A Research on the Seasonal Effect on the Price of Live Pig: Based on SARIMA Model

Zhang Haochun(Michael), Yu Linxuan(Shawn), Cao Yukun(Henry), Lu Mingyang(Leo)
Beijing National Day School

June 9, 2023

Keywords— Price of live pig, Hypothesis Test, Time series, ARIMA, SARIMA

1 Introduction

The past forty years since the Reform and Opening Up have seen a dramatic rise of Chinese pig production, from less than 8 million tons before 1978 to over 50 tons in the 2010s[1]. In 2022, the exact number is 55.41 million tons, which is significantly larger than the world's second biggest pork producer, the EU who produced 22.46 million tons. China is not only the greatest pork producer, but also the greatest single consumer of pork on the work [2]. The fluctuation of pig price can be found in the national bureau of statistics of China. For example, due to the effect of swine fever, in August 2020, pig price rose to a highest recorded price of 37.15 yuan/kg . Yet, as this unusual fluctuation had drawn scholar's attention in and abroad China, it is worth noting that seasonal factor is also identified as one among the most influential factors of Chinese pork price.

Therefore, our research aims to examine existence of the statistical relationship between seasonality and the price of live pig. The research will be useful in various ways. First, it will increase our understanding about the formation of pork price fluctuation, and allows us to further predict the pork price and the subsequent economic profits or losses. The research may be useful to policy makers to offer some insight into the live pig market, and can be used as reference for planning intervention measures and policy formulation, ultimately protecting the economic benefits of pig farmers and the industry as a whole.

The research is based on the statistics provided by Chinese government from January, 2003 to April, 2023. The months are then divided into seasons according to climate. In the first part of our research, we had performed a two sample Z-test to examine whether season has an effect on price of live pig or not. At the significance level of 0.01, there is convincing evidence that the correlation exists. After that, we applied the SARIMA model to analyze an unstationary time series that contains an periodic component. Quantile-quantile plot (Q-Q plot) is used to show the residuals of the model. Ljung box tests are also conducted to examine the existence of possible outliers. Finally, we changed the period parameter and hypothesized that multiple periodic trends other than seasonal trends exist in the time series.

2 Research

Recent years, the price of live pig varies frequently with a roughly periodic pattern. Therefore, to study the fluctuation pattern and growth trend of our country's pork price, and put forward corresponding measures to regulate it, is not only the need for the healthy development of the pig breeding industry, but also the need for stabilizing residents' meat consumption. Many foreign scholars has studied on the price fluctuations of live pigs and pork. In order to explain the price fluctuations of agricultural products such as live pigs, Kaldor introduced a theory about the relative difference in the elasticity of supply and demand to analyze the market stability conditions, and named it "cobweb theory". Ezekiel[4] used the cobweb model dynamic analysis method to study the fluctuation process and results of pig supply and price after they deviate from the equilibrium state. Talpaz[5] proposed the high-frequency cobweb theory, pointing out that pork and other agricultural products may have several short cycles in a long cycle, and short cycles need to be identified in a long cycle. For recent studies, Liu in 2021 had identified different factors that affected the fluctuation of pig price, including seasonal factors, and claimed that seasonal changes of pork price is caused by the "seasonality" and "festivity" of pork demands

In a research provided by Huang[7], the researchers used ARIMA model to predict stock market, analyzing the data by first using the ADF test, and the original model showed a p-value greater than 0.05, so he used a d-order difference, and the new result had a p-value way less than 0.05. Then the author took a white noise test, and partial autocorrelation residual test and resulted in a not white noise. This research provide experience of applying ARIMA model in economic field. There are similarities between stock and price of live pig, since they are both random walk and are effected by multiple factors that could include periodic factors. Huang's research of applying ARIMA in stock predicting can also be used in analyzing the price of live pig.

3 Data Collection and Analysis

3.1 Data collection

From the official website of the National Bureau of Statistics of China we have collected price of live pig in China from 2003 January to 2023 April. The raw data is shown in the Appendix.

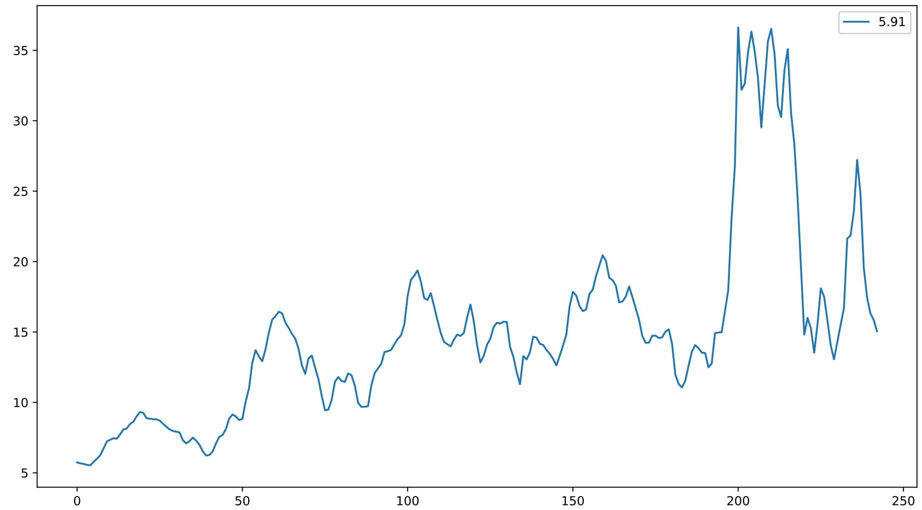

Figure 1: scaterplot of price of live pig over time

Figure 1 is a scaterplot of the raw data. From the data we observed that the price of live pig after 2018 is significantly higher. It is caused by the COVID-19 pandanmic, according to our background research. The data before 2008 is lower, this is because the PRRSV epidemic happened in 2008 and raised up the price of live pig. In order to exclude the factors mentioned above we decide to focus our study mainly on the data from 2007 to 2018. The new scaterplot of the new data set is shown in figure 2.

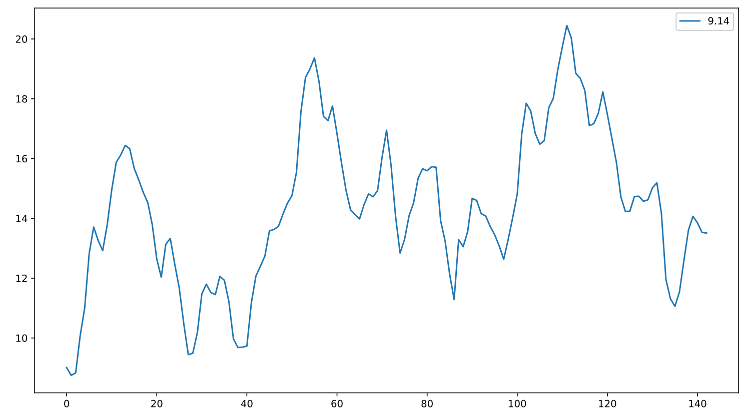

Figure 2: scaterplot of price of live pig from 2007 to 2018

3.2 Two Sample Z test on the seasonal effect on price of live pig

From the data set we can compare the mean price of live pig in a specific season and the mean price of live pig in the whole data set. A two sample Z test is presented.

H_0 : There is no significant difference between the mean price of live pig in autumn,from July to September and the mean price of live pig in the whole data set.
H_a : The mean price of live pig in autumn is lower than the mean price of live pig in the whole data set.

	Autumn	Year
sample size	36	144
mean	13.49667	14.30722
standard deviation	3.053688	2.589309

Table 1: mean value and standard deviation of the data of the sample

Table 1 shows the mean value and standard deviation of the data. The corresponding test statistics and p-value are $Z = -1.4662641$ and $p - value = 0.07128817$.

We can reject the null hypothesis under a 10% significance level. The mean price of live pig in autumn is lower than the mean price of live pig in the whole year. This proves that there is a seasonal effect on the price of live pig in China. However, the result of the hypothesis test is not persuasive enough. In order to provide more evidence on the existence of seasonal effect on the price of live pig, a time series analysis should be done.

3.3 Applying ARIMA and SARIMA model on the price of live pig

ARMA model, also known as autoregressive-moving average model, is a combination of autoregressive and moving average that analyzes and predicts time series. ARMA model can only deal with stationary process. ARIMA model, as an generalization of ARMA model, can be used to comprehened unstationary process. ARIMA model can turn unstationary process into stationary process through differencing the data. SARIMA model is an extension of ARIMA model that specifies in dealing with data that contains a periodic component which ARIMA model don't considerate. By comparing the results of ARIMA and SARIMA model we are able to see if the seasonal effect exist on the data.

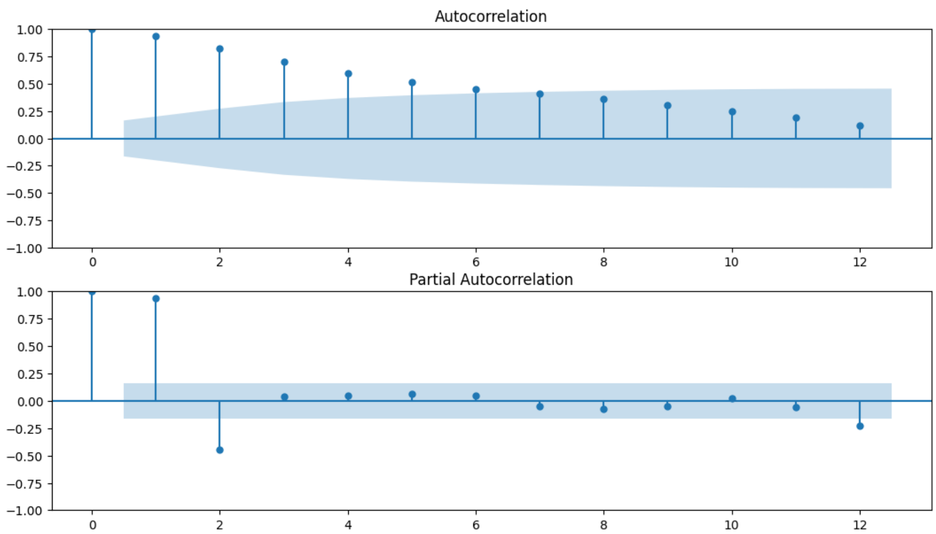

Figure 3: ACF and PACF of raw data

Figure 3 is the autocorrelation and partial autocorrelation coefficient plot. The autocorrelation coefficient does not cut off or tail off, and the partial autocorrelation coefficient shows a strong correlation with 1 lag. Thus, the raw data set is not stationary, and it needs to be differenced. The partial autocorrelation coefficient shows that a first order difference is enough to make the data stationary. The scatterplot and ACF and PACF plot are shown in Figure 4 and Figure 5.

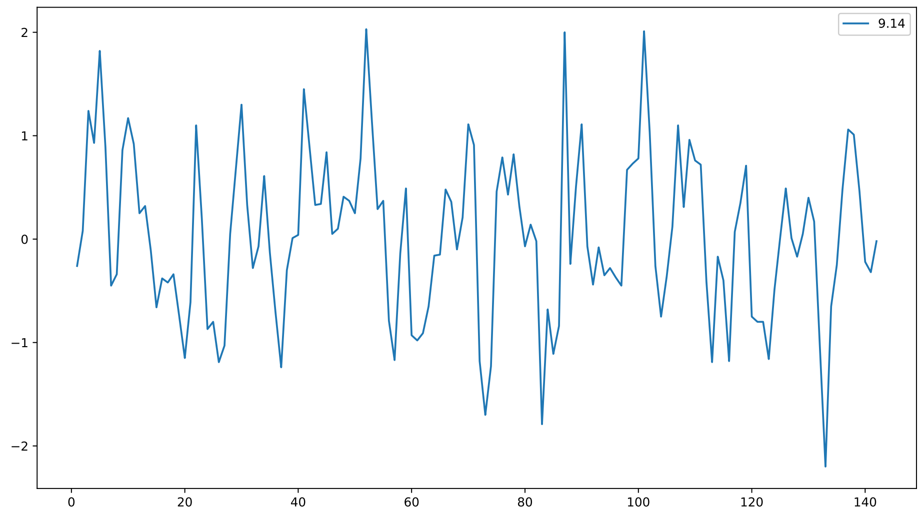

Figure 4: price of live pig after first order difference

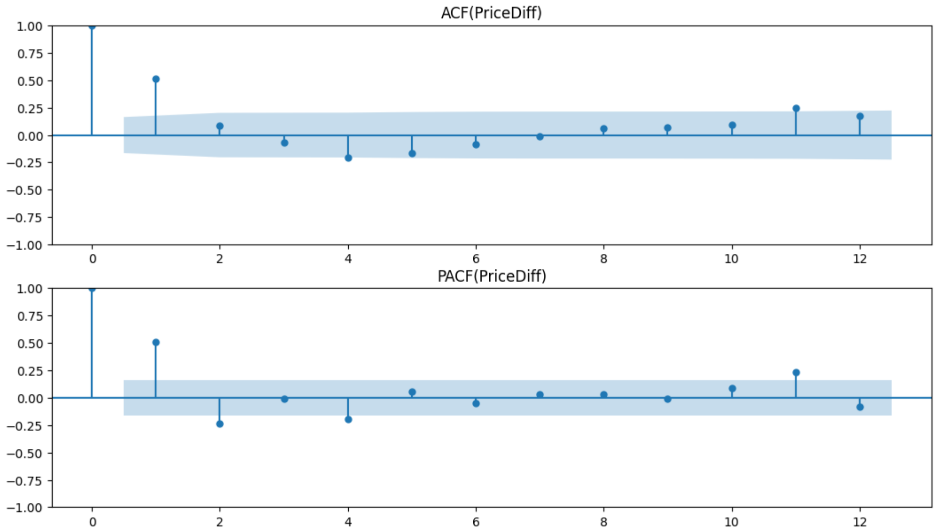

Figure 5: ACF and PACF of differenced data

After the first order difference both ACF and PACF tail off. We applied a ADF test to the differenced data and the results are as bellow:

$$H_0 : \text{the process has a unit root}$$
$$H_a : \text{the process is weakly stationary}$$
$$\text{Test statistic=-4.160, P-value=0.001}$$

Under a 99% significance level, we can conclude that the data after first order difference is weakly stationary. Thus, an ARIMA model with coefficient (2,1,1) or a SARIMA model with coefficient(2,1,1,2,1,2,12) can be applied. The coefficient is determined combining the automatic parameter tuning and the ACF and PACF plot. The graph of ARIMA and SARIMA model using first 132 month as train set and last 12 month as test set is shown in figure 6 and figure 7. The blue line represents the real data and the orange line represents the data calculated by the ARIMA model

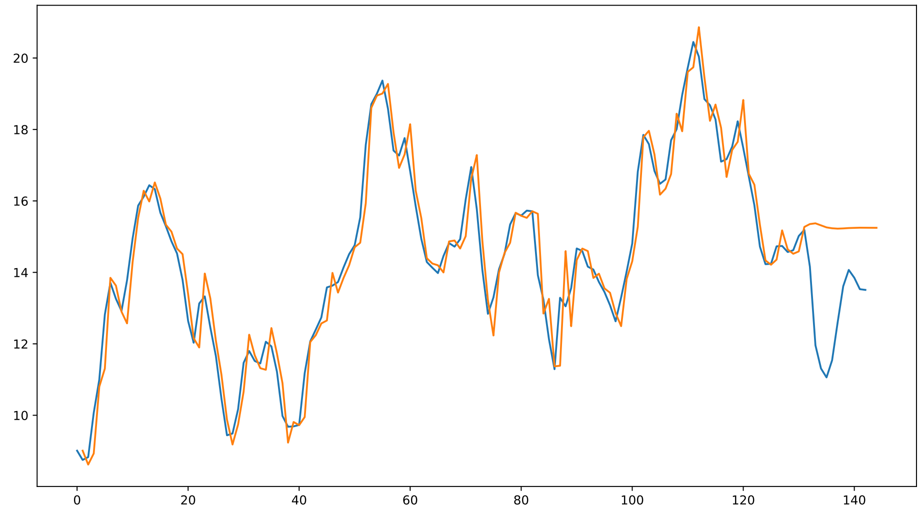

Figure 6: ARIMA model forcasting the price of live pig

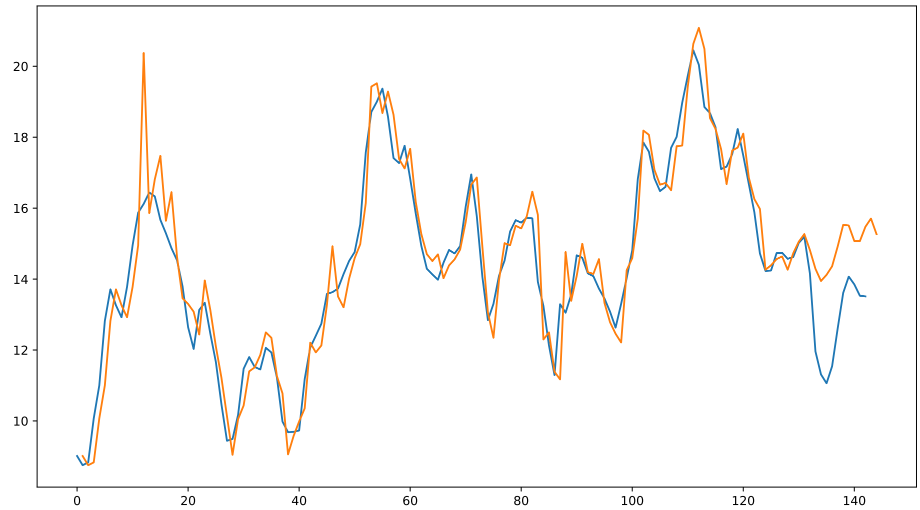

Figure 7: SARIMA model forcasting the price of live pig

Figure 6 shows that ARIMA model failed to predict the price of live pig. The train set were fitted well by the model, but the model predicts a flat trend which is impossible on the price of live pig. There could be a overfitting in the ARIMA model. Figure 7 shows that SARIMA model partially reflects the trend of price, and correctly predicts the turning point of the price, but there's still a nonnegligible error in the prediction. The error could be caused by unpredictable random factors such as weather and disease, or there could be factors that are not considered in the model. However, the scatterplot shows that adding seasonal factor into ARIMA model does improve the prediction.

3.4 Residual analysis

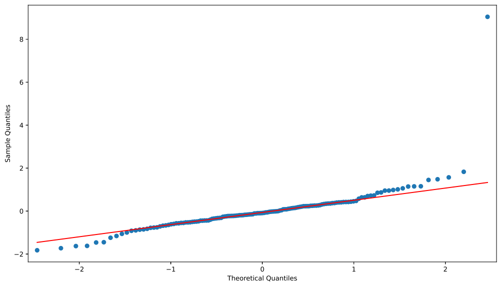

Figure 8: ARIMA Residual QQplot

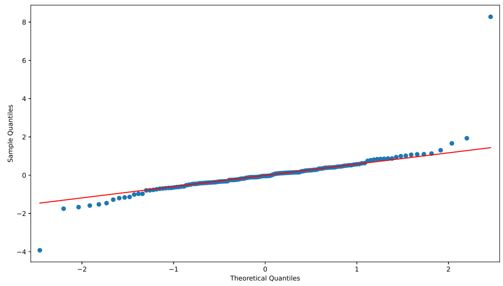

Figure 9: SARIMA Residual QQplot

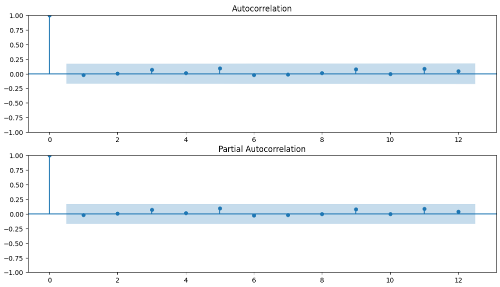

Figure 10: ARIMA Residual QQplot

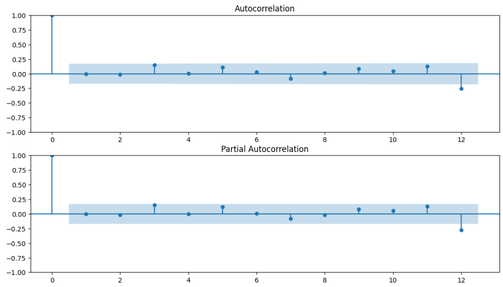

Figure 11: SARIMA Residual QQplot

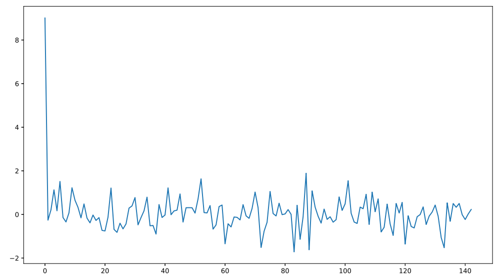

Figure 12: ARIMA Residual QQplot

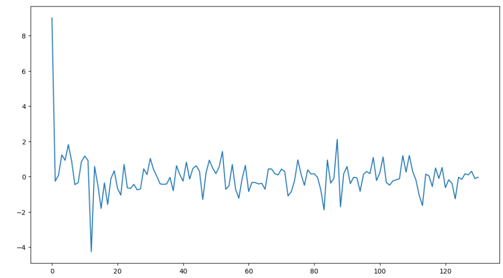

Figure 13: SARIMA Residual QQplot

Figure 8 to 13 are quantile-quantile plots, ACF and PACF plots, and scaterplots for the residual of ARIMA and SARIMA model. The figure shows that the residual are approaximately normally distributed and are independent of each other. However, when we applied a ljung box test for the residual of SARIMA model, there is a significant decrease in p-value when lags=12 which is the seasonal period we choosed in the SARIMA model. The result of the ljung box test is shown in table 2.

$$H_0 : \text{The distribution of residual is a white noise.}$$
$$H_a : \text{The distribution of residual is not independent.}$$

lags	test statistics	P-value
1	0.040897	0.839738
6	4.462300	0.614373
12	18.092090	0.112923
18	26.354471	0.091925
24	28.963085	0.221523

Table 2: ljung box test result for SARIMA residual

Similar pattern is not observed in the ARIMA model. Interestingly, as we change the period parameter in the SARIMA model, the number of lags which the decrease in p-value occur changes correspondingly. Our group suspects that there are multiple periodic trend in the time series, seasonal trend is just a component of it. These periodic trend forms a periodic trend with a long period together. The period of the trend can be up to years. However, we are unable to discover that trend due to lack of data. Our data of price of live pig in the past 20 years in China might be insufficient with respect to the long period, since the length of the period can be very long.

4 conclusion and future suggestion

4.1 Conclusion

Our project is aiming to find how the price of live pig is affected by the season. We collect data of the price of live pig in 20 years(from 2003 January to 2023 April) through official website of the National Bureau of Statistics of China and make a scatterplot to view the data clearly. To find whether the there is a seasonal effect, we do a two sample z test. The null hypothesis is there is no difference between the price of pig in summer and in other seasons and the alternative hypothesis that there is a difference. We get a p-value of 0.07 and conclude that there is a seasonal effect on the price of live pig but we need more evidence. Therefore, we apply ARIMA and SARIMA model on the price of the live pig and draw the graphs. We apply a ADF test to the differenced data and the null hypothesis is the process has a unit root and the alternative hypothesis is that the process is weakly stationary. The p-value is 0.001 so that we can confirm the alternative hypothesis. In predicting future change of price in the live pig, as we do prediction without a seasonal factor, the ARIMA model failed due to a flat trend which is impossible. As we added the seasonal factor, the graph have the same periodic trend same as before, which improves our prediction. Finally, we extents the research and suspect through the SARIMA model that the price of live pig is affected by multiple periodic factors.

4.2 Future suggestion

For improving our experiment, extenting the data set is helpful. Now we only have 20 years, and the several beginning years and ending years are not representative due to prevalent diseases or other situations that affected whole China. The persuasive data are only about 12 years, so we need more data before 2003. Moreover, the method we use, ARIMA and SARIMA, are not that accurate. We can use neutral network as a more precise method. Neural networks are artificial systems that were inspired by biological neural networks. These systems learn to perform tasks by being exposed to various datasets and examples without any task-specific rules. The idea is that the system generates identifying characteristics from the data they have been passed without being programmed with a pre-programmed understanding of these datasets. Using neutral networks means using artificial system to find the relationship between the price of live pig and seasons, which is much better from testing by ourselves.

References

[1] Zhang, S., Wu, X., Han, D., Hou, Y., Tan, J., Kim, S. W., Li, D., Yin, Y., & Wang, J. (2021). PORK PRODUCTION SYSTEMS IN CHINA: A REVIEW OF THEIR DEVELOPMENT, CHALLENGES AND PROSPECTS IN GREEN PRODUCTION. Frontiers of Agricultural Science and Engineering, 8(1), 15. https://doi.org/10.15302/j-fase-2020377

[2] United States Department of Agriculture, 2023. Livestock and poultry: world markets and trade.

[3] JIANG Liuyi, DING Jiayun, ZHOU Zhengping. (2014). Research on pork price fluctuation law and regulatory countermeasures in China-Empirical analysis based on H-P filter method. Explorations in Economic Issues, 9, 96–101. http://www.cqvip.com/QK/95595X/201409/662015370.html

[4] Ezekiel, M. (2010). The Cobweb Theorem. http://academic.oup.com/qje/article/52/2/255/1862362

[5] Talpaz, H. (1974). Multi-Frequency Cobweb Model: Decomposition of the Hog Cycle. American Journal of Agricultural Economics, 56(1), 38–49. https://doi.org/10.2307/1239345

[6] Liu, Yangfang. (2021). Analysis and Trend Anticipation of the Fluctuations of Chinese Swine Price. Chinese Swine Industry, 7(6), 18-20.

[7] Huang, M. (2023). Stock price predicting research based on ARIMA model. Nei Jiang Ke Ji.

Appendix

Raw data

accumulated month	market price of live pig
1	5.91
2	5.74
3	5.67
4	5.63
5	5.56
6	5.53
7	5.78
8	6
9	6.25
10	6.74
11	7.23
12	7.34
13	7.46
14	7.42
15	7.73
16	8.08
17	8.15
18	8.46
19	8.63
20	9.01
21	9.32
22	9.26
23	8.88
24	8.84
25	8.8
26	8.8
27	8.71
28	8.48
29	8.27
30	8.08
31	7.97
32	7.92
33	7.87
34	7.31
35	7.09
36	7.25
37	7.5
38	7.29
39	6.99
40	6.53
41	6.23
42	6.26
43	6.52
44	7.08
45	7.56
46	7.68
47	8.1
48	8.85
49	9.14
50	9.01
51	8.75
52	8.83
53	10.07
54	11
55	12.82
56	13.71
57	13.26
58	12.92

59	13.78
60	14.95
61	15.87
62	16.12
63	16.44
64	16.33
65	15.67
66	15.29
67	14.87
68	14.53
69	13.79
70	12.64
71	12.03
72	13.13
73	13.33
74	12.46
75	11.66
76	10.47
77	9.44
78	9.49
79	10.17
80	11.47
81	11.8
82	11.52
83	11.45
84	12.06
85	11.93
86	11.22
87	9.98
88	9.68
89	9.69
90	9.73
91	11.18
92	12.07
93	12.4
94	12.74
95	13.58
96	13.63
97	13.73
98	14.14
99	14.51
100	14.76
101	15.54
102	17.57
103	18.71
104	19
105	19.37
106	18.58
107	17.41
108	17.27
109	17.76
110	16.83
111	15.85
112	14.94
113	14.29
114	14.13
115	13.98
116	14.46
117	14.82
118	14.72
119	14.93
120	16.04
121	16.95

122	15.77
123	14.07
124	12.84
125	13.3
126	14.09
127	14.52
128	15.34
129	15.66
130	15.59
131	15.73
132	15.71
133	13.92
134	13.24
135	12.13
136	11.29
137	13.29
138	13.05
139	13.56
140	14.67
141	14.6
142	14.16
143	14.08
144	13.73
145	13.45
146	13.08
147	12.63
148	13.3
149	14.03
150	14.81
151	16.82
152	17.85
153	17.59
154	16.84
155	16.48
156	16.6
157	17.7
158	18.01
159	18.97
160	19.73
161	20.45
162	20.04
163	18.85
164	18.68
165	18.28
166	17.1
167	17.17
168	17.52
169	18.23
170	17.48
171	16.68
172	15.88
173	14.72
174	14.23
175	14.24
176	14.73
177	14.74
178	14.57
179	14.62
180	15.02
181	15.19
182	14.16
183	11.96
184	11.31

185	11.06
186	11.54
187	12.6
188	13.61
189	14.07
190	13.85
191	13.53
192	13.51
193	12.49
194	12.77
195	14.92
196	14.96
197	14.98
198	16.45
199	17.97
200	23.08
201	26.77
202	36.62
203	32.2
204	32.63
205	34.89
206	36.33
207	34.92
208	33.02
209	29.52
210	32.63
211	35.63
212	36.52
213	34.77
214	31.08
215	30.26
216	33.58
217	35.09
218	30.56
219	28.33
220	24.42
221	19.64
222	14.81
223	16
224	15.27
225	13.53
226	15.54
227	18.1
228	17.51
229	15.83
230	14.06
231	13.06
232	14.31
233	15.51
234	16.69
235	21.63
236	21.85
237	23.56
238	27.22
239	24.82
240	19.58
241	17.45
242	16.33
243	15.86
244	15.05

Data source

Code

Code for ARIMA

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import statsmodels.api as sm
from scipy import stats
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.stats.diagnostic import acorr_ljungbox

from statsmodels.graphics.api import qqplot

#data input
price = pd.read_excel("D:\Statistics\project\price.xlsx")
real = pd.read_excel("D:\Statistics\project\data.xlsx")
price_diff = price.diff(1).dropna()
print (price)
price.plot(figsize=(12, 8))
price_diff.plot(figsize=(12,8))
#ACF,PACF
fig = plt.figure(figsize=(12, 8))
ax1 = fig.add_subplot(211)
fig = sm.graphics.tsa.plot_acf(price.values.squeeze(), lags=12, ax=ax1)
ax2 = fig.add_subplot(212)
fig = sm.graphics.tsa.plot_pacf(price, lags=12, ax=ax2)
fig = plt.figure(figsize=(12, 8))
ax3 = fig.add_subplot(211)
fig = sm.graphics.tsa.plot_acf(price_diff.values.squeeze(), lags=12, ax=ax3)
ax3.set_title('ACF(PriceDiff)')
ax4 = fig.add_subplot(212)
fig = sm.graphics.tsa.plot_pacf(price_diff, lags=12, ax=ax4)
ax4.set_title('PACF(PriceDiff)')
#ARIMA
arma_mod110 = ARIMA(price, order=(2, 1, 1)).fit()
print(arma_mod110.params)
#aic,bic,hqic and durbin watson test
print(arma_mod110.aic, arma_mod110.bic, arma_mod110.hqic)
print(sm.stats.durbin_watson(arma_mod110.resid.values))
#Resid
res = acorr_ljungbox(arma_mod110.resid, lags=[1,6,12,24,36], return_df=True)
print(res)
fig = plt.figure(figsize=(12, 8))
ax = fig.add_subplot(111)
ax = arma_mod110.resid.plot(ax=ax)
resid = arma_mod110.resid
print(stats.normaltest(resid))
fig = plt.figure(figsize=(12, 8))
ax = fig.add_subplot(111)
fig = qqplot(resid, line="q", ax=ax, fit=True)
fig = plt.figure(figsize=(12, 8))
ax1 = fig.add_subplot(211)
fig = sm.graphics.tsa.plot_acf(resid.values.squeeze(), lags=12, ax=ax1)
ax2 = fig.add_subplot(212)
fig = sm.graphics.tsa.plot_pacf(resid, lags=12, ax=ax2)
fig = plt.figure(figsize=(12, 8))
plt.plot(real)
plt.plot(arma_mod110.predict(1,144))

plt.show()
```

Code for SARIMA

```
import numpy as np
import pandas as pd
import warnings
warnings.filterwarnings("ignore")
from tqdm import tqdm_notebook
from itertools import product
from matplotlib import pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from arch.unitroot import ADF
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.seasonal import STL
from statsmodels.stats.diagnostic import acorr_ljungbox
from statsmodels.tsa.stattools import arma_order_select_ic
from statsmodels.tsa.arima.model import ARIMA
```

```python
from statsmodels.tsa.statespace.sarimax import SARIMAX
from statsmodels.graphics.api import qqplot
import statsmodels.api as sm
from scipy import stats




#price
price = pd.read_excel("D:\Statistics\project\price.xlsx")
real = pd.read_excel("D:\Statistics\project\data.xlsx")


# price(yeardiff)
price_year = price.diff(12).dropna()

# price first orderdiff after yeardiff
price_season = price_year.diff(1).dropna()

#price after first order diff
price_diff = price.diff(1).dropna()

fig, ax = plt.subplots(2, 2)

ax[0][0].plot(price_season)
ax[0][0].set_title('price_season')
ax[0][1].plot(price_year)
ax[0][1].set_title('price_year')
ax[1][0].plot(price)
ax[1][0].set_title('price')
ax[1][1].plot(price_diff)
ax[1][1].set_title('price_diff')

#ACF,PACF
fig, ax = plt.subplots(4, 2)
fig.subplots_adjust(hspace=0.5)

plot_acf(price_season, lags=12, ax=ax[0][0])
ax[0][0].set_title('ACF(price_seasondiff)')
plot_pacf(price_season, lags=12, ax=ax[0][1])
ax[0][1].set_title('PACF(price_seasondiff)')

plot_acf(price_year, lags=12, ax=ax[1][0])
ax[1][0].set_title('ACF(price_season)')
plot_pacf(price_year, lags=12, ax=ax[1][1])
ax[1][1].set_title('PACF(price_season)')

plot_acf(price, lags=12, ax=ax[2][0])
ax[2][0].set_title('ACF(Price)')
plot_pacf(price, lags=12, ax=ax[2][1])
ax[2][1].set_title('PACF(Price)')

plot_acf(price_diff, lags=12, ax=ax[3][0])
ax[3][0].set_title('ACF(PriceDiff)')
plot_pacf(price_diff, lags=12, ax=ax[3][1])
ax[3][1].set_title('PACF(PriceDiff)')


#ADF test
adf_diff = ADF(price_diff)
print("price_diff")
print(adf_diff.summary().as_text())
adf=ADF(price)
print("price")
print("price",adf.summary().as_text())
adf=ADF(price_season)
print("price_seasondiff")
(adf.summary().as_text())
adf=ADF(price_year)
print("price_year")
print(adf.summary().as_text())


#white noise test
res = acorr_ljungbox(price, lags=[15], return_df=True)
print(res)
```

```python


#generating parameter
p=range(0,12)
d=range(0,2)
q=range(0,12)
P_season=range(0,3)
D_season=range(0,2)
Q_season=range(0,3)

params_list=list(product(p,d,q,P_season,D_season,Q_season))

def paramSearch(data:np.array,params_list):
    result = []
    best_bic = 10086
    for param in tqdm_notebook(params_list):
        model = SARIMAX(price,order = (param[2],param[1],param[1]),seasonal_order = (
    param[2],param[1],param[2],12)).fit(disp=-1)
        bic = model.bic
        if bic < best_bic :
            best_bic = bic
            best_model = model
            best_param = param
        order1 = (param[0],param[1],param[2])
        order2 = (param[3],param[4],param[5])
        param='SARIMA{0}x{1}'.format(order1,order2)
        print(param)
        result.append([param,model.bic])

#resultTable = paramSearch(price , params_list)


#SARIMA
model = SARIMAX(price, order=(2,1,1), seasonal_order=(1,1,1,12)).fit()
bic=model.bic
model.summary()
print(model.params)
print(bic)

#Resid
res = acorr_ljungbox(model.resid, lags=[1,6,12,18,24], return_df=True)
print(res)
fig = plt.figure(figsize=(12, 8))
ax = fig.add_subplot(111)
ax = model.resid.plot(ax=ax)
resid = model.resid
print(stats.normaltest(resid))
fig = plt.figure(figsize=(12, 8))
ax = fig.add_subplot(111)
fig = qqplot(resid, line="q", ax=ax, fit=True)
fig = plt.figure(figsize=(12, 8))
ax1 = fig.add_subplot(211)
fig = sm.graphics.tsa.plot_acf(resid.values.squeeze(), lags=12, ax=ax1)
ax2 = fig.add_subplot(212)
fig = sm.graphics.tsa.plot_pacf(resid, lags=12, ax=ax2)
fig = plt.figure(figsize=(12, 8))
plt.plot(real)
plt.plot(model.predict(1,143))
plt.show()
```