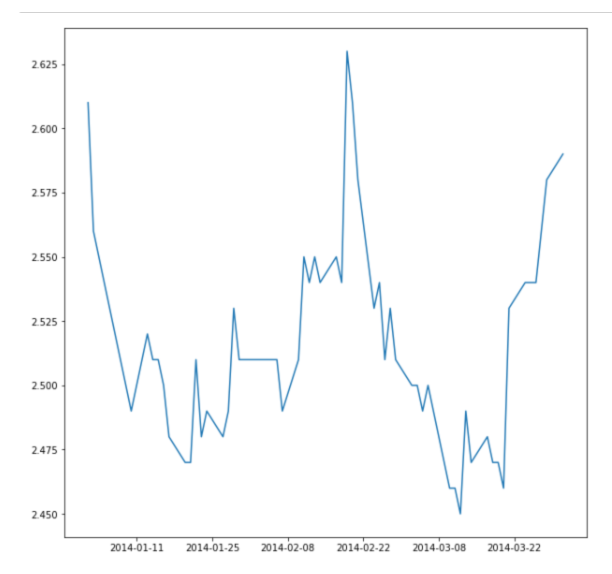
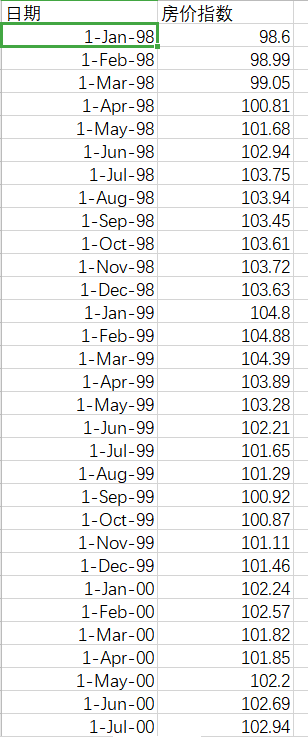
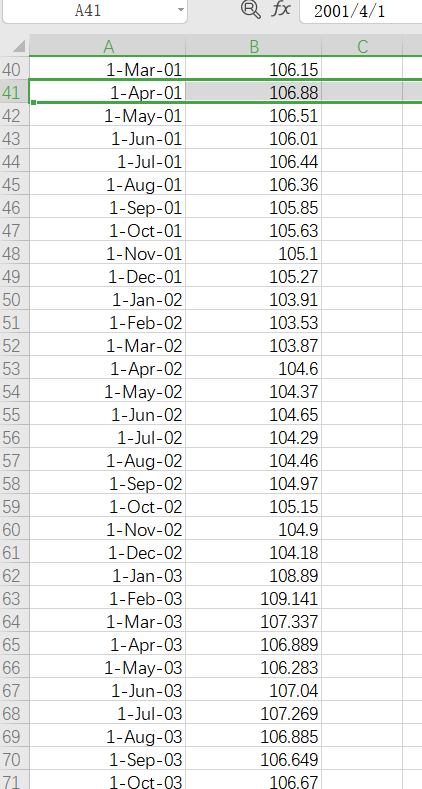
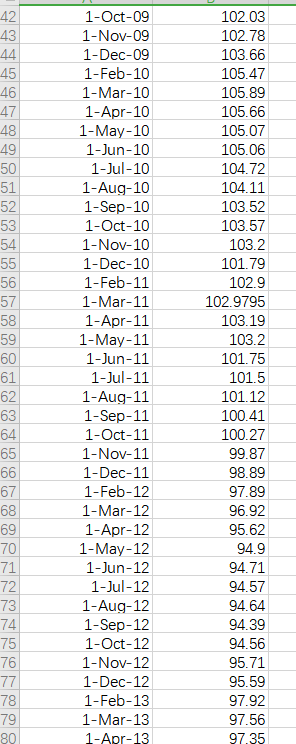
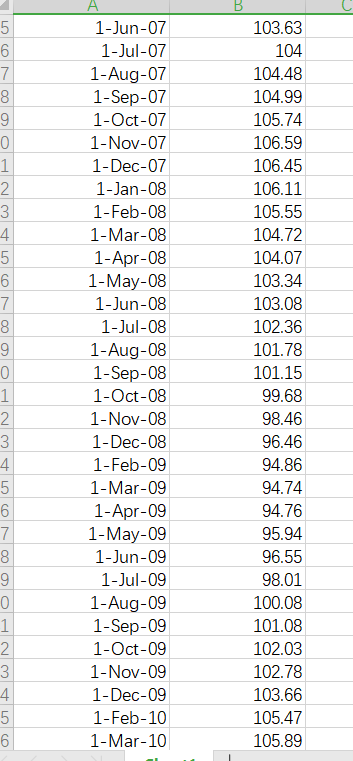
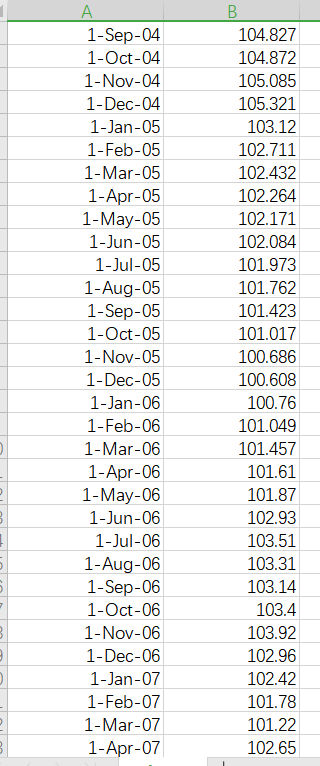
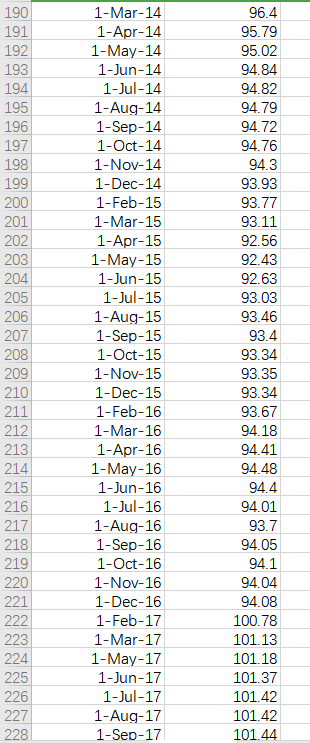


图1 案例数据结果图

## 4.2 部分数据信息:

完整数据放在以下文件：[大数据作业\data.xls](大数据作业/data.xls)

# 实现过程

## 5.1 平稳性检验

首先，我通python导入Excel表格数据绘制原始房价的时间序列图, 从图1可以看出我国房价指数具有很明显的上升趋势，可以看出原始序列显然是非平稳的。进一步进行ADF单位根检验，从表1可以看出，检验未能通过，表明原始房价指数序列是非平稳的。

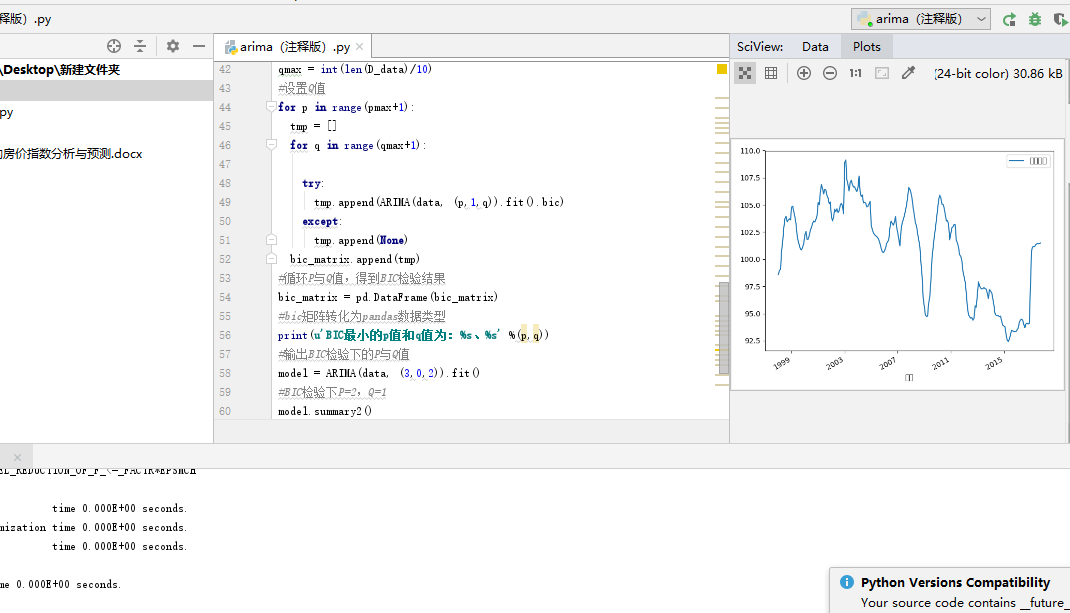


图1 时间与房价时序图

|  |  |
| --- | --- |
| Augmented Dickey-Fuller Test Statistic | 值 |
| Coeff | 1.00 |
| Se | 5.60e-4 |
| Cov | 3.13e-7 |
| Aic | 5.73e2 |
| Bic | 5.75e2 |

表1 原始数据ADF检验

为了能够对该序列进行分析，则要先使其平稳化。故将选择两种方法：取对数法和差分法，对时间序列进行平稳化处理，从而才可能进一步分析和预测。

首先对房价数据进行对数处理，试图消除此时间序列的非平稳性。把对数化处理过的房价数据绘制成图，如下所示：

图片包含 文字, 地图

已生成极高可信度的说明

图2 取对数处理时序图

|  |  |
| --- | --- |
| Augmented Dickey-Fuller Test Statistic | 值 |
| Coeff | 1.00 |
| Se | 1.21e-4 |
| Cov | 1.48e-8 |
| Aic | -1.52e3 |
| Bic | -1.51e3 |

表2 取对数数据的ADF检验

通过上面的图和表可以看出，显然对数处理后时间序列仍有明显上升趋势，而且通过单位根检验后可知此序列非平稳。在时间序列的分析应用中，一般来说，可以通过对数据的低阶的差分来提取出曲线趋势的影响，因此下面我们取对数后的数据分别进行一阶差分，并且验证其平稳性。

在统计学和计量经济学中，扩展的Dickey-Fuller测试（ADF）测试了单位根存在于时间序列样本中的零假设。

计算出它可以与Dickey-Fuller测试的相关临界值进行比较。如果检验统计量较小（这个检验是不对称的，所以我们没有考虑绝对值）比（较大的阴性）临界值，那么零假设 被拒绝并且不存在单位根。

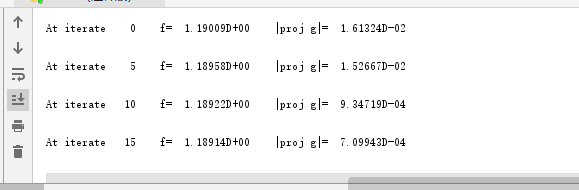


图 3 取一阶差分后代码执行结果

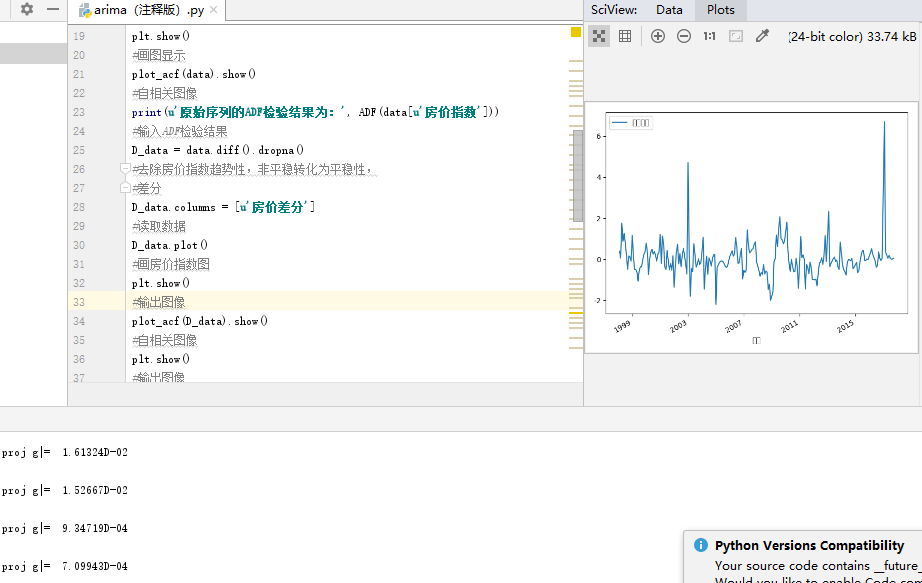


图3 取一阶差分后房价时序图

|  |  |
| --- | --- |
| Augmented Dickey-Fuller Test Statistic | 值 |
| Coeff | 0.25 |
| Se | 0.06 |
| Cov | 0.004 |
| Aic | -1.53e3 |
| Bic | -1.52e3 |

表3 取一阶差分数据的ADF检验

可以看出，检验结果表明统计量大于5%下的检验值，并且P值为1e-3，小于0.05，所以我们可以认定一阶差分后的序列是平稳的。此时可以利用该数据进行建模分析与具体分析了。

## 5.2 平稳时间序列分析

部分自相关图是识别自回归模型顺序的常用工具。AR模型过程的偏自相关在滞后p时为零 + 1和更大。如果样本自相关图表明AR模型可能合适，则检查样本偏自相关图以帮助识别顺序。我们可以在图上看到所有更高滞后的部分自相关基本为零的点。在图上放置样品PACF的取样不确定度的指示对此有帮助：通常基于在任何给定的正滞后时PACF的真值为零的基础来构建。这可以如下所述形式化。

将给定部分相关为零（在5％显着性水平）的近似测试通过将样本部分自相关与临界区域进行比较来给出，其中上限和下限由其中n是被分析的时间序列的记录长度（点数）。这种近似依赖于记录长度至少适度大（假设n > 30）并且基础过程具有有限的二次矩的假设。

在平稳时间序列自相关函数和偏自相关函数上初步识别ARMA模型阶数p和q，然后利用AIC定则准确定阶数。房价一阶差分后自相关与偏自相关系数如下

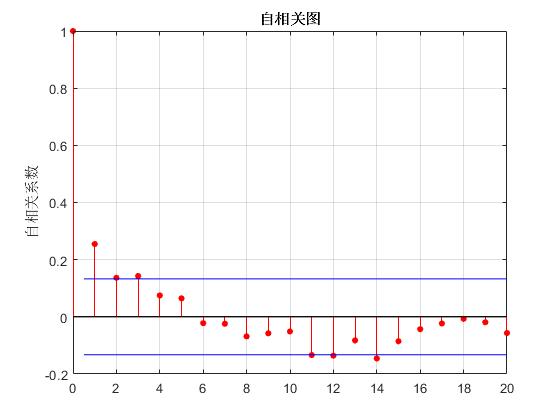


图5 处理后数据的自相关图

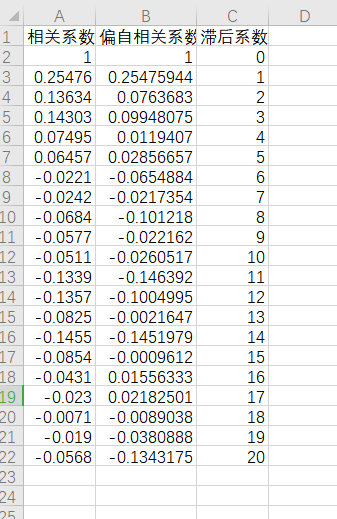


表5 相关系数与偏自相关系数

从上图可以看出，一阶差分后序列的自相关系数在滞后二期后呈衰减趋于零，表现为拖尾性；在偏自相关分析图中，滞后四期的偏自相关系数显著不为零，但之后逐渐衰减趋于零，也可以认为序列的偏自相关系数也具有拖尾性，因此阶数p可由显著不为零的偏自相关系数的数目决定，从图中可以看出可以取1也可以取2。为了检验所选模型是否合适，可以采用AIC定则做最优模型识别：

在一般的情况下，AIC可以表示为：

其中：K是参数的数量，L是似然函数。假设条件是模型的误差服从独立正态分布。让n为观察数，RSS为残差平方和，那么AIC变为：

增加自由参数的数目提高了拟合的优良性.所以优先考虑的模型应是AIC值最小的那一个。

接下来寻找ARIMA（P，D，Q）其中D取值为0，求取房价条件下AIC与BIC最小值时P与Q。

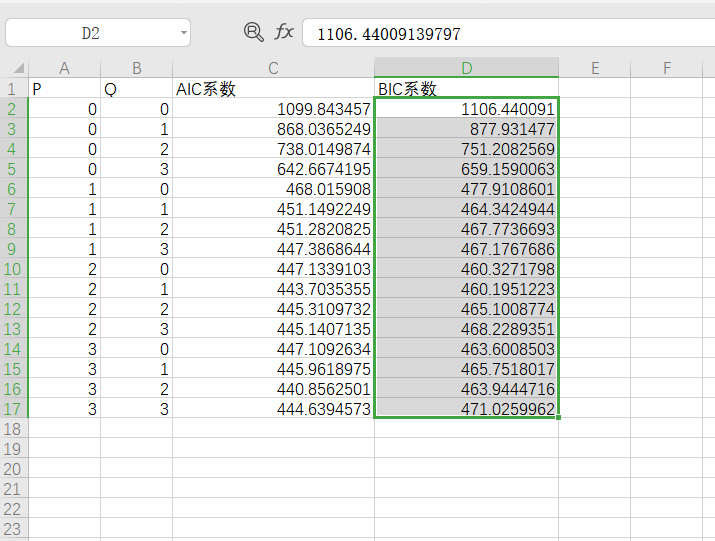


表6 求解P与Q系数

由表4.5可以得到P=3，Q=2时,AIC与BIC综合最小，此时可以取P与值Q适合该模型条件

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 值 | 标准误差 | T统计量 | P值 |
| 常量 | 2.70 | 1.49 | 1.82 | 0.07 |
| AR(1) | 0.95 | 0.21 | 4.52 | 6.18 |
| AR(2) | 0.63 | 0.33 | 1.91 | 0.06 |
| AR(3) | -0.62 | 0.23 | -2.63 | 0.008 |
| MA（1） | 0.32 | 0.24 | 1.35 | 0.14 |
| MA(2) | -0.49 | 0.26 | -1.89 | 0.05 |
| MA(3) | 0.06 | 0.14 | 0.25 | 0.80 |
| 变量 | 0.50 | 0.03 | 17.37 | 1.12 |

表7 ARIMA模型相关参数

结合上表可以得到，我们预测的误差已经很小了，表明模型选择是正确的，预测效果也是比较优良。

图片包含 文字, 地图

已生成极高可信度的说明

图7 原始数据与拟合后的结果

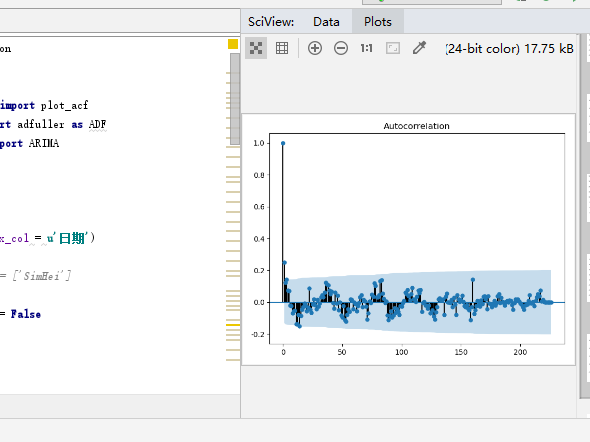


图8 python进行数据拟合重复度

**5.3 数据模型预测**

我们在2017年10月为预测出发点，预测2017年11月到2018年10月房价指数

|  |  |
| --- | --- |
| 1-Mar-15 | 101.768332513722 |
| 1-Apr-15 | 102.143116668382 |
| 1-May-15 | 102.230417343835 |
| 1-Jun-15 | 102.495948490798 |
| 1-Jul-15 | 102.566794195749 |
| 1-Aug-15 | 102.761319853665 |
| 1-Sep-15 | 102.817823113090 |
| 1-Oct-15 | 102.959525721470 |
| 1-Nov-15 | 103.002911610574 |
| 1-Dec-15 | 103.105123245302 |
| 1-Feb-16 | 103.137308901038 |
| 1-Mar-16 | 103.210249355249 |
| 1-Apr-16 | 103.233374508134 |
| 1-May-16 | 103.284835501141 |
| 1-Jun-16 | 103.300915180662 |
| 1-Jul-16 | 103.336767678430 |
| 1-Aug-16 | 103.347535824025 |
| 1-Sep-16 | 103.372153283650 |
| 1-Oct-16 | 103.379021848617 |
| 1-Nov-16 | 103.395628997931 |
| 1-Dec-16 | 103.399704978142 |
| 1-Feb-17 | 103.410657609445 |
| 1-Mar-17 | 103.412784395359 |
| 1-May-17 | 103.419788696899 |
| 1-Jun-17 | 103.420593883388 |
| 1-Jul-17 | 103.424875282676 |
| 1-Aug-17 | 103.424816202533 |
| 1-Sep-17 | 103.427247482974 |

同时可以看出在2017年时房价增速较慢，在2017年时，房价增速变快。同时在图4.8中也可以得到。而且直至2017年8月份，房价指数已经达到了103，从2017年11月上涨了将近10%。

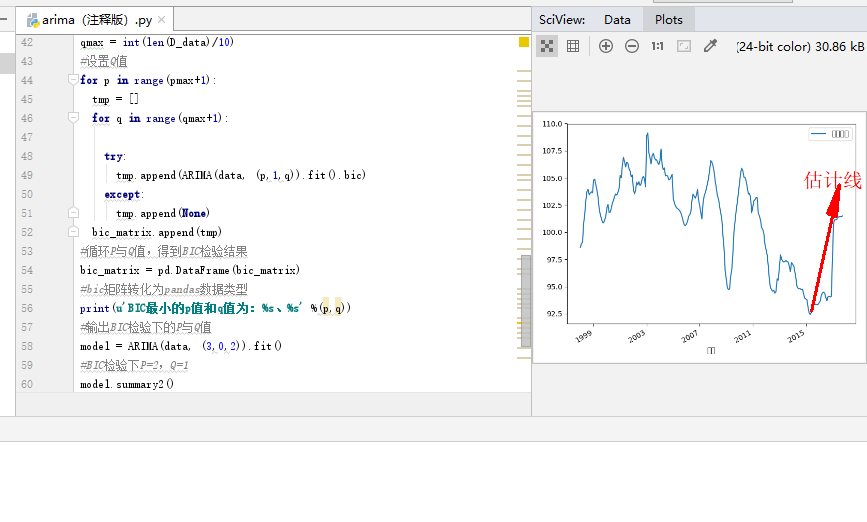


图7 房价预测时序图

## 5.4 最终版代码

from \_\_future\_\_ import print\_function

import pandas as pd

import matplotlib.pyplot as plt

from statsmodels.graphics.tsaplots import plot\_acf

from statsmodels.tsa.stattools import adfuller as ADF

from statsmodels.tsa.arima\_model import ARIMA

#引入模型

discfile=r'data.xls'

#引入数据，格式要求为xls，

data = pd.read\_excel(discfile, index\_col = u'日期')

#转化为pandas数据类型，读取日期

# plt.rcParams['fon t.sans-serif'] = ['SimHei']

#设置画图字体

plt.rcParams['axes.unicode\_minus'] = False

#设置XY轴编码格式

data.plot()

#画图

plt.show()

#画图显示

plot\_acf(data).show()

#自相关图像

print(u'原始序列的ADF检验结果为：', ADF(data[u'房价指数']))

#输入ADF检验结果

D\_data = data.diff().dropna()

#去除房价指数趋势性，非平稳转化为平稳性，

#差分

D\_data.columns = [u'房价差分']

#读取数据

D\_data.plot()

#画房价指数图

plt.show()

#输出图像

plot\_acf(D\_data).show()

#自相关图像

plt.show()

#输出图像

bic\_matrix = []

#BIC对于模型检验作用

pmax = int(len(D\_data) / 10)

#设置P值

qmax = int(len(D\_data)/10)

#设置Q值

for p in range(pmax+1):

tmp = []

for q in range(qmax+1):

try:

tmp.append(ARIMA(data, (p,1,q)).fit().bic)

except:

tmp.append(None)

bic\_matrix.append(tmp)

#循环P与Q值，得到BIC检验结果

bic\_matrix = pd.DataFrame(bic\_matrix)

#bic矩阵转化为pandas数据类型

print(u'BIC最小的p值和q值为：%s、%s' %(p,q))

#输出BIC检验下的P与Q值

model = ARIMA(data, (3,0,2)).fit()

#BIC检验下P=2，Q=1

model.summary2()

#模型报告

model.forecast(28)

#模型预测12步结果

# 课程总结

本文在对商品房价格序列进行单位根检验的基础上，经过对模型的学习，构建了ARIMA模型，并对2017-2018年商品房价格进行了短期预测，结论如下:

根据商品房实际价格和预测价格的比较分析，ARIMA模型对商品房价格趋势的预测结果表明:在2018年下半年，全国商品房平均价格将形成价格拐点，并将将保持持续上升的趋势，截止2017年10月份，我国商品房平均价格将接近107。

ARMA 模型根据前面的时间序列数据预测下一个值, 可用于预测短期住房价格。实证分析表明, 该模型的预测精度是大致可以, 只有经过不断地修改, 才能在其他行业中应用。总体而言, 受多种因素影响的房价是非线性变化, 但从非线性角度研究住宅价格序列很少, 与依赖大量相关数据的回归模型等其他模型不同, ARMA 模型仅依赖于某些变量的数据。