ACF and PACF

31 August 2024 14:11

$Y_t = f(Y_{t-1}, Y_{t-2}, \ldots)$

$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3}$ } Auto regression

$t \rightarrow$ Current time point

Assumption: The value in the current time point is used to predict the value in next time point

Auto-Correlation function

$ACF(1) = Corr(y_{t+1}, y_t) = Corr(Y_t, Y_{t-1})$

$ACF(k) = Corr(Y_{t+k}, Y_t) = Corr(Y_t, Y_{t-k})$

* ACF → Determines average correlation b/w TS observation and its past values

PACF → $Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3}$

$Corr(Y_t, Y_{t-3} \mid Y_{t-1}, Y_{t-2})$

} Partial Auto Correlation Function

* When the PAC dia or it = 0, the time point is considered and the other lags are eliminated.

* Latter order AC's are less reliable

→ Avoid using PACF for seasonal trend analysis

→ Seasonal lags can be used for prediction

Seasonality and Stationarity

31 August 2024 14:49

1) The properties of the series do not move → stationary
2) It has pattern whereas that does not depend on time

Y_t : $E(Y_t)$ does not depend on t

$Var(Y_t)$ does not depend on t

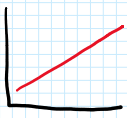
$Corr(Y_t, Y_{t-s})$ does not depend on t

$E(Y_t)$ → Expectation

Stationarity

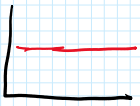
$Y_t = a + bt + ct^2$ Y_t

$d = 0$, Y_t



$\frac{d}{dt}(Y_t) = b + 2ct$ $Y_t - Y_{t-1}$

$d = 1$, first difference



$\frac{d^2}{dt^2}(Y_t) = 2c$ $(Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$

$d = 2$, Second difference

Differencing:

$Y_t = a + bt$

$Y_{t-1} = a + b(t-1)$

$Y_t - Y_{t-1} = b$ → first difference

It is a process of transforming a ts into stationary

Seasonal differencing → Y_{t-12} [Last year same day]

Differencing / Seasonal Differencing

$Y_t = \phi Y_{t-1} + \varepsilon_t$

$H_0 = \phi = 1$ ie $Y_t = Y_{t-1} + \varepsilon_t \rightarrow Y_t - Y_t = \varepsilon_t$ [~~Stationary~~]

$H_a \neq 1$ ie [~~Non stationary~~]

Dicky fuller Test (Augmented Dicky Fuller Test)

Time Series Processing

31 August 2024 16:28

Autoregressive process (p) or AR(p)

$$Y_t = \varphi Y_{t-1} + \varphi_2 Y_{t-2} + \dots \varphi_p Y_{t-p} + \varepsilon_t$$

Identification of AR(1) model or AR(1) Process

① PACF (K) ≈ 0 $K > p$

$$\hat{Y}_t = \hat{\phi}_1 Y_{t-1} + \hat{\phi}_2 Y_{t-2} + \hat{\phi}_3 Y_{t-3} \dots \hat{\phi}_p Y_{t-p}$$ } Forecasting Auto Regressive processes

$\hat{\varepsilon}_t = Y_t - \hat{Y}_t$: forcast error at t

Moving Average Process (q)

$$Y_t = \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_3 \varepsilon_{t-3} \dots \theta_q \varepsilon_{t-q}$$

Identification of MA(q)

↳ ACF (k) ≈ 0 $K > q$

Auto Regression Moving Average (p, q)

ex: ARMA (2,0) → Auto Regression process of order 2
AR(2)

ex: ARMA (0,1) → Moving Average process of order 1
MA(1)

Mean Reversion, The model tends to return to mean.

Stationary Data

AR (P)

MA (q)

ARMA (P,q)

SARIMA (P,d,Q), (P,D,Q)

S → Seasonal

P, D, Q is same as p,d,q whereas the seasonal element is added

TREND

ARIMA (p,r,q)

Seasonal

SARIMA (P, d, q), (P,D, Q)

Program exp 1

01 September 2024 15:30

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                      CO2 ppm   No. Observations:                  168
Model:                 ARIMA(2, 1, 1)   Log Likelihood                -162.164
Date:                Sun, 01 Sep 2024   AIC                            332.328
Time:                        15:27:42   BIC                            344.800
Sample:                    01-01-1965   HQIC                           337.391
                         - 12-01-1978
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.5317      0.050     30.597      0.000       1.434       1.630
ar.L2         -0.8284      0.051    -16.391      0.000      -0.927      -0.729
ma.L1         -0.8233      0.063    -13.088      0.000      -0.947      -0.700
sigma2         0.4028      0.042      9.679      0.000       0.321       0.484
===================================================================================
Ljung-Box (L1) (Q):                   1.52   Jarque-Bera (JB):                 1.92
Prob(Q):                              0.22   Prob(JB):                         0.38
Heteroskedasticity (H):               0.68   Skew:                             0.23
Prob(H) (two-sided):                  0.15   Kurtosis:                         3.24
===================================================================================

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

Forecast

$$Y_t = 1.53 Y_{t-1} - 0.83 Y_{t-2} - 0.82\ \varepsilon_{t-1} + \varepsilon$$

In [71]:
```
ARIMA_pred=ARIMA_predictions.cumsum()
#ARIMA_pred
```

In [72]:
```
plt.plot(train,label='Training Data')
```

$$P_t = (P_t - P_{t-1}) + (P_{t-1} - P_{t-2}) + (P_{t-2} - P_{t-3})$$ } Cumulative sum

$$= Y_t + Y_{t-1} + Y_{t-2}$$

Program exp 2

01 September 2024 18:19

2 log likelihood → SSE + 2 coefficients

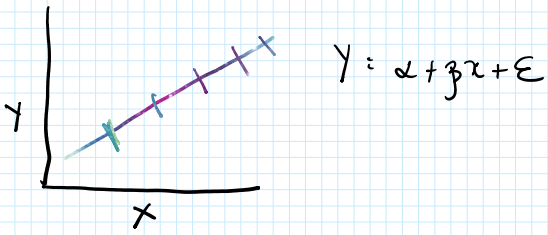
likelihood : $\prod_{i=1}^{n} f(y_i)$

$= \prod \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - (\alpha+\beta x_i)}{\sigma}\right)^2}$

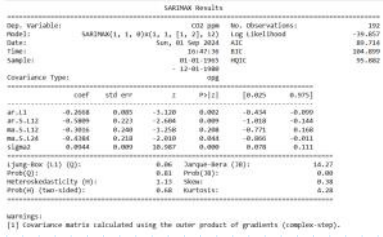
$= \prod \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{\sum_{i=1}^{n} y_i (\alpha+\beta x_i)^2}{\sigma^2}}$

: −2 log likelihood

$\frac{1}{\sigma^2}\sum\left(y-(\alpha+\beta x)^2\right)$



$Y = \alpha + \beta x + \varepsilon$



$D=1, d=1$

$y_t = (P_t - P_{t-1}) - (P_{t-12} - P_{t-13})$

$(P_t - P_{t+12}) - (P_{t-1} - P_{t-13})$

$Y_t = -0.26 Y_{t-1} - 0.69 Y_{t-12} - 0.37 \varepsilon_{t-12} - 0.67 \varepsilon_{t-24} + \varepsilon_t$

$P_t = P_{12} - P_{t-1} + P_{t-13} - 0.26\left(P_{t-1} - P_{t-15} - P_{t-2} + P_{t-14}\right) - 0.67\left(P_{t-12} - P_{t-24} - P_{t-3} + P_{t-25}\right)$

Key point → Even for the statistically insignificant coefficient can perform well in the model.

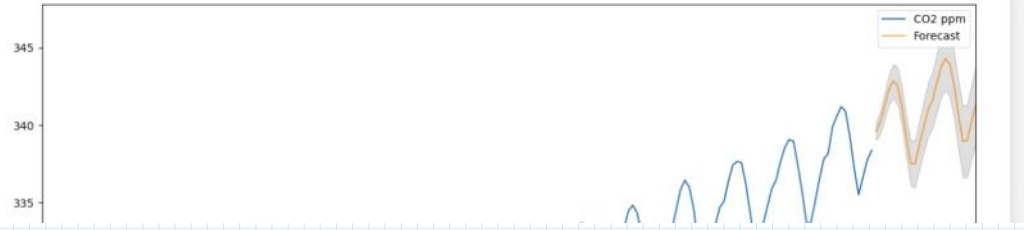
AIC = SSE + 2(HYPER PARAMETERS)

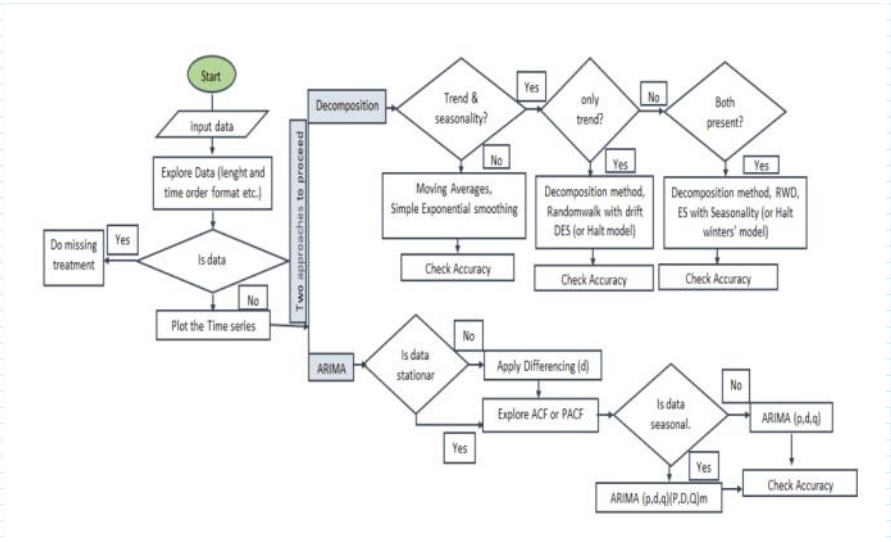
```
1982-11-01    340.205074
1982-12-01    341.323511
Freq: MS, Name: predicted_mean, dtype: float64

In [68]: pred95 = model_Sarima.get_forecast(steps=24)

In [69]: pred95=pred95.conf_int()

In [70]: pred95
```

The Confidence Interval gives the range for the predicted values

	lower CO2 ppm	upper CO2 ppm
1981-01-01	338.986512	340.191575
1981-02-01	339.420455	340.914739
1981-03-01	340.337395	342.118382
1981-04-01	341.302874	343.319650
1981-05-01	341.731837	343.962170
1981-06-01	341.294832	343.719353
1981-07-01	339.828506	342.432933
1981-08-01	337.822521	340.595165
1981-09-01	336.051761	338.982996
1981-10-01	335.945486	339.027157
1981-11-01	337.104292	340.329391
1981-12-01	338.214169	341.576583
1982-01-01	339.227049	342.762972
1982-02-01	339.748205	343.438420

```
]: axis = df.plot(label='Observed', figsize=(15, 8))
forecast.plot(ax=axis, label='Forecast', alpha=0.7)
axis.fill_between(forecast.index, pred95['lower CO2 ppm'], pred95['upper CO2 ppm'], color='k', alpha=.15)
axis.set_xlabel('Year-Months')
axis.set_ylabel('CO2 ppm')
plt.legend(loc='best')
plt.show()
```




Start
input data
Explore Data (lenght and time order format etc.)
Is data
Yes
Do missing treatment
No
Plot the Time series
Two approaches to proceed
Decomposition
Trend & seasonality?
Yes
No
Moving Averages, Simple Exponential smoothing
Check Accuracy
only trend?
No
Yes
Decomposition method, Randomwalk with drift DES (or Halt model)
Check Accuracy
Both present?
Yes
Decomposition method, RWD, ES with Seasonality (or Halt winters' model)
Check Accuracy
ARIMA
Is data stationar
No
Yes
Apply Differencing (d)
Explore ACF or PACF
Is data seasonal.
No
ARIMA (p,d,q)
Yes
ARIMA (p,d,q)(P,D,Q)m
Check Accuracy

Reference

01 September 2024 14:24

Tufte

ATSA

Forcasting Principles and Practice

Forecasting: Principles and Practice (3rd ed) (otexts.com)

→ Syllabus book for TSF

Programming lang - R, Python.

Book of Y by Judea Perl [Lauret in Computer science

Time series analysis by James Hamelton.

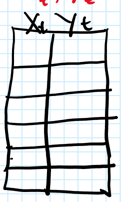
Exams → 1 to 4

VAR Vector Auto Regressive model

21 September 2024 14:15

time - t

X_t, Y_t



$$X_t = \phi_{11} X_{t-1} + \phi_{12} Y_{t-1} + \delta_t$$

$$Y_t = \phi_{21} X_{t-1} + \phi_{22} Y_{t-1} + \varepsilon_t$$

Granger Causality:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} \delta_t \\ \varepsilon_t \end{bmatrix}$$

Z_t ϕ_t Z_{t-1} E_t

$$Z_t = \phi_t Z_{t-1} + \phi_2 Z_{t-2} + E_t$$

Tensor:

tensorflow: datastructure that has the flow of the numbers in the neural network

Cross Correlation: Correlation b/w one variable and the past/lag of the other variable

$$Y_t = c + \phi Y_{t-1}$$

$$Y_{t-1} = c + \phi Y_{t-2}$$

$$Y_t - Y_{t-1} = \phi (Y_{t-1} - Y_{t-2})$$

$$\nabla t = \phi \nabla t_{t-1}$$

→ The P value must be < 0.05 level

→ confidence

→ The intercept is biggthence largest store

for order = (7,0)

Results for equation Store_1041

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	758.3079	79.915	9.489	0.000	601.678	914.938
L1.Store_1044	-0.8404	0.138	-6.104	0.000	-1.110	-0.571
L1.Store_1041	0.5908	0.090	6.540	0.000	0.414	0.768
L2.Store_1044	0.6229	0.136	4.596	0.000	0.357	0.889
L2.Store_1041	-0.5778	0.090	-6.398	0.000	-0.755	-0.401
L3.Store_1044	-0.2683	0.139	-1.926	0.054	-0.541	0.005
L3.Store_1041	0.2115	0.094	2.244	0.025	0.027	0.396
L4.Store_1044	0.4057	0.156	2.600	0.009	0.100	0.712
L4.Store_1041	-0.3745	0.106	-3.522	0.000	-0.583	-0.166
L5.Store_1044	-0.3819	0.136	-2.814	0.005	-0.648	-0.116

The session 5 python notebook

Results for equation Store_1041

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	758.3079	79.915	9.489	0.000	601.678	914.938
L1.Store_1044	-0.8404	0.138	-6.104	0.000	-1.110	-0.571
L1.Store_1041	0.5908	0.090	6.540	0.000	0.414	0.768
L2.Store_1044	0.6229	0.136	4.596	0.000	0.357	0.889
L2.Store_1041	-0.5778	0.090	-6.398	0.000	-0.755	-0.401
L3.Store_1044	-0.2683	0.139	-1.926	0.054	-0.541	0.005
L3.Store_1041	0.2115	0.094	2.244	0.025	0.027	0.396
L4.Store_1044	0.4057	0.156	2.600	0.009	0.100	0.712
L4.Store_1041	-0.3745	0.106	-3.522	0.000	-0.583	-0.166
L5.Store_1044	-0.3819	0.136	-2.814	0.005	-0.648	-0.116
L5.Store_1041	0.1913	0.091	2.097	0.036	0.013	0.370

$$A_t = \phi_{11} A_{t-1} + \phi_{12} S_{t-1}$$

$$S_t = \phi_{21} A_{t-1} + \phi_{22} S_{t-1}$$

A = advertising

S - Sales

High advertising at t_{-1} → High sales at t

low sales at t_{-1} → low adv at t

} example

```
import math
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(test['Store_1044'],pred['Store_1044'])
rmse = math. sqrt(mse)
print('Store_1044:',mse)
print('Store_1044:',rmse)
```

C = 471 ratio = 0.57

```
Store_1044: 73057.78180640825
Store_1044: 270.2920306009932
```

```
## Calculating the RMSE for Store 1041
mse = mean_squared_error(test['Store_1041'],pred['Store_1041'])
rmse = math. sqrt(mse)
print('Store_1041:',mse)
print('Store_1041:',rmse)
```

C = 723 ratio = 0.51

```
Store_1041: 139450.76236055826
Store_1041: 373.4310677495356
```

VARMAX

21 September 2024 16:31

Y_t → non stationary

$Y_t - Y_{t-1}$ is stationary

Common trend, Co-integration:

∴ Spurious Regression / false regression

ex population and unemployment

There is no correlation b/w them whereas these variables are integrated

Case Study

22 September 2024 14:17

model → fit → Forecast

model : SARIMAX
Exponential smoothing → hardcode parameters or use optimize = true
Prophet

Generate ACF
PACF
Calculate RMSE, MAP
} main questions

1) Covid data

Target variable : Hospitalized / hospitalization

① Converting data into timeseries (do not use parsing, as there would be missing values for the period)
② Box plot → The outliers is the trend
③ Boxplots for the months gives the trend
④ decomposition series to check seasonality
⑥ check for stationarity
⑦ Rolling standard deviation ($\bar{X}$-S charts or control chart Mean · standard deviation
$\bar{X}$-r chart or control chart Mean · range)
⑧ Statistical test for stationarity (adfuller test)
⑨ ACF and PACF → [determine the periodicity and then build the decomposition series]
⑩ Model selection and Hyperparameters
⑪ Determine the equation
⑫ Predict the series
⑬ Incorporate seasonality to improve the model [SARIMA model]
⑭ RMSE calculation MAPE calculation
⑮ Calculate AIC
⑯ Select the P, D, Q, p, d, q and plot SARIMAX

$$Y_t = (H_t - H_{t-1}) - (H_{t-7} - H_{t-8})$$

⑰ Bumps in the SARIMA is due to weekly effect
⑱ RMSE and MAPE for the ARIMA and SARIMA
⑲ Residual analysis → Q-Q plot, correlogram } → Model assessment
⑳ forecasing hospitalization based on another variable ['Positive' or 'death']
㉑ Use SARIMAX
㉒ Model predictions
㉓ Plotting the residuals

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y})^2}$$

$$MAPE = \frac{1}{n} \cdot \sum \frac{|y - \hat{y}|}{y}$$

Note : The variable could be either exogenous or endogenous
exogenous → independent
endogenous → dependent