National Taiwan University
Institute of Industrial Engineering
Jakey BLUE

Time Series Analytics

111-1 Homework #07

Due at 23h59, November 27, 2022; files uploaded to NTU-COOL

1. (10%) Given the model $y_t = a + bt + c_t + x_t$, where a, b are constants, c_t is deterministic and periodic with period s and x_t is a SARIMA$(p, 0, q) \times (P, 1, Q)_s$. What is the model for $w_t = y_t - y_{t-s}$?

2. (10%) Identify the following as certain multiplicative SARIMA models:
   (a) $y_t = 0.5y_{t-1} + y_{t-4} - 0.5y_{t-5} + a_t - 0.3a_{t-1}$
   (b) $y_t = y_{t-1} + y_{t-12} - y_{t-13} + a_t - 0.5a_{t-1} - 0.5a_{t-12} + 0.25a_{t-13}$

3. (10%) If the characteristic function of an AR time series model is
   $$(1 - 1.6B + 0.7B^2)(1 - 0.8B^{12})$$
   (a) Is the model stationary?
   (b) Identify the model as a certain SARIMA model.

4. (15%) Suppose $y_t = y_{t-4} + a_t$ with $y_t = a_t$, for $t = 1, 2, 3, 4$.
   (a) Find the variance function for y_t.
   (b) Find the autocorrelation function for y_t.
   (c) Identify the model for y_t as a certain SARIMA model.

5. (15%) Consider the famous time series data "co2" (monthly carbon dioxide through 11 years in Alert, Canada).
   (a) Fit a deterministic regression model in terms of months and time. Are the regression coefficients significant? What is the adjusted R-squared? (Note that the month variable should be treated as categorical and transformed into 11 dummy variables.)
   (b) Identify, estimate the SARIMA model for the co2 level.
   (c) Compare the two models above, what do you observe?

Dhanabordee Meluintharanggur T11902203 TSA HW# 07

1. (10%) Given the model $y_t = a + bt + c_t + x_t$, where a, b are constants, c_t is deterministic and periodic with period s and x_t is a SARIMA$(p, 0, q) \times (P, 1, Q)_s$. What is the model for $w_t = y_t - y_{t-s}$?

Given $y_t = a + bt + c_t + x_t$

a, b : constant

c_t : deterministic, period s

x_t : SARIMA $(p,0,q) \times (P,1,Q)_s$

Find the model for $w_t = y_t - y_{t-s}$

$w_t = a + bt + c_t + x_t - [a + b(t-s) + c_{t-s} + x_{t-s}]$

$w_t = bs + c_t - c_{t-s} + x_t - x_{t-s}$

$w_t = bs + \nabla_s x_t$

w_t follows the ARMA$(p,q) \times (P,Q)_s$ with "bs" as an added constant

2. (10%) Identify the following as certain multiplicative SARIMA models:
(a) $y_t = 0.5y_{t-1} + y_{t-4} - 0.5y_{t-5} + a_t - 0.3a_{t-1}$
(b) $y_t = y_{t-1} + y_{t-12} - y_{t-13} + a_t - 0.5a_{t-1} - 0.5a_{t-12} + 0.25a_{t-13}$

a) $Y_t = 0.5Y_{t-1} + Y_{t-4} - 0.5Y_{t-5} + a_t - 0.3a_{t-1}$

$Y_t - Y_{t-4} = 0.5(Y_{t-1} - Y_{t-5}) + a_t - 0.3a_{t-1}$

I, AR, MA

∵ note that $s = 4$

Y_t follows ARIMA$(1,0,1) \times (0,1,0)_4$ model, with $\Phi_1 = 0.5$, $\theta_1 = 0.3$, $s = 4$

b) $Y_t = Y_{t-1} + Y_{t-12} - Y_{t-13} - 0.5a_{t-1} - 0.5a_{t-12} + 0.25a_{t-13}$

$(Y_t - Y_{t-1}) - (Y_{t-12} - Y_{t-13}) = -0.5a_{t-1} - 0.5a_{t-12} + (0.5)(0.5)a_{t-13}$

Y_t follows ARIMA$(0,1,1) \times (0,1,1)_{12}$ model with $\theta_1 = 0.5$, $\theta_{s1} = 0.5$, $s = 12$

3. (10%) If the characteristic function of an AR time series model is

$$(1 - 1.6B + 0.7B^2)(1 - 0.8B^{12})$$

(a) Is the model stationary?
(b) Identify the model as a certain SARIMA model.

AR Time Series $(1 - 1.6B + 0.7B^2)(1 - 0.8B^{12})$

ϕ_1 ϕ_2 Φ

a) Since $\Phi_1 = 0.8$, the seasonal part of the model is stationary

Check the roots of the nonseasonal AR(2) part.

The stationary conditions for AR(2) are

$\phi_1 + \phi_2 < 1$ $\quad -1.6 + 0.7 < 1$: ✓

$\phi_2 - \phi_1 < 1$ $\quad -1.6 - 0.7 < 1$: ✓

$|\phi_2| < 1$ $\quad |0.7| < 1$: ✓

Thus the nonseasonal part is stationary

∴ Since both the seasonal and nonseasonal parts are stationary, the complete model is stationary

b) Identify the model as a certain ARIMA model

The model is ARIMA$(2,0,0) \times (1,0,0)_{12}$ with $\phi_1 = 1.6$, $\phi_2 = -0.7$, $\Phi = 0.8$, $s = 12$

$0 = (1 - 1.6B + 0.7B^2)(1 - 0.8B^{12})\, Y_t$

$0 = (1 - 0.8B^{12} - 1.6B + (1.6)(0.8)B^{13} + 0.7B^2 - (0.7)(0.8)B^{14})\, Y_t$

$Y_t = (0.8B^{12} + 1.6B - (1.6)(0.8)B^{13} - 0.7B^2 + (0.7)(0.8)B^{14})\, Y_t$

$Y_t = 1.6Y_{t-1} - 0.7Y_{t-2} + 0.8Y_{t-12} - (1.6)(0.8)Y_{t-13} + (0.7)(0.8)Y_{t-14} + a_t$

4. (15%) Suppose $y_t = y_{t-4} + a_t$ with $y_t = a_t$, for $t = 1, 2, 3, 4$.
   (a) Find the variance function for y_t.
   (b) Find the autocorrelation function for y_t.
   (c) Identify the model for y_t as a certain SARIMA model.

$y_t = y_{t-4} + a_t$, $y_t = a_t$ for $t = 1,2,3,4$

a) Variance function of y_t

Consider $t = 4k+r$

$r = 1,2,3,4$ // indicating quarter

$k = 0,1,2,3\ldots$ // indicating year

$E[Y] = 0$

$$\begin{aligned} Y_t &= Y_{t-4} + a_t \\ &= (Y_{t-8} + a_{t-4}) + a_t \\ &= (Y_{t-12} + a_{t-8}) + a_{t-4} + a_t \\ &\vdots \\ &= a_t + a_{t-4} + a_{t-8} + \ldots + a_{r+8} + a_{r+4} + a_r \end{aligned}$$

There are $k+1$ ocurences of a_r in Y_t

The variance is $\text{Var}[Y_t] = (k+1)\sigma_a^2$ //

b) Find the autocorrelation function of y

Let s be $> t$

Consider $s = 4j+i$

$i = 1,2,3,4$ // indicating quarter

$j = 0,1,2,3\ldots$ // indicating year

$\text{Cov}(Y_t, Y_s) = \text{Cov}(a_t + a_{t-4} + a_{t-8} + \ldots + a_{r+8} + a_{r+4} + a_r)$

In the case of $r \neq i$ there won't be overlaps occuring between the two sets of noises a and thus the covariance will be 0.

In the case of $r = i$, then it's when s and t is referring to the same year, same quarter. Thus the noises a's will be identical from r until t

The $\text{cov}(Y_t, Y_s)$ will be $\text{Var}(Y_t)$

Thus $\text{Cov}(Y_t, Y_s) = \begin{cases} (k+1)\sigma_a^2 & ; \text{when } r = i \\ 0 & ; \text{otherwise} \end{cases}$

Because when r is equal to i, it means that $s-t$ is divisible by 4.

The correlation $\text{Corr}(Y_t, Y_s) = \begin{cases} \frac{\sqrt{k+1}}{\sqrt{j+1}} & \text{if } s = 4j+i \text{ and } t = 4k+r \quad \text{where } r = 1,2,3,4 \\ 0 & \text{otherwise} \end{cases}$ //

When t increases, the autocorrelation of the same quarter in consecutive years will have high positive correlation. For different quarters, will be uncorrelated.

c) Identify the model for Y_t as a certain seasonal ARIMA model

Recall that $Y_t - Y_{t-4} = a_t$ is $\text{SARIMA}(0,0,0) \times (0,1,0)_4$ //

So, the yearly seasonal difference is white noise.

TSA HW #07

Name: Dhanabordee Mekintharanggur

Student ID: T11902203

[1]
```
! pip install --upgrade Cython
!pip install --verbose statsmodels==0.13.2
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: Cython in /usr/local/lib/python3.7/dist-packages (0.29.32)
Using pip 21.1.3 from /usr/local/lib/python3.7/dist-packages/pip (python 3.7)
Value for scheme.platlib does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: /usr/local/lib/python3.7/dist-packages
sysconfig: /usr/lib/python3.7/site-packages
Value for scheme.purelib does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: /usr/local/lib/python3.7/dist-packages
sysconfig: /usr/lib/python3.7/site-packages
Value for scheme.headers does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: /usr/local/include/python3.7/UNKNOWN
sysconfig: /usr/include/python3.7m/UNKNOWN
Value for scheme.scripts does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: /usr/local/bin
sysconfig: /usr/bin
Value for scheme.data does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: /usr/local
sysconfig: /usr
Additional context:
user = False
home = None
root = None
prefix = None
Non-user install because site-packages writeable
Created temporary directory: /tmp/pip-ephem-wheel-cache-kem1fo9m
Created temporary directory: /tmp/pip-req-tracker-l3did5jb
Initialized build tracking at /tmp/pip-req-tracker-l3did5jb
Created build tracker: /tmp/pip-req-tracker-l3did5jb
Entered build tracker: /tmp/pip-req-tracker-l3did5jb
Created temporary directory: /tmp/pip-install-9hk06z4m
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
2 location(s) to search for versions of statsmodels:
* https://pypi.org/simple/statsmodels/
* https://us-python.pkg.dev/colab-wheels/public/simple/statsmodels/
Fetching project page and analyzing links: https://pypi.org/simple/statsmodels/
Getting page https://pypi.org/simple/statsmodels/
Found index url https://pypi.org/simple
Looking up "https://pypi.org/simple/statsmodels/" in the cache
Request header has "max_age" as 0, cache bypassed
Starting new HTTPS connection (1): pypi.org:443
https://pypi.org:443 "GET /simple/statsmodels/ HTTP/1.1" 200 47419
Updating cache with response from "https://pypi.org/simple/statsmodels/"
Caching due to etag
  Found link https://files.pythonhosted.org/packages/76/a0/80cc74654f30eb01a95194c6d486f4ce0dcee268c687d571b8aa7984478f/statsmodels-0.4.0.tar.gz#sha256=b11280a773c4ceb95e8ee4fdf32a8d2e3
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/29/7b/86b65e5d2567b8795d2403ff290dc3dd0dc88e41856509129c418f7a06f1/statsmodels-0.4.0.win-amd64
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/d2/fc/d9a766ad568e16be2311ecb239b035e8b50ee3724b7ebc61391f14941ce1/statsmodels-0.4.0.win-amd64
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/a1/c6/2349e2b4a6791a99ae607fba1597be1047cf738e7cbcdb02219d5c4eaedd/statsmodels-0.4.0.win-amd64
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/3e/d9/9d25a99db2a261da2e49cfb6a5ca03310a61310219b28f8612b6fcd582ff/statsmodels-0.4.0.win32-py2
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/e8/da/2c5171f41ec68c34845cbb084071fe88441cf6dbc61f95a7156716a7f132/statsmodels-0.4.0.win32-py2
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/e8/a9/174b29b14ffedb23c54d9038ed3ee58c4f85e8eb64296e7d1ba0f76597b7/statsmodels-0.4.0.win32-py3
  Found link https://files.pythonhosted.org/packages/65/96/6d99d8055dee6aba982775aa577d79db195e26a4ad534482a293440dc95b/statsmodels-0.4.0.zip#sha256=527396d220f84d60e50d1a9b91f3dbe364c8
  Found link https://files.pythonhosted.org/packages/92/df/a6be8bc880acdb2b05c1f8854d1c1885f34f17988b8014c00772cc861b9f/statsmodels-0.4.1.tar.gz#sha256=c1c959edb7314a132b65239b5ece103f0
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/64/cd/4f0d8d7d50c4116503f8cffae15649e6a22b677e477163cefd8481cf8706/statsmodels-0.4.1.win-amd64
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/b5/41/6645e18c904e5c5fd8d10f20a3a2db00e5c8a32fb5ba0aa9b8d0c534f6ac/statsmodels-0.4.1.win-amd64
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/bf/88/80f17534eb38a56bba6e6828f3f4586f889b40a0d5607f13e6c400a42613/statsmodels-0.4.1.win-amd64
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/c6/f3/c4b01588e8c1ffa70b74d1a84a050202c1da6a1a5b834db0d6a9d01f1035/statsmodels-0.4.1.win32-py2
  Skipping link: unsupported archive format: .exe: https://files.pythonhosted.org/packages/f6/27/4980cfa781f1b73912940dff6a9090e9ade8dcd4ea390009248f63996f21/statsmodels-0.4.1.win32-py2
```

[2]
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
import pandas as pd
import math
import calendar
import statsmodels.api as sm
import statsmodels.tsa as smtsa
import statsmodels.graphics as smg
from random import gauss
from random import seed
from pandas import Series
from sklearn import preprocessing
from sklearn.preprocessing import MinMaxScaler
```

[3]
```
# Figure settings
sb.set(rc={'figure.figsize':(14,7)})
sb.set_style({'axes.grid' : True})
```

[5]
```
# read the co2 data and save as a dataframe
data=pd.read_csv('/content/TSA HW07.co2.csv')

# convert months to indices to be grouped as categories
dataMonthConverted = data.copy()
dataMonthConverted['month'] = [list(calendar.month_abbr).index(month) for month in data['month']]

data
```

	time_trend	month	co2_level
0	1994.000000	Jan	363.05
1	1994.083333	Feb	364.18
2	1994.166667	Mar	364.87
3	1994.250000	Apr	364.47
4	1994.333333	May	364.32
...	...	...	...
127	2004.583333	Aug	368.69

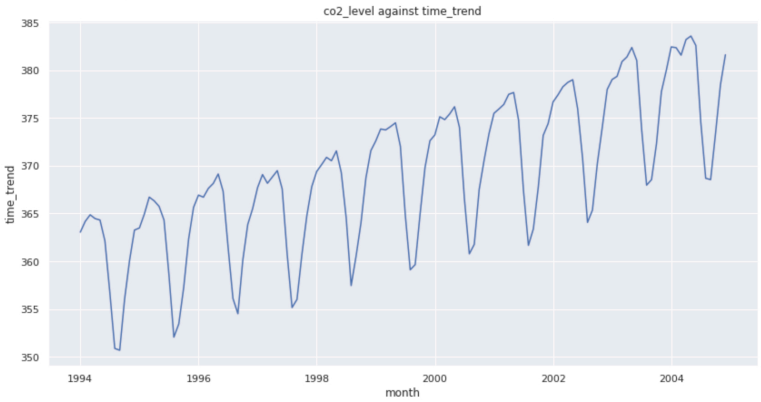
128	2004.666667	Sep	368.55
129	2004.750000	Oct	373.39
130	2004.833333	Nov	378.49
131	2004.916667	Dec	381.62

132 rows × 3 columns

```
[6] t = data['time_trend']
    y = data['co2_level']

    plt.xlabel("month")
    plt.ylabel("time_trend")
    plt.title('co2_level against time_trend')
    sb.lineplot(x = t, y = y, color = "b")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5a50141690>
```



Question 5: (15%) Consider the famous time series data “co2” (monthly carbon dioxide through 11 years in Alert, Canada).

a) Fit a deterministic regression model in terms of months and time. Are the regression coefficients significant? What is the adjusted R-squared? (Note that the month variable should be treated as categorical and transformed into 11 dummy variables.)

Note: the month of March has been dropped to colinearity among variables

```
[52] monthsData = pd.get_dummies(data)
     monthsData.head()
     monthsData = monthsData.drop('month_Mar', axis=1)
     monthsData
```

	time_trend	co2_level	month_Apr	month_Aug	month_Dec	month_Feb	month_Jan	month_Jul	month_Jun	month_May	month_Nov	month_Oct	month_Sep
0	1994.000000	363.05	0	0	0	0	1	0	0	0	0	0	0
1	1994.083333	364.18	0	0	0	1	0	0	0	0	0	0	0
2	1994.166667	364.87	0	0	0	0	0	0	0	0	0	0	0
3	1994.250000	364.47	1	0	0	0	0	0	0	0	0	0	0
4	1994.333333	364.32	0	0	0	0	0	0	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
127	2004.583333	368.69	0	1	0	0	0	0	0	0	0	0	0
128	2004.666667	368.55	0	0	0	0	0	0	0	0	0	0	1
129	2004.750000	373.39	0	0	0	0	0	0	0	0	0	1	0
130	2004.833333	378.49	0	0	0	0	0	0	0	0	1	0	0
131	2004.916667	381.62	0	0	1	0	0	0	0	0	0	0	0

132 rows × 13 columns

```
[54] X = monthsData[[
         'time_trend',
         "month_Jan",
         "month_Feb",
         "month_Apr",
         "month_May",
         "month_Jun",
         "month_Jul",
         "month_Aug",
         "month_Sep",
         "month_Oct",
         "month_Nov",
         "month_Dec"]]
     Y = data[['co2_level']]
     X = sm.add_constant(X)
     model = sm.OLS(Y, X)
     results = model.fit()
     print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              co2_level   R-squared:                       0.990
Model:                            OLS   Adj. R-squared:                  0.989
Method:                 Least Squares   F-statistic:                     997.7
Date:                Wed, 23 Nov 2022   Prob (F-statistic):          2.93e-113
Time:                        02:54:17   Log-Likelihood:                -151.49
No. Observations:                 132   AIC:                             329.0
Df Residuals:                     119   BIC:                             366.4
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
------------------------------------------------------------------------------
const       -3289.5775     44.183    -74.454      0.000   -3377.064   -3202.091
time_trend      1.8321      0.022     82.899      0.000       1.788       1.876
month_Jan      -0.9637      0.342     -2.815      0.006      -1.642      -0.286
month_Feb      -0.2955      0.342     -0.863      0.390      -0.973       0.382
month_Apr       0.2673      0.342      0.781      0.436      -0.411       0.945
month_May       0.5637      0.342      1.646      0.102      -0.114       1.242
month_Jun      -1.6398      0.342     -4.789      0.000      -2.318      -0.962
month_Jul      -8.2489      0.342    -24.088      0.000      -8.927      -7.571
month_Aug     -14.4052      0.342    -42.059      0.000     -15.083     -13.727
month_Sep     -13.7842      0.343    -40.240      0.000     -14.463     -13.106
month_Oct      -9.2242      0.343    -26.923      0.000      -9.903      -8.546
month_Nov      -4.8914      0.343    -14.273      0.000      -5.570      -4.213
month_Dec      -2.3004      0.343     -6.711      0.000      -2.979      -1.622
==============================================================================
Omnibus:                        3.248   Durbin-Watson:                   0.983
Prob(Omnibus):                  0.197   Jarque-Bera (JB):                3.189
Skew:                           0.376   Prob(JB):                        0.203
Kurtosis:                       2.883   Cond. No.                     1.26e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.26e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Are the regression coefficients significant?

The p-value can be usec to determine whether the coefficients are significant. If the p-value is less than the signifiance level then this sample data has enough evidence to reject the null hypothesis. The null hypothesis states that there's no correlation between the dependent variables. By having a p value lower than the significance level, the null hypothesis is rejected, meaning the coefficients from this model are significant.

Most coefficients has p-values are below the significance level of 2.5%. The months of Feb and May has p-values exceeing 2.5% and thus their coefficients are regarded as insignificant.

What is the adjusted R-squared?

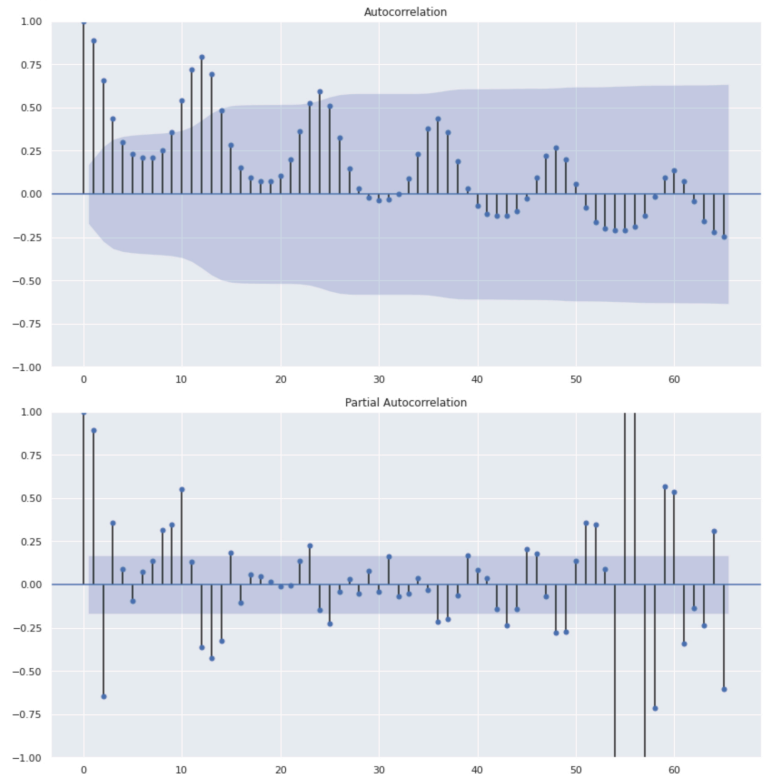
The Adjusted R-squared determines how well the model fits the data (model accuracy) with the percentage of variance in each field.

The Adjusted R-squred found is 0.989, signifying that the model is 98.9% effective.

(b) Identify, estimate the SARIMA model for the co2 level.

```
[55] smg.tsaplots.plot_acf(data['co2_level'], lags = len(data['co2_level'])/2 -1);
     smg.tsaplots.plot_pacf(data['co2_level'], lags = len(data['co2_level'])/2 -1);
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/graphics/tsaplots.py:353: FutureWarning: The default method 'yw' can produce PACF values outside of the [-1,1] interval. After 0.13, the
  FutureWarning,
```



The ACF plot shows a tail-off intraseasonal trend each year.

Additionally, the ACF plot also shows a big exponential interseasonal trend observable from the lower highs.

Note that there since there are 12 months in a year, we can define $s = 12$

This behavior suggests that the model follows $ARIMA(1,0,0)x(1,0,0)_{12}$

(c) Compare the two models above, what do you observe?

```
[56] # create the sampled ARIMA(1,0,0)x(1,0,0)_12 model
     model = smtsa.statespace.sarimax.SARIMAX(Y, exog = X, order = (1,0,0), seasonal_order = (1,0,0,12))
     results = model.fit()
     print(results.summary())
```

```
                                     SARIMAX Results
==========================================================================================
```

```
Dep. Variable:                          co2_level   No. Observations:                  132
Model:             SARIMAX(1, 0, 0)x(1, 0, 0, 12)   Log Likelihood                -130.743
Date:                            Wed, 23 Nov 2022   AIC                            293.485
Time:                                    02:54:19   BIC                            339.610
Sample:                                         0   HQIC                           312.228
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const      -3289.5775     73.782    -44.585      0.000   -3434.187   -3144.968
time_trend     1.8321      0.037     49.639      0.000       1.760       1.904
month_Jan     -0.9637      0.279     -3.454      0.001      -1.511      -0.417
month_Feb     -0.2955      0.217     -1.362      0.173      -0.721       0.130
month_Apr      0.2673      0.269      0.995      0.320      -0.259       0.794
month_May      0.5637      0.324      1.739      0.082      -0.072       1.199
month_Jun     -1.6398      0.293     -5.597      0.000      -2.214      -1.066
month_Jul     -8.2489      0.276    -29.937      0.000      -8.789      -7.709
month_Aug    -14.4052      0.325    -44.366      0.000     -15.042     -13.769
month_Sep    -13.7842      0.303    -45.537      0.000     -14.378     -13.191
month_Oct     -9.2242      0.337    -27.344      0.000      -9.885      -8.563
month_Nov     -4.8914      0.312    -15.663      0.000      -5.503      -4.279
month_Dec     -2.3004      0.288     -7.981      0.000      -2.865      -1.736
ar.L1          0.5078      0.090      5.665      0.000       0.332       0.683
ar.S.L12      -0.1082      0.107     -1.007      0.314      -0.319       0.102
sigma2         0.4337      0.060      7.202      0.000       0.316       0.552
===================================================================================
Ljung-Box (L1) (Q):                   1.26   Jarque-Bera (JB):                 0.96
Prob(Q):                              0.26   Prob(JB):                         0.62
Heteroskedasticity (H):               1.14   Skew:                             0.18
Prob(H) (two-sided):                  0.67   Kurtosis:                         3.20
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
/usr/local/lib/python3.7/dist-packages/statsmodels/base/model.py:606: ConvergenceWarning: Maximum Likelihood optimization failed to converge. Check mle_retvals
  ConvergenceWarning)
```

Discussion for Question 3

Model COmparison

Since the SARIMA model takes into consideration the seasonality of data as oppose to the regression model. The co2 data itself displays a clear seasonal trend. Thus, the $ARIMA(1,0,0)x(1,0,0)_{12}$ should be better at modeling the data.

To confirm this, let's use AIC and BIC values to confirm.

Minimum AIC and BIC values are used as model selection criteria.

Akaike's Information Criteron (AIC) is an information criteria that punishes the number of parameters used.

AIC is defined as: AIC = $-2ln(\hat{L}) + 2r = nln(2\pi) + n * ln(\hat{\sigma_a^2} + n + 2(p+q+1)$

BIC is defined as:

$BIC = k\ln(n) - 2\ln(\widehat{L})$

The smaller the AIC and BIC, the better the model fitting is.

Previously when fitting the data with OLS regression, the AIC and BIC value obtained as

AIC: 329.0

BIC: 366.4

Using the $ARIMA(1,0,0)x(1,0,0)_{12}$ to sample the data resutls in AIC and BIC value as follows

AIC 293.485

BIC 339.610

It can be observed that the $ARIMA(1,0,0)x(1,0,0)_{12}$ has better performance than OLS regression as the AIC and BIC of ARIMA model is lower.

This is possibly due to the big number of parameters (each month, the time_trend, and the co2_level).

Suppose that there is in increase in the parameters required to model the data, then hypothetically, the model should perform worse with higher AIC and BIC.

For example, please refer to the following ARIMA model summary:

```
[57] # create the sampled ARIMA(3,0,0)x(3,0,0)_12 model
     model = smtsa.statespace.sarimax.SARIMAX(Y, exog = X, order = (8,0,0), seasonal_order = (8,0,0,12))
     results = model.fit()
     print(results.summary())
```

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                          co2_level   No. Observations:                  132
Model:             SARIMAX(8, 0, 0)x(8, 0, 0, 12)   Log Likelihood                -132.865
Date:                            Wed, 23 Nov 2022   AIC                            325.730
Time:                                    02:54:43   BIC                            412.214
Sample:                                         0   HQIC                           360.873
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const      -3289.5775    118.331    -27.800      0.000   -3521.503   -3057.652
time_trend     1.8321      0.059     30.963      0.000       1.716       1.948
month_Jan     -0.9637      0.170     -5.672      0.000      -1.297      -0.631
month_Feb     -0.2955      0.186     -1.586      0.113      -0.661       0.070
month_Apr      0.2673      0.193      1.388      0.165      -0.110       0.645
month_May      0.5637      0.186      3.034      0.002       0.200       0.928
month_Jun     -1.6398      0.177     -9.289      0.000      -1.986      -1.294
month_Jul     -8.2489      0.192    -43.052      0.000      -8.624      -7.873
month_Aug    -14.4052      0.220    -65.430      0.000     -14.837     -13.974
month_Sep    -13.7842      0.210    -65.782      0.000     -14.195     -13.374
month_Oct     -9.2242      0.212    -43.603      0.000      -9.639      -8.810
month_Nov     -4.8914      0.206    -23.789      0.000      -5.294      -4.488
month_Dec     -2.3004      0.215    -10.677      0.000      -2.723      -1.878
ar.L1          0.3462      0.152      2.270      0.023       0.047       0.645
ar.L2          0.2254      0.151      1.498      0.134      -0.070       0.520
ar.L3          0.0527      0.169      0.312      0.755      -0.278       0.384
ar.L4          0.0386      0.183      0.211      0.833      -0.320       0.398
ar.L5          0.0551      0.182      0.303      0.762      -0.301       0.411
ar.L6          0.0330      0.208      0.159      0.874      -0.375       0.441
ar.L7         -0.0447      0.168     -0.265      0.791      -0.375       0.286
ar.L8         -0.0569      0.155     -0.366      0.714      -0.361       0.247
ar.S.L12      -0.1058      0.232     -0.457      0.648      -0.560       0.348
ar.S.L24      -0.2493      0.246     -1.014      0.310      -0.731       0.232
ar.S.L36      -0.3842      0.220     -1.749      0.080      -0.815       0.046
ar.S.L48      -0.2211      0.267     -0.827      0.408      -0.745       0.303
ar.S.L60      -0.0568      0.261     -0.218      0.828      -0.455       0.568
```

```
ar.S.L72      -0.5641      0.217     -2.594      0.009      -0.990      -0.138
ar.S.L84       0.0631      0.323      0.196      0.845      -0.569       0.695
ar.S.L96      -0.1130      0.281     -0.402      0.688      -0.664       0.438
sigma2         0.4067      0.127      3.201      0.001       0.158       0.656
===================================================================================
Ljung-Box (L1) (Q):                   0.71   Jarque-Bera (JB):                 2.96
Prob(Q):                              0.40   Prob(JB):                         0.23
Heteroskedasticity (H):               1.66   Skew:                             0.33
Prob(H) (two-sided):                  0.10   Kurtosis:                         3.31
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

With the $ARIMA(8,0,0)x(8,0,0)_{12}$

AIC 325.730

BIC 412.214

The AIC is now almost equivalent to the one obtained with the OLS regression, and the BIC is not significantly higher.

Moreover, some coefficients has p-values that lie outside the significant level and fails to reject the null hypothesis, further confirming the negative effects of increasing the number of parameters on the model performance.

Conclusion

1. Juding from the characteristics of the ACF plot, the SARIMA model can be estimated.
2. The SARIMA model $ARIMA(1,0,0)x(1,0,0)_{12}$ is able to perform better than the OLS regression model in fitting the data judging from AIC and BIC.
3. As the number of parameters increase, the model performance decreases.
4. The $ARIMA(1,0,0)x(1,0,0)_{12}$ model takes into consideration the seasonality of the data, and thus is a better fit for the CO2 data provided. This is supported by the lower AIC and BIC values compared to the OLS regression.