对上海私车牌照拍卖价格的时间序列分析

应用统计学 16 2016310868 郑昊亮

目录

1 引入

1986 年 11 月，上海第一辆 “Z” 字私人自备车牌照代码 “沪-AZ0001” 号诞生。这被认为是中国私家车开行的标志，也被视为改革开放的一个里程碑。此门一开，小汽车进入一般家庭便是历史大势，城市交通也面临着冲击与挑战。1994 年，一个对机动车进行总量控制的制度应运而生：上海政府发布规定，以有底价、不公开拍卖的方式，对私车牌照额度进行市场化分配。这就是上海私车牌照拍卖制度的起源。

本文主要分析的对象即是上海私车牌照历史拍卖价格的时间序列数据，通过进行建模，试图获得价格变动的内在规律，并在此基础上预测未来的价格。通过本文的分析，首先可以更好地反映上海私车牌照拍卖价格二十多年来的变化情况，通过对价格的预测，一方面能够反映出未来上海对汽车牌照的需求变化，另一方面也能为未来的拍卖者提供一个价格依据。

2 数据

2.1 数据集介绍

数据集来自 “上海市非营业性课程额度拍卖” 官网中的 “历史投标结果” 栏目，网址为 http://chepai.alltobid.com/contents/22/276.html 。该网页上的数据从 2002 年 1 月开始到 2018 年 12 月，每月公布一次，共 203 条（其中 2008 年 2 月的数据为缺失值）；每条数据包括：时间、投放数量、最低成交价（元）、平均成交价（元）、最低成交价截止时间、投标人数。

由于最低成交价截止时间对分析没有帮助，因此没有计入数据集中。在 R 中储存的数据框如下所示，

```
head(data)
```

```
##    Time licenses lowest.price avg.price applicants
## 1 2-Jan     1400         13600     14735       3718
## 2 2-Feb     1800         13100     14057       4590
## 3 2-Mar     2000         14300     14662       5190
## 4 2-Apr     2300         16000     16334       4806
## 5 2-May     2350         17800     18357       4665
## 6 2-Jun     2800         19600     20178       4502
```

其中 Time 为时间，licenses 为当月投放的牌照数量，lowest.price 为最低成交价，avg.price 为平均成交价，applicants 为当月投标人数。虽然主要研究的对象为平均成交价，但由于投放的拍照数量、投标人数等变量可能可以帮助解释价格的变化，也就是进行时间序列回归。

```
avg.price = ts(data[,4], start = c(2002,1), frequency = 12)
avg.price[74] = (avg.price[73] + avg.price[75])/2
```

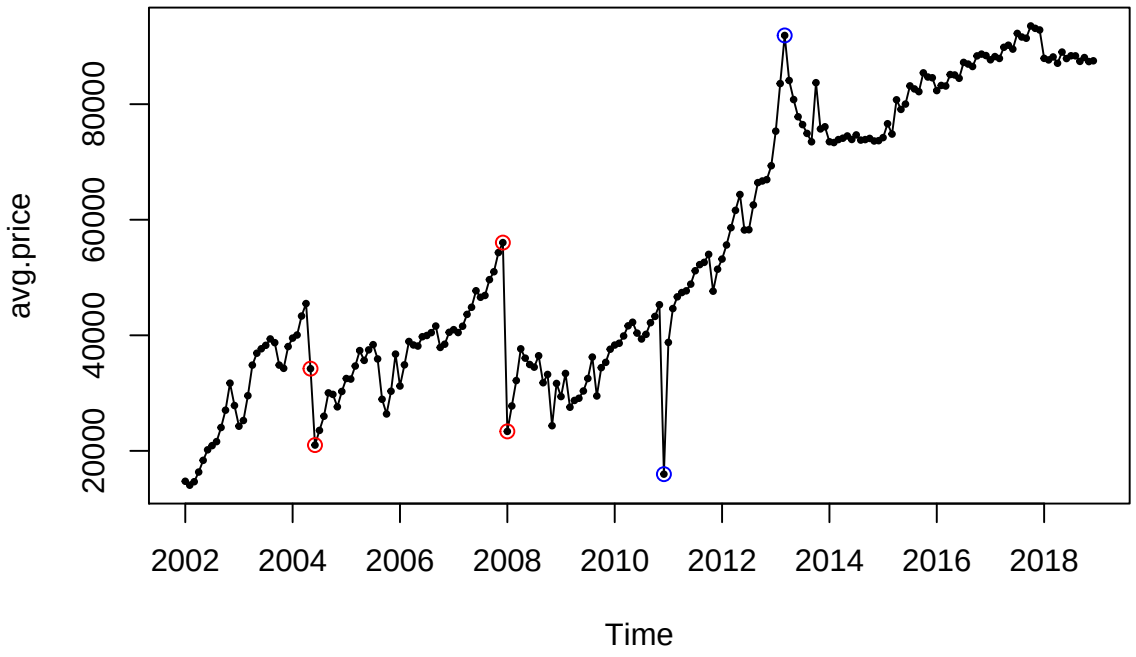
对于**缺失值**，由于仅一个且其相邻位置没有出现明显的变化规律，因此用一般的插补方法进行插补即可（例如取相邻位置的加和平均）。

2.2 数据展示与分析

作拍卖价格的时间序列图如下：

```
plot(avg.price, xaxt="n", type = "o", pch =20, cex = 0.6)
axis(1,at=seq(2002,2018,2),las=1)

timepoint1 = c(29,30,72,73); timepoint2 = c(108,135)
timepoints1 = time(avg.price)[timepoint1]; timepoints2 = time(avg.price)[timepoint2]
points(timepoints1, avg.price[timepoint1], col="red")
points(timepoints2, avg.price[timepoint2], col="blue")
```



可以看到，数据有明显的上升趋势，同时在**红点**处出现了剧烈的变化，可能要引入干预分析；在**蓝点**处出现了突出点，可能要引入异常值。红点对应的具体时间点为：**2004 年 4 月与 5 月**以及 **2007 年 12 月与 2008 年 1 月**；蓝点对应的时间点为 **2010 年 12 月**和 **2013 年 3 月**。

对历史背景进行分析，发现：

1、2004 年 5 月 1 日，《道路交通安全法》实施，规定申请机动车登记应提交五种证明，并无拍卖牌照的说法。**2004 年 5 月 24 日**，时任商务部部长助理的黄海在央视表示，上海的私车牌照拍卖违反了《道路交通安全法》。这是上海该项措施诞生以来，受到的最为明确的批评，一时令其合法性之争，在社会各界公开化、沸腾化。当年 7 月 7 日举行的上海市政府新闻发布会上，时任上海市政府法制办主任的徐强表示，经请示全国人大、国务院等，各方都认为上海拍卖私牌的做法没有违法。上海作为国际大都市，在不同发展阶段，对交通采取一定的特殊管理措施完全正当。

这就解释了为何 2004 年 4 月开始车牌价格剧烈下降，并在几个月后回复到原来的水平。考虑到最后上海市政府对合法性进行了阐明，因此对价格的影响应该是短期的，可以考虑在此处引入**脉冲响应干预**。

2、**2008 年 1 月 3 日**，上海私车牌照额度拍卖新办法出炉。从上海国际拍卖公司公布的竞拍新规来看，投标拍卖将分为两个阶段进行，使得竞买人在投标过程中能够修改竞标价格；同时，拍卖方还

将通过网上、电话或现场即时公布所有当前投标信息，包括当前投标人数和当前时间的最低中标价格等关键信息。这一力求从两个方面来抑制车牌额度继续上涨的动因。一是提供多次出价机会，二是信息全部公开透明。

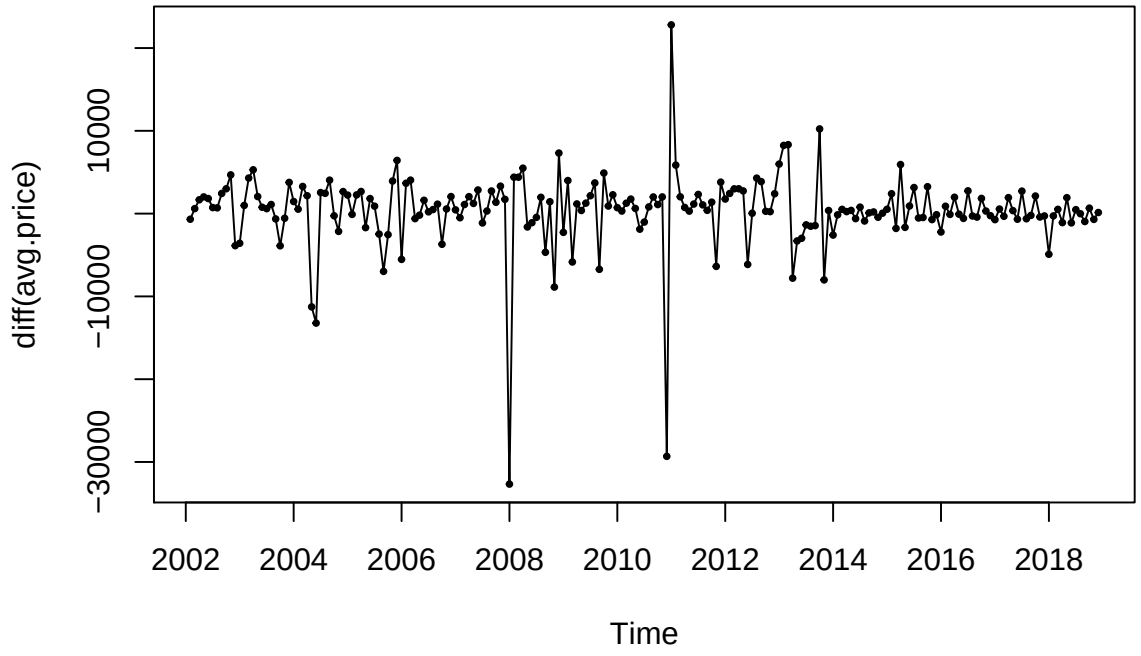
这就解释了为何车牌价格从 2007 年 12 月的高点在 2008 年 1 月突然跌到低谷。考虑到这样的新规在一段时间内都影响着拍卖的规则，拍卖者也需要不断适应这样的规则，因此考虑在此引入**阶梯响应干预**。

3、至于 2010 年 12 月价格的急剧下降，网友均表示这是“大爆冷门”，没有明显的原因。可以看到在 2011 年 1 月，价格又恢复到原来水平，因此考虑在此引入**可加异常值**。2013 年 3 月的价格高点也同理。

3 模型的建立

由于数据具有明显的上升趋势，先做一阶差分，差分后结果如下所示

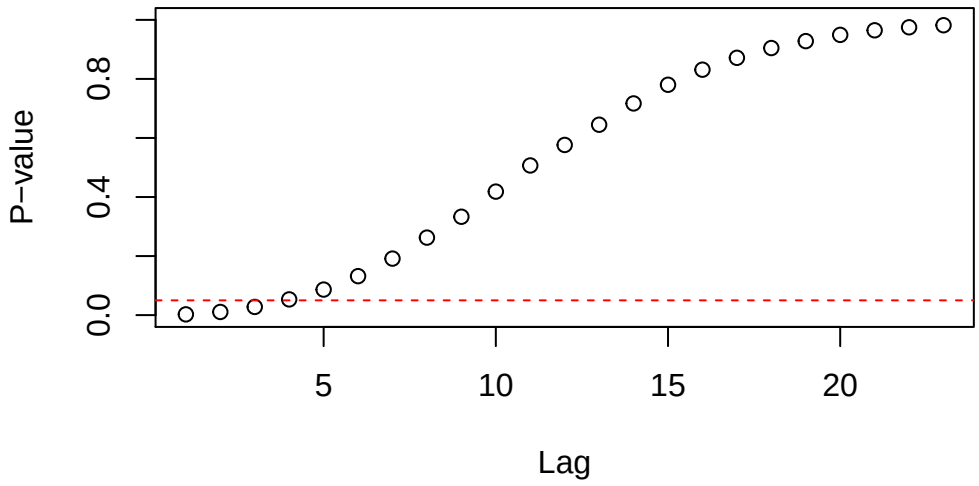
```
plot(diff(avg.price), xaxt="n", type = "o", pch =20, cex = 0.6)
axis(1,at=seq(2002,2018,2),las=1)
```



可以看到，经过差分之后，除了上文提到的几个特殊的时间点之外，在其余时间点上时间序列还是较为平稳的。

该数据没有表现出明显的 ARCH 效应，为了确认，进行 McLeod-Li 检验：

```
McLeod.Li.test(y = diff(avg.price))
```



可以看到在 4 阶之后结果都不显著，与观察基本一致，因此无需引入 ARCH 效应。（之所以在前 3 阶出现显著，可能与干预和异常值导致的表面上的“波动集群”有关）

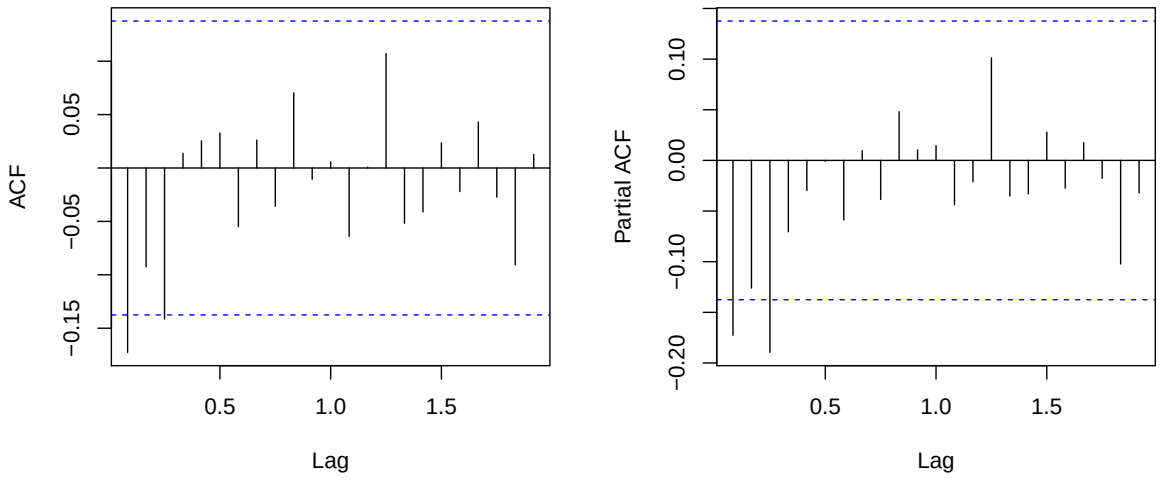
先不引入干预分析与可加异常值，按照一般的 ARIMA 模型建立过程，得到以下模型一。

3.1 模型一（ARIMA(1,1,1) ）

3.1.1 模型的识别

为了对模型一进行识别，绘制差分后的数据的 acf，pacf 和 eacf 图如下：

```
op = par(mfrow=c(1,2))
acf(diff(avg.price), main="")
pacf(diff(avg.price), main="")
```



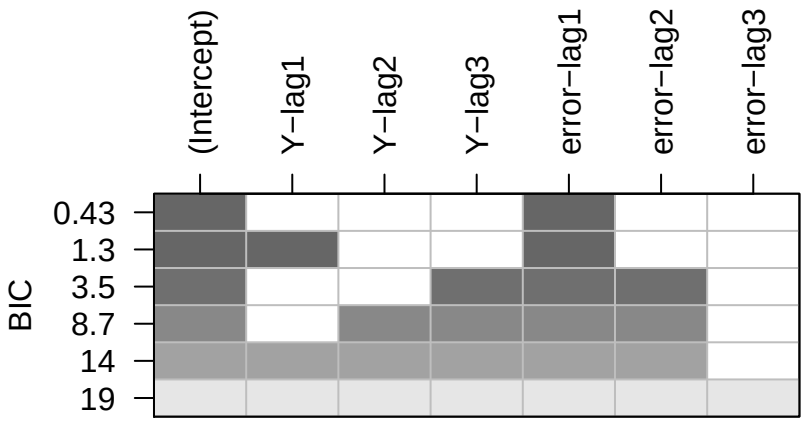
```
eacf(diff(avg.price), ar.max = 4, ma.max = 9)
```

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9
## 0 x o o o o o o o o o
## 1 x o x o o o o o o o
## 2 x x x o o o o o o o
## 3 x o o o o o o o o o
## 4 x o o x o o o o o o
```

从 acf 和 pacf 中可以看出，截尾的阶数主要在 1 阶附近，而 eacf 中没有明显的三角形状，这可能与没有去除掉外界冲击的效应有关。

如下图所示，从最优子集的角度讲，只引入 MA1 或引入 MA1 和 AR1 是较为合适的。

```
plot(armasubsets(y= diff(avg.price),nar=3,nma=3,ar.method="ols"))
```



3.1.2 模型的拟合与检验

先选择 ARIMA（1，1，1）进行拟合，模型结果如下：

```
model1 = arima(avg.price,order=c(1,1,1)); model1
```

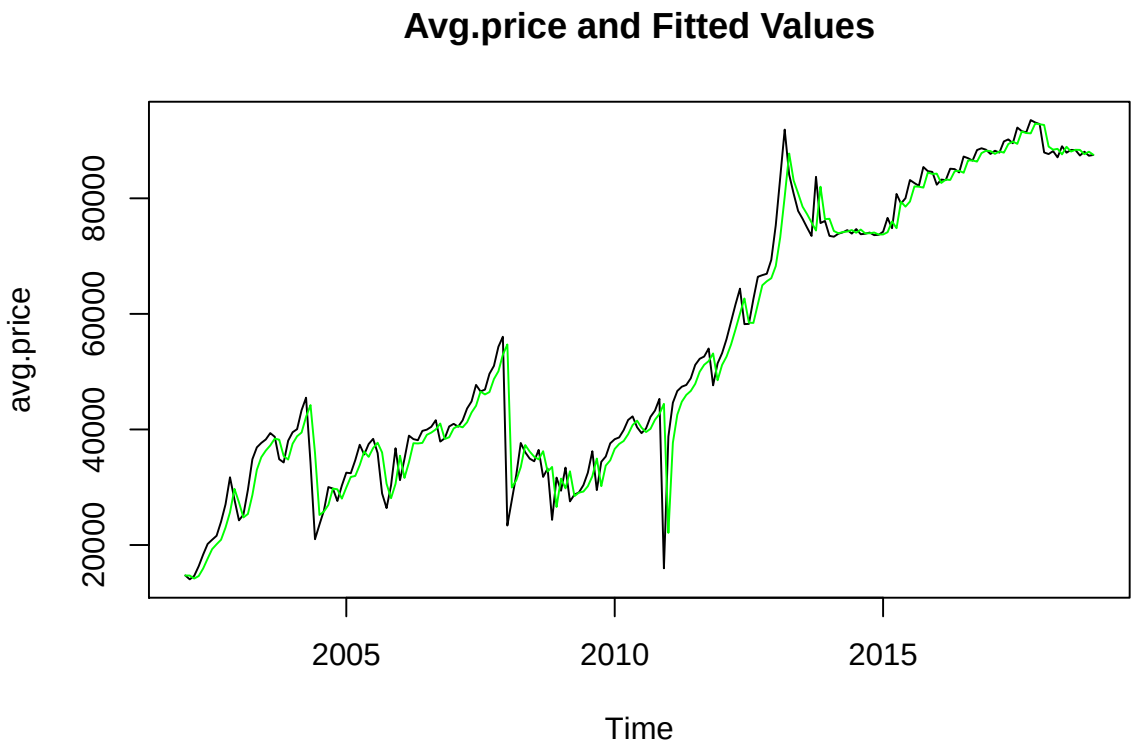
```
##
## Call:
## arima(x = avg.price, order = c(1, 1, 1))
##
## Coefficients:
##          ar1      ma1
##       0.4475  -0.6762
## s.e.  0.1507   0.1198
##
```

```
## sigma^2 estimated as 20485800:  log likelihood = -1996.88,  aic = 3997.76
```

结果显示 ar 系数和 ma 系数均显著异于 0，AIC 值为 3997.76。

原数据与拟合值的时间序列图如下：

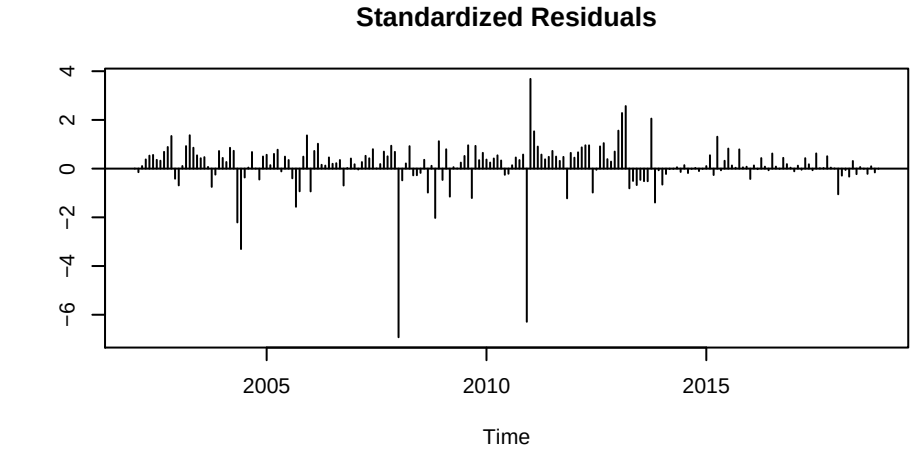
```
fit = avg.price - model1$residuals
plot(avg.price, main = "Avg.price and Fitted Values")
lines(fit,col="green")
```

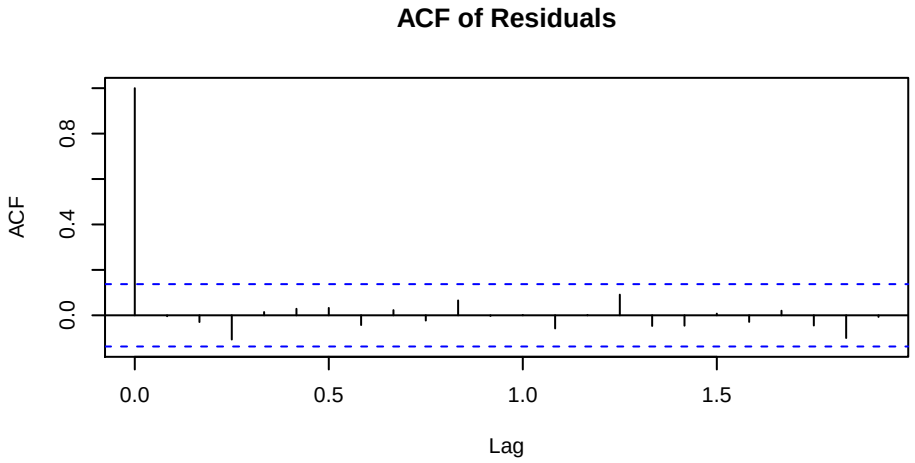


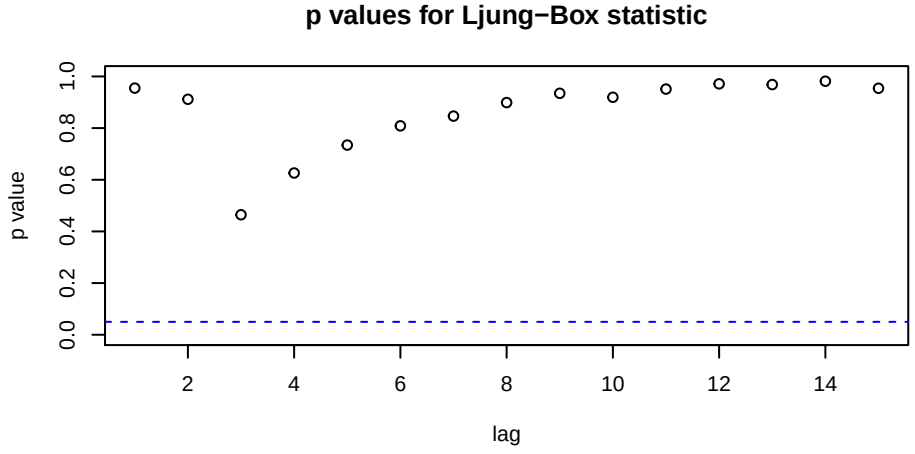
黑线代表原数据，绿线代表拟合值。可以看出绿线与黑线实则有一定的“缝隙”，并且在干预处以及异常值的位置，拟合效果并不好。

运用标准化残差的时间序列图、残差的 ACF 图和 Ljung-Box 检验图对模型进行检验：

```
tsdiag(model1,gof=15)
```







Ljung-Box 显示在 15 阶以内的残差没有显著的自相关性；但标准化残差中出现几处明显突起，数值大于 3，这应该与没有引入干预效应和异常值有关。

从以上简单的分析中可以看到，对于这样的数据，不引入干预分析，仅用 ARIMA（1，1，1）进行建模，可以取得一定的效果，但还有很多瑕疵。所以接下来的重点是引入干预分析，得到更完善的模型二，再与此处简单的模型一做对比，看干预分析的优势所在。

3.2 模型二（引入干预分析与异常值）

本模型具有两次干预，设拍卖价格的时间序列为 $\{Y_t\}$，则一般模型由下式给出：

$$Y_t = m_t + m_t' + N_t$$

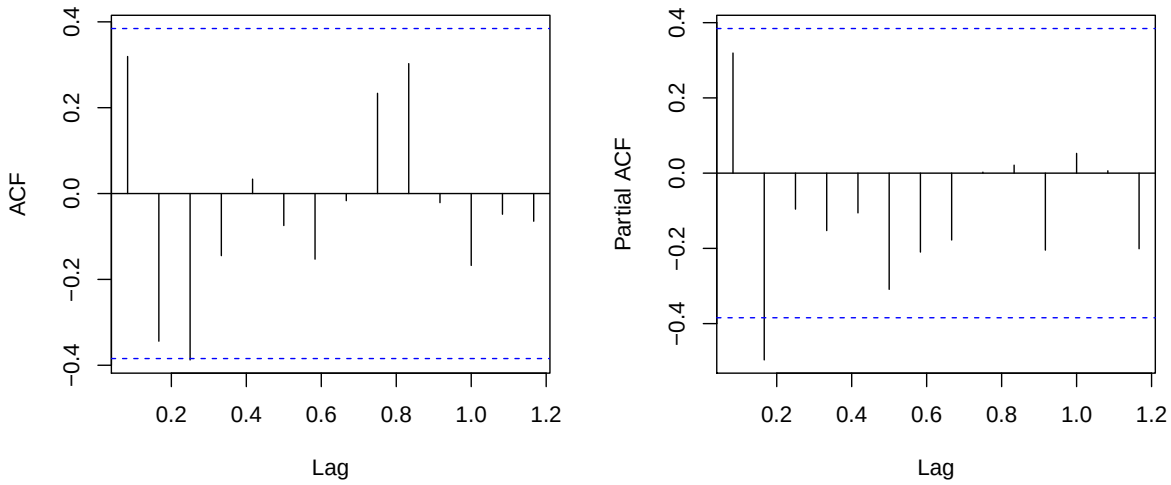
其中，m_t 代表第一次干预（2004 年 5 月 24 日）导致的均值函数的变化，m_t' 代表第二次干预（2008 年 1 月 3 日）导致的均值函数的变化，$\{N_t\}$ 的模型是 ARIMA 过程。

过程 $\{N_t\}$ 代表着未受干预影响的基础时间序列，称作无扰过程。假设时间序列在时刻 T 第一次受到干预，称时间序列 $\{Y_t, t < T\}$ 为预干预数据，可用其识别无扰过程 $\{N_t\}$ 的模型。接下来的第一步就是识别无扰过程的模型。

3.2.1 预干预期间数据建模

认为 T 为 2004 年 4 月，则在此之前为预干预期间，为了对无扰过程的模型进行识别，绘制这一期间差分后的数据的 acf，pacf 和 eacf 图如下：

```
op = par(mfrow=c(1,2))
acf(diff( window(avg.price,end=c(2004,3)) ), main="")
pacf(diff( window(avg.price,end=c(2004,3)) ), main="")
```



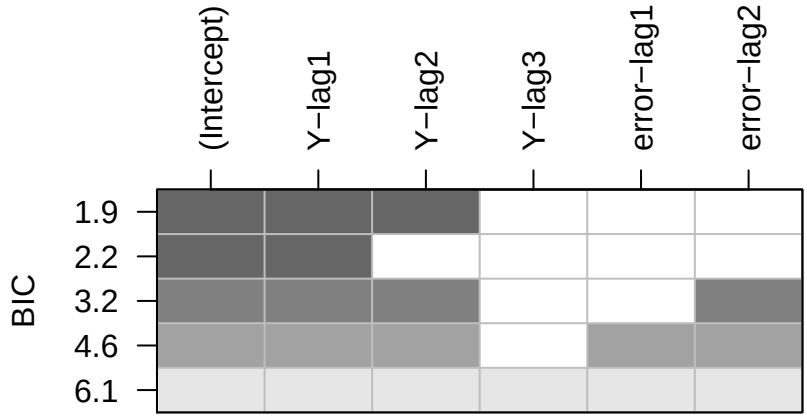

```
eacf(diff( as.numeric(window(avg.price,end=c(2004,3))) ), ar.max = 5, ma.max = 7)
```

```
## AR/MA
##   0 1 2 3 4 5 6 7
## 0 o o o o o o o o
## 1 x x o o o o o o
## 2 o o o o o o o o
## 3 o o o o o o o o
## 4 x o o o o o o o
## 5 o o o o o o o o
```

从 acf 中可以看出，数值基本都在虚线范围内，暗示可能只包含 e_t 项。从 pacf 中可以看出，只有在二阶出现显著，基本是截尾现象，暗示可能是 AR 过程。而 eacf 中没有明显的三角形状，但以 (2,1) 为顶点可以构成一个由 “o” 构成的三角形。由以上的分析认为，AR(2) 模型可能比较适合。

如下图所示，从最优子集的角度讲，引入 AR2 是较为合适的。

```
plot(armasubsets(y= diff( window(avg.price,end=c(2004,3)) ),nar=3,nma=2,ar.method="ols"))
```



3.2.2 引入干预效应

要处理时间序列中的干预分析及异常值，需要利用 **arimax** 函数，该函数是 arima 函数的扩展。假设干扰影响过程的均值，相对未受干扰的均值函数的偏差被称作**传递函数**。传递函数又由动态部分和协变量两部分组成，动态部分为 ARMA 滤波器的形式，表示如下：

$$\frac{(a_0 + a_1 B + \cdots + a_q B^q)}{(1 - b_1 B - b_2 B^2 \cdots - b_p B^p)} \text{协变量}$$

airmax 函数中的参数 **xtransf** 表示传递函数中的协变量，参数 **transfer** 表示传递函数中的动态部分的 (p,q) 阶数。

我的模型设定为：

$$m_t = \frac{a_0 + a_1 B}{1 - b_1 B} P_t^{(T)}$$

$$m_t' = \frac{a_0' + a_1' B}{1 - b_1' B} P_t^{(T')}$$

其中 T 表示 2004 年 5 月的第一次干预，T' 表示 2008 年 1 月的第二次干预。
模型拟合如下：

```
model2.1 = arimax(avg.price, order = c(2,1,0),
            xtransf=data.frame(I524=1*(seq(avg.price)==29),I13=1*(seq(avg.price)>=73)),
            transfer = list( c(1,1),c(1,1) ),
            fixed = c(NA,NA,0.95,NA,NA,0.45,NA,NA) )
# 由于在模型拟合中发现ARMA滤波器的AR项的估计会出现不收敛的情况;
# 因此我直接为b1和b'1赋值0.95和0.45, 没有进行估计, 所以结果的标准误为0
model2.1
```

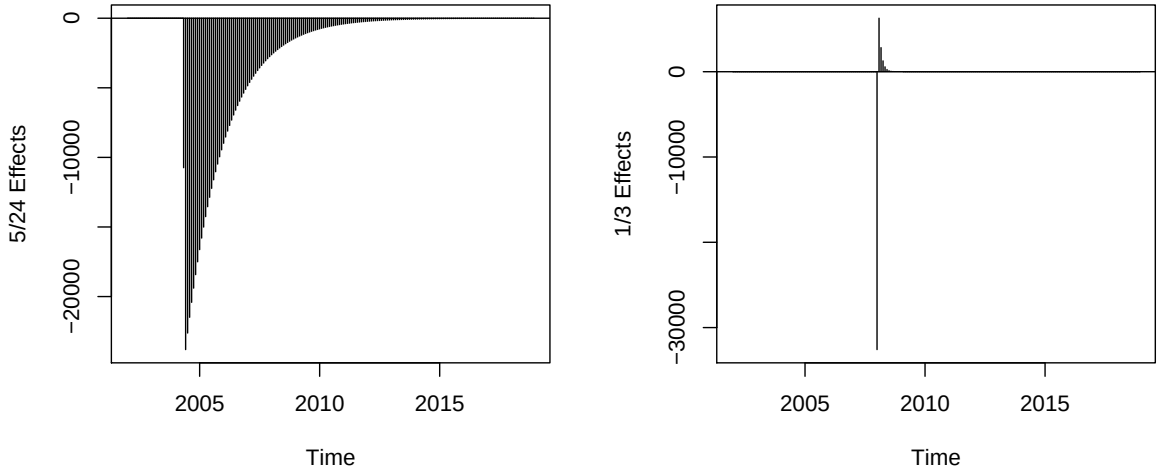
```
##
## Call:
## arimax(x = avg.price, order = c(2, 1, 0), fixed = c(NA, NA, 0.95, NA, NA, 0.45,
##     NA, NA), xtransf = data.frame(I524 = 1 * (seq(avg.price) == 29), I13 = 1 *
##     (seq(avg.price) >= 73)), transfer = list(c(1, 1), c(1, 1)))
##
## Coefficients:
##           ar1      ar2  I524-AR1    I524-MA0    I524-MA1  I13-AR1
##       -0.2455  -0.0767      0.95  -10741.395  -13606.060     0.45
## s.e.   0.0701   0.0705      0.00    3704.742    3696.112     0.00
##          I13-MA0    I13-MA1
##       -32581.076  20933.665
## s.e.    3724.008   3717.556
##
## sigma^2 estimated as 13979446:  log likelihood = -1948.42,  aic = 3908.84
```

可以看到 ar1 和 ar2 的系数并不显著；但是干预效应的系数均十分显著，说明引入干预效应是正确的。（I524 表示 5 月 24 日导致的第一次干预，I13 表示 1 月 3 日导致的第二次干预）

两次干预造成影响的估计如下图所示：

```
op = par(mfrow=c(1,2))
MAY24=1*(seq(avg.price)==29)
plot(ts(MAY24*(13606/0.95)+
        filter(MAY24,filter=.95,method='recursive',side=1)*(-10741-13606/0.95),
        frequency=12,start=2002),type='h',ylab='5/24 Effects')
abline(h=0)

JAN3=1*(seq(avg.price)==73)
plot(ts(JAN3*(-20933/0.45)+
        filter(JAN3,filter=.45,method='recursive',side=1)*(-32581+20933/0.45),
        frequency=12,start=2002),type='h',ylab='1/3 Effects')
abline(h=0)
```



可见 2004 年 5 月 24 日的第一次干预随着时间的推移影响逐渐减少；而 2008 年 1 月 3 日的第二次干预在当月有很大的降低价格的影响，在其后反而出现反作用，然后影响迅速消失，可能说明新的规则并没有很好的起到降低价格的作用。

在已建立模型的基础上，接下来就可以引入可加异常值。

3.2.3 引入可加异常值

首先运用 detectAO 函数检测 model2.1 中的可加异常值

```
detectAO(model2.1)
```

```
##               [,1]       [,2]       [,3]
## ind     108.000000 109.000000 134.000000
## lambda2  -8.139106   6.179851   4.082509
```

结果显示存在三个可加异常值，其位置正是 2010 年 12 月和 2013 年 3 月附近，这与我在第二节中的分析完全一致。

接下来利用 arimax 中的 **xreg** 参数来引入可加异常值，方法为加入 Dec10、Jan11 和 Mar13 三个示性变量代表上述三个可加异常值发生的位置，其他参数不变，模型拟合如下：

```
model2.2 = arimax(avg.price, order = c(2,1,0),
           xtransf=data.frame(I524=1*(seq(avg.price)==29),I13=1*(seq(avg.price)>=73)),
           transfer = list( c(1,1),c(1,1) ),
           xreg=data.frame( Dec10=1*(seq(avg.price)==108), Jan11=1*(seq(avg.price)==109),
                            Mar13=1*(seq(avg.price)==135)),
           fixed = c(NA,NA,NA,NA,NA,0.95,NA,NA,0.45,NA,NA) )

model2.2
```

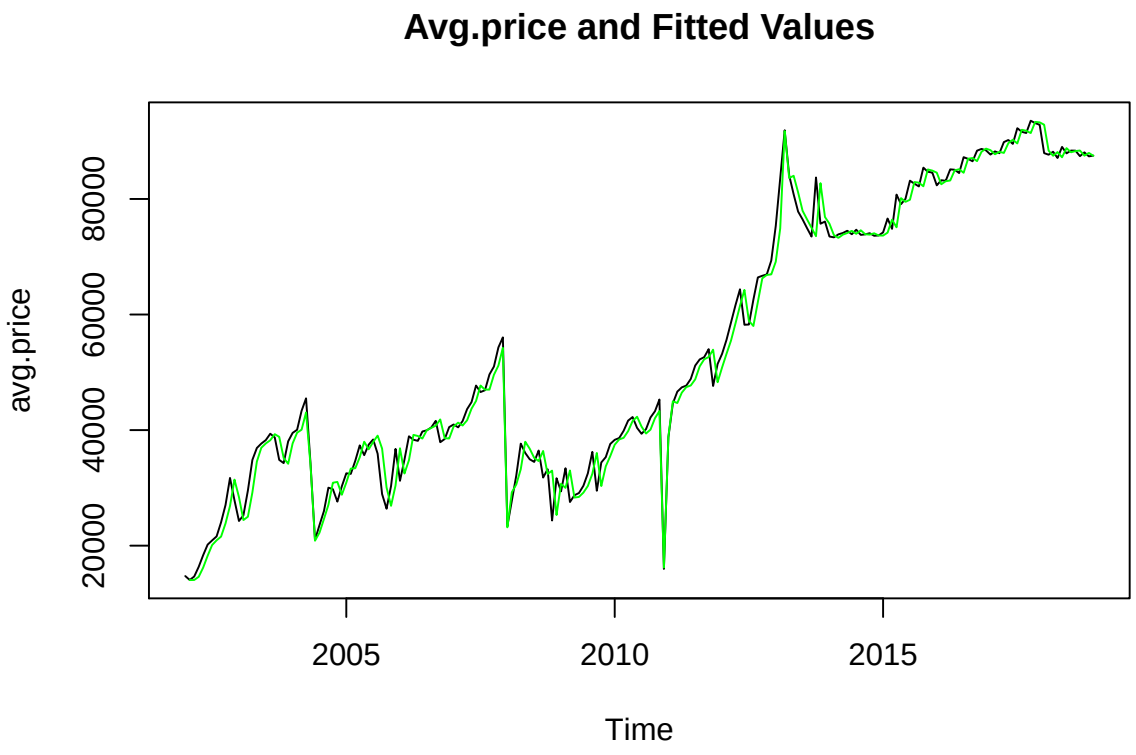
```
##
## Call:
```

```
## arimax(x = avg.price, order = c(2, 1, 0), xreg = data.frame(Dec10 = 1 * (seq(avg.price) ==
##     108), Jan11 = 1 * (seq(avg.price) == 109), Mar13 = 1 * (seq(avg.price) ==
##     135)), fixed = c(NA, NA, NA, NA, NA, 0.95, NA, NA, 0.45, NA, NA), xtransf = data.frame(I
##     (seq(avg.price) == 29), I13 = 1 * (seq(avg.price) >= 73)), transfer = list(c(1,
##     1), c(1, 1)))
##
## Coefficients:
##           ar1     ar2       Dec10      Jan11     Mar13  I524-AR1
##       -0.0903  0.0427  -28980.938  -6213.236  8663.501      0.95
## s.e.   0.0726  0.0720    2177.036   2177.871  2018.412      0.00
##         I524-MA0    I524-MA1  I13-AR1     I13-MA0    I13-MA1
##       -11456.583  -14019.248     0.45  -33011.606  20541.184
## s.e.    2651.046    2640.172     0.00    2680.503   2666.707
##
## sigma^2 estimated as 7120015:  log likelihood = -1880.25,  aic = 3778.5
```

可以看到 ar1 和 ar2 的系数并不显著；但是干预效应的系数仍十分显著，并且可加异常值的系数都十分显著，说明引入可加异常值的正确的。

原数据与拟合值的时间序列图如下：

```
plot(avg.price, main = "Avg.price and Fitted Values")
lines(fitted(model2.2),col="green")
```

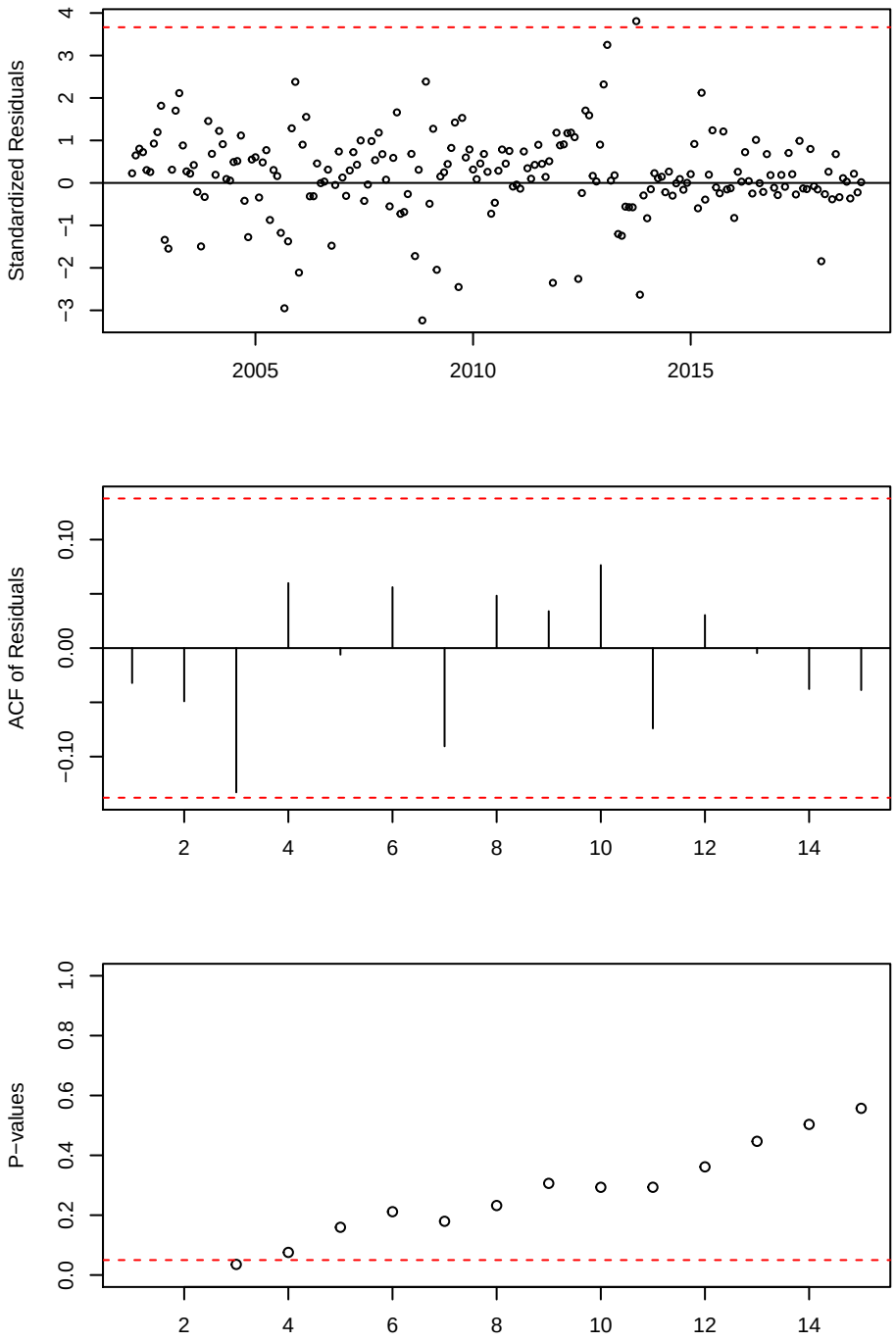


黑线代表原数据，绿线代表拟合值。可以看出绿线与黑线基本吻合，并且在干预处以及异常值的位置拟合效果也很好。

3.2.4 模型的检验

运用标准化残差的时间序列图、残差的 ACF 图和 Ljung-Box 检验图对模型进行检验：

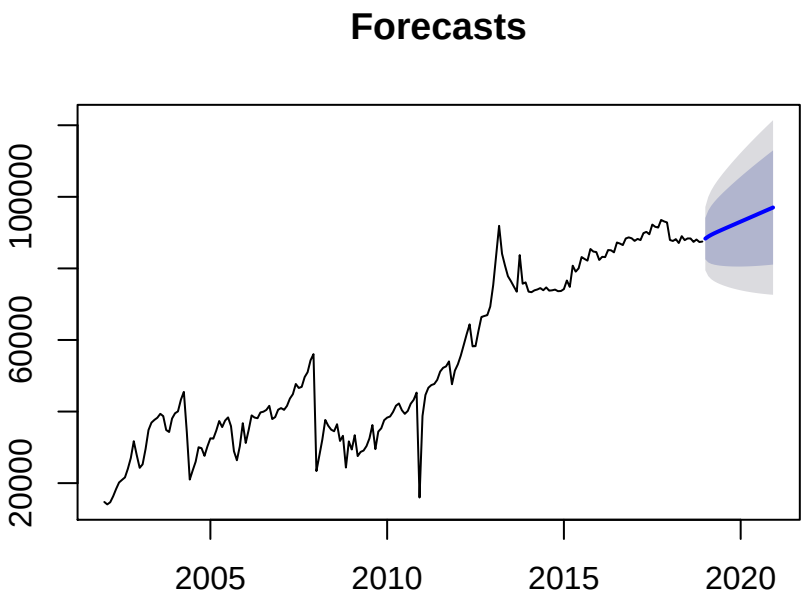
```
tsdiag(model2.2,gof=15)
```



标准化残差基本都在 3 以内，且没有表现出规律性；残差的 ACF 也没有显著项；Ljung-Box 检验在考虑的阶数大于 3 以后也都不显著。以上的检验结果表明模型二拟合效果良好，克服了模型一出现的一些问题。

3.2.5 预测

通过以上的分析，确定了最终的模型——模型二；接下来就可以进一步利用模型进行预测了，结果如下：



黑线为原始数据，蓝线为预测值，阴影代表预测极限。可以看到拍卖价格可能会进一步上升，但是上升的速率会小于 2015 年前的高速上涨。

对于具体的预测值，比如 2019 年 1 月，其最佳预测为 88326.13，95% 置信水平下的置信区间为 (79481.63,97170.64)。由于报告截止前 2019 年 1 月的拍卖价格已在网站中公示，其平均成交价为 89565，符合预测结果，因此说明我的模型是有一定实际价值的。

因为已有 2019 年 1 月的数据，因此其他的预测可以利用预测的更新来提高准确度。

4 分析与结论

4.1 模型的比较分析

从模型拟合的角度讲，通过比较两个模型 “原数据与拟合值的时间序列图” 可以看出，模型二的拟合值与原数据的 “贴合” 程度是要优于模型一的，这体现为两方面：一是模型一的绿线与黑线有一定的缝隙，而模型二的绿线与黑线基本吻合；二是模型一在干预处以及异常值的位置拟合效果并不好，而模型二在干预处以及异常值的位置拟合效果也很好。

另外，虽然模型二引入了更多需要估计的参数，比模型一更为复杂。但模型二的 AIC 为 3778.5，模型一的 AIC 为 3997.76, 模型二的 AIC 实则远小于模型一。这说明引入干预分析与可加异常值极大地提高了拟合效果，因此 AIC 变小。

模型二优于模型一的地方也可以在标准化残差的时间序列图中看出，模型一的残差有几个位置大于 3，这就是没有考虑干预效应和异常值导致的；而模型一的残差基本都在 3 以内，说明没有问题。

最后，由于我在尝试中发现，数据集中其他几个变量对拍卖价格的相关性并不强，引入这几个

变量的信息进行时间序列回归无法得到很好的效果，因此在本文中没有涉及与讨论。可能其他的变量（如上海市的可支配收入，上海市人口等）对拍卖价格的预测有更好的帮助，这值得进一步分析和探究，也是模型可以进一步改进的地方。

4.2 关于车牌价格的结论

首先是可以认为上海私车牌照拍卖价格符合**带有两个干预效应以及三个可加异常值的ARIMA(2,1,0) 过程**。

其次是通过干预分析对影响的估计发现，2008 年 1 月引入的新规则并没有很好的起到降低价格的作用。

最后是对价格的预测可以看出，拍卖价格可能会进一步上升，但是上升的速率会小于 2015 年前的高速上涨。并且可以利用该模型对未来的拍卖价格进行一定的预测，为拍卖者在出价上提供一定的辅助。

5 参考资料

[1] 数据来源：http://chepai.alltobid.com/contents/22/276.html
爬虫数据：https://www.kaggle.com/bogof666/shanghai-car-license-plate-auction-price
[2] 上海车牌拍卖制度 “变形记”[J]. 决策, 2008(4):57-60.
[3] 《上海推出私车牌照额度拍卖新办法》 网址：http://www.shanghai.gov.cn/nw2/nw2314/nw32419/nw32422/index.html