Week 2 Project

Jason Feng

Problem 1

Compare the conditional distribution of the Multivariate Normal, to the OLS equations. Are these values the same? Why? Use the data in problem1.csv to prove your answer empirically.

Answer

After compare the conditional distribution of the Multivariate Normal and the OLS equations, I think the values are **Same**.

I think there are two reasons that the values are same. First the test sample the size is large enough to get the result that close to correct answers by both methods.

The second reason is when I calculate the OLS equation, I find the equation have some similar part compare with the equation for conditional distribution. I also find I subtract the y_mean for y and x_mean for x which can make my calculation more close to accurate.

Here is the prove by **math** function:

For conditional Variance:

$E[(Y-f(x))^2] = E[(Y - E(Y|X) + E(Y|X) - f(X))^2]$

$= E[E\{(Y - E(Y|X) + E(Y|X) - f(X))^2|X\}]$

$= E[Var(Y|X)] + E[(E(Y|X) - f(X))^2]$

For $f(X) = E(Y|X)$ the second term becomes zero

$E[(Y-f(x))^2] = E[Var(Y|X)]$

$Y = E[Y|X] + (Y - E(Y|X))]$

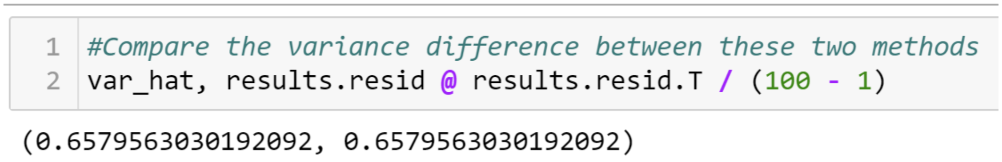
$= E(Y|X) + E(error|X)$ → error to 0

$= E(Y|X)$

When I used problem1.csv data to calculate the variance hat and mean hat for conditional distribution of the Multivariate Normal and the residuals and predict variance and mean by OLS equation, the result are **Same**

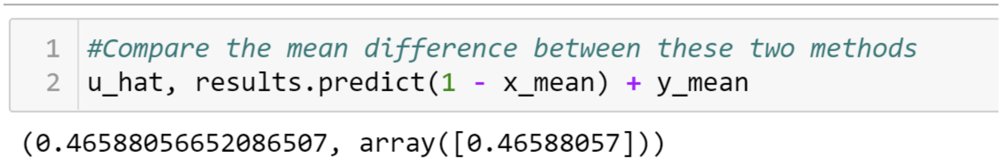
Here is the results:

	Conditional Distribution	OLS
Variance	0.6579563030192092	0.6579563030192092
Mean	0.46588056652086507	0.46588057



```
#Compare the variance difference between these two methods
var_hat, results.resid @ results.resid.T / (100 - 1)
```

```
(0.6579563030192092, 0.6579563030192092)
```



```
#Compare the mean difference between these two methods
u_hat, results.predict(1 - x_mean) + y_mean
```

```
(0.46588056652086507, array([0.46588057]))
```

I think according to our data test, it can prove my thought that the value for these two methods are same.

Problem 2

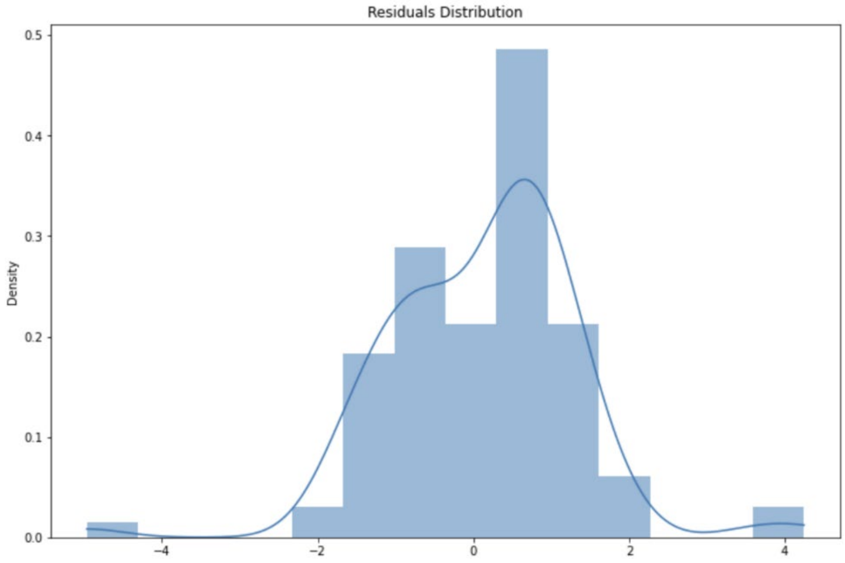
Fit the data in problem2.csv using OLS and calculate the error vector. Look at it's distribution. How well does it fit the assumption of normally distributed errors?

Answer

After fit the data using OLS and calculate the error vector, I find the distribution it **not fit** the assumption of normally distributed errors. In other words, the error distribution is not normally distributed.

Here is the result:

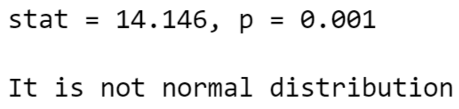
First I draw a histplot to show the residuals distribution:



However, I cannot determine the residuals is normal distribution or not. So I do two normal distribution tests. These two tests show the residuals is **not normal distribution**.

Here is the test result:

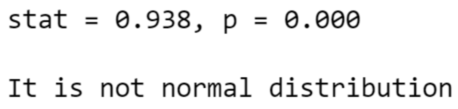
Normal test:



```
stat = 14.146, p = 0.001

It is not normal distribution
```

Shapiro test:



```
stat = 0.938, p = 0.000

It is not normal distribution
```

Fit the data using MLE given the assumption of normality. Then fit the MLE using the assumption of a T distribution of the errors. Which is the best fit?

Answer

After fit the data using MLE given the assumption of normality and T distribution, I think **T distribution** best fit.

Here is the result for AIC for both assumptions:

	Normality	T distribution
AIC	325.98419338057477	317.1186022971749

I find the T distribution assumption has lower AIC; so, I think T distribution assumption best fit.

What are the fitted parameters of each and how do they compare? What does this tell us about the breaking of the normality assumption in regards to expected values in this case?

Answer

I compare the standard deviation for each method and I think this it the fitted parameters. I find they are very close for both method for mean and standard deviation.

	Residuals	Normality	T distribution
SD	1.198394128	1.19839547	0.9735875321845389

This result tells me breaking of the normality assumption in regards to expected values will not influence our prediction a lot.

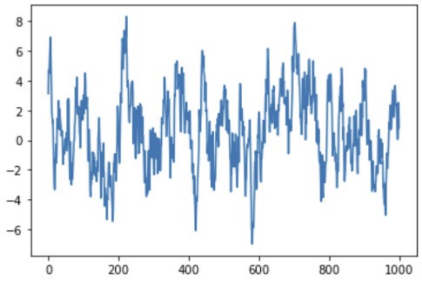
Problem 3

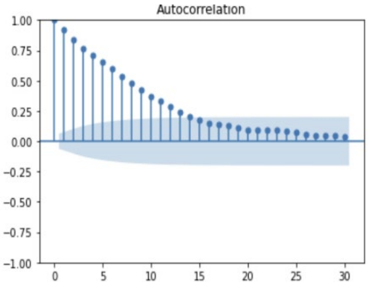
Simulate AR(1) through AR(3) and MA(1) through MA(3) processes. Compare their ACF and PACF graphs. How do the graphs help us to identify the type and order of each process?

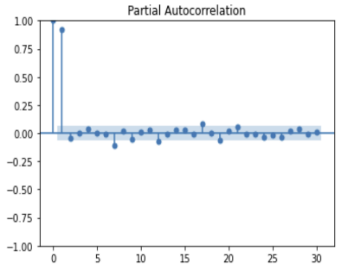
Answer

Here is the ACF and PACF graph for AR(1) through AR(3)

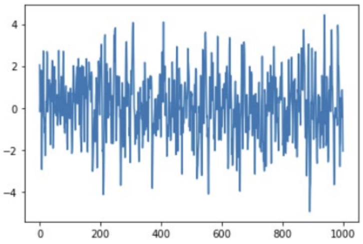
AR(1) ACF PACF

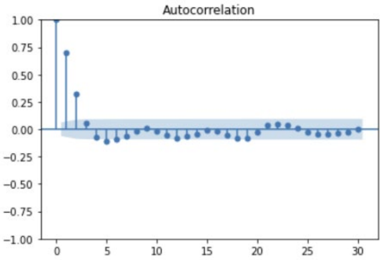


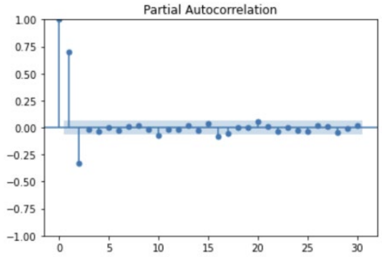




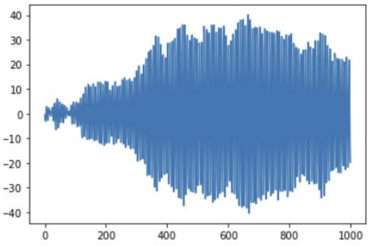
AR(2) ACF PACF

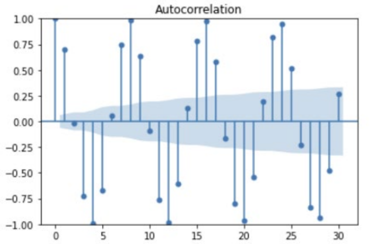


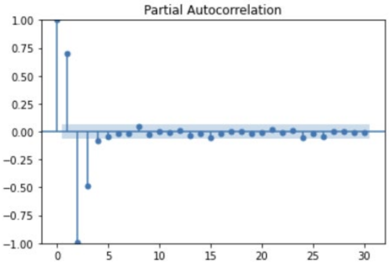




AR(3) ACF PACF

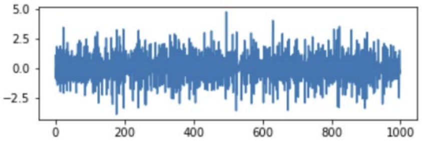




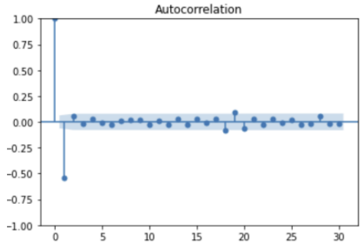


Here is the ACF and PACF graph for MA(1) through MA(3)

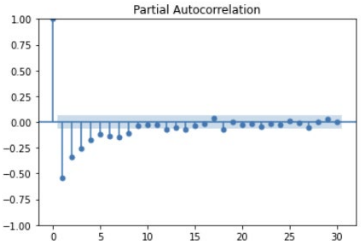
MA(1)



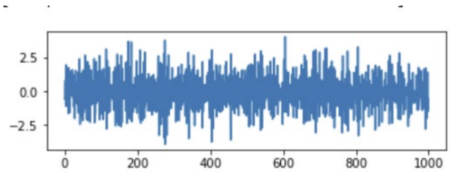
ACF



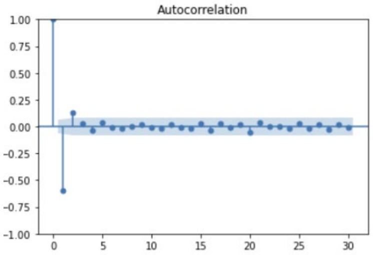
PACF



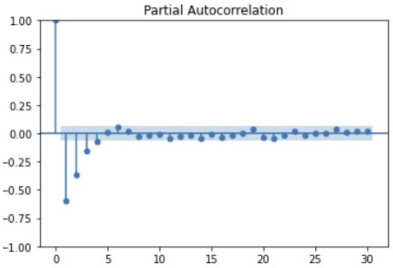
MA(2)



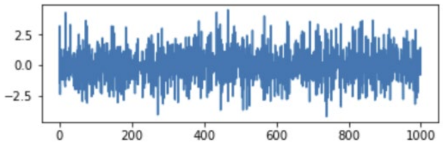
ACF



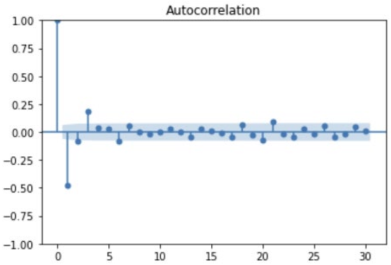
PACF



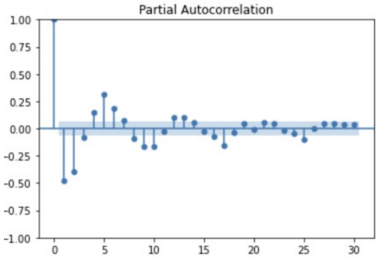
MA(3)



ACF



PACF



How do the graphs help us to identify the type and order of each process?

Answer

I think the ACF and PACF plot can very useful for us to identify the type and order of each process.

Identify the type:

1. The ACF for AR and MA are very different. The ACF for AR has many long line that excess the shadow; however, the ACF for MA only have few.

2. The PACF for AR and MA are very different either. It opposite compare to ACF. The PACF for MA has many long line that excess the shadow; however, the PACF for AR only have few.

Order:

AR:

ACF: ACF for AR order is very different. For AR(1), almost all line are same direction. For AR(2), many of the line are different direction; however, there is one direction are short. For AR(3), the line from two direction are close to equal and both have short and long line.

PACF: PACF for AR order is hard to find different. For my perspective, I can use how many long line to determine. For example, there two long line for AR(1), 3 long line for AR(2), 4 long line for AR(3)

MA:

ACF: ACF for MA order is hard to find different. For my perspective, I can use how many long line (Excess the shadow) to determine. For example, there two long line for MA(1), 3 long line for MA(2), 4 long line for MA(3)

PACF: PACF for MA order is very different. For MA(1), Almost all line are same direction. For MA(2), there are some short line are different direction. For MA(3), there are some long line and short line are different direction.