# Project Week02

Hongren Jiang

January 2022

# 1 Problem 1

## 1.1 Question

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.

## 1.2 Answer

The following are the conditional expected value and variance for the multivariate distribution:

$$E[P(X_1|X_2 = x_2)] = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2)$$

$$Var[P(X_1|X_2 = x_2] = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}$$

The following are the expected value and variance for the OLS regression:

$$E[Y|X = x] = E[\beta X + c + e|X = x] = \beta x + c$$

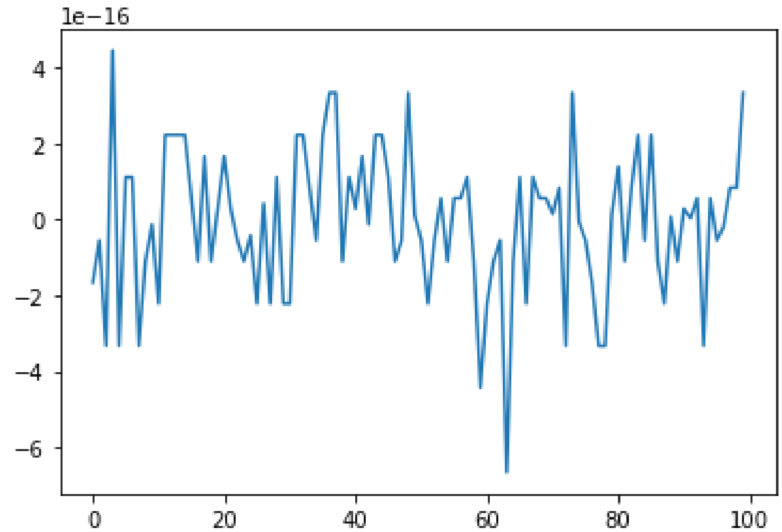
$$Var[Y|X = x] = Var[\beta X + c + e|X = x] = Var[e]$$

Therefore, we only need to calculate the above values and compare them. Through the algorithm provided by problem1.ipynb, we can get:

$$Var[P(X_1|X_2 = x_2)] \approx 0.657956$$

$$Var[Y|X = x] \approx 0.651377$$

They are almost the same.
We use images to visually represent the difference between two expectation vectors




As you can see, the difference between the two is very small. Therefore, they can be considered equal.

# 2 Problem 2

## 2.1 Question

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?

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?

What are the fitted parameters of each 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?

## 2.2 Answer

The following are the OLS equation and fitting result:

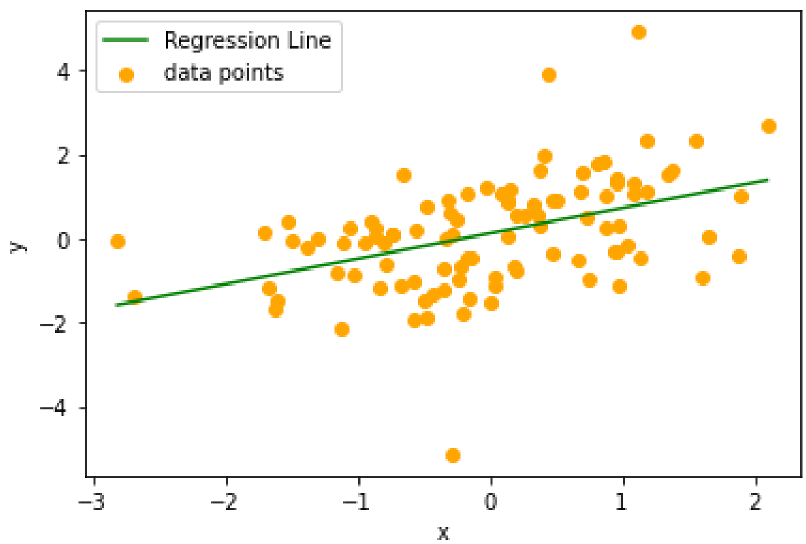
$$\hat{y} = \hat{\beta}x + c$$

$$\hat{\beta} \approx 0.605205$$

$$c \approx 0.119836$$

$$\hat{y} = 0.605205x + 0.119836$$

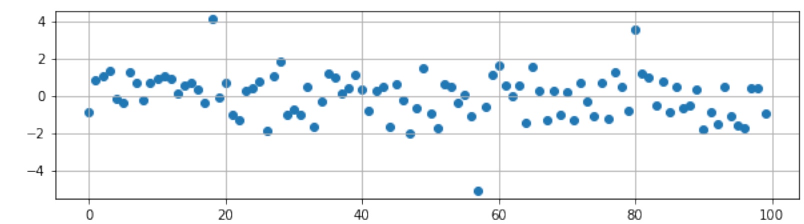
Here are the picture of regression line and data point:

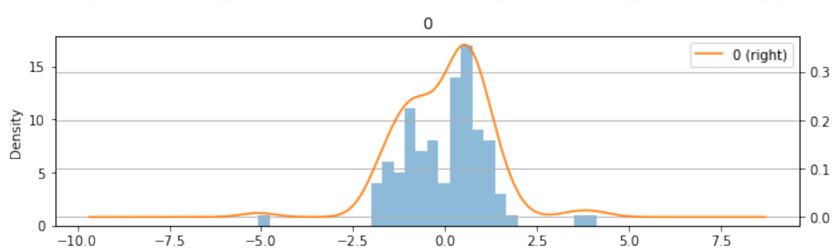



The residual equation is as follows

$$error = y - \hat{y}$$

Here are the graphs:






They do not look like they follow a normal distribution. Therefore, I do the K-S test:

$$Pvalue \approx 0 < 0.05$$

The errors do not follow a normal distribution. Then, I use MLE method estimate the regression parameters, and get the following answers: Based on the assumption of normality:

$$\hat{y} = 0.605190x + 0.119836$$

Based on the assumption of t-distribution:

$$\hat{y} = 0.594180x + 0.119835$$

These results are very similar.

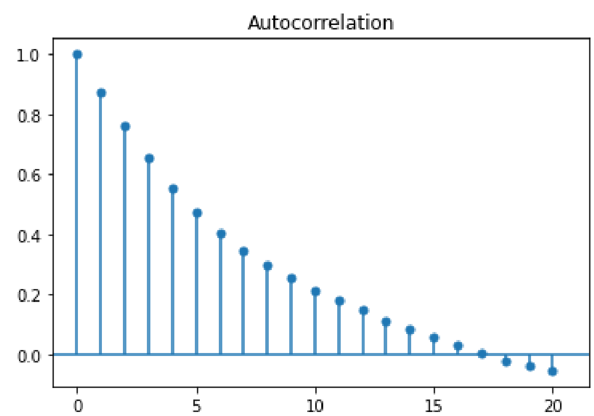
# 3 Problem 3

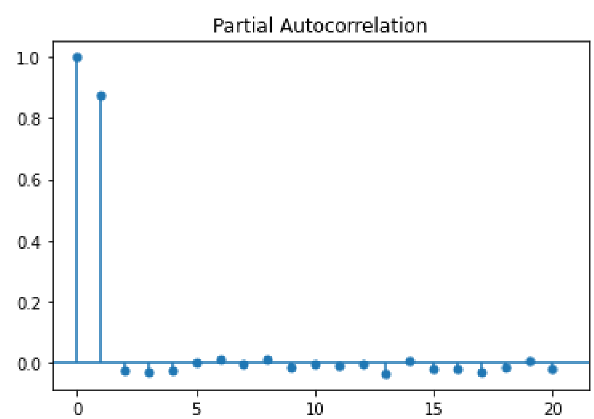
## 3.1 Question

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?

## 3.2 Answer

In this section I first conclude that the ACF (absolute value) of the AR model decreases gently, while the PACF (absolute value) decreases abruptly at the lag order. The MA model has the opposite pattern. The following are graphs of AR(1), and you can find graphs of other AR models in Jupyter Notebook files.



The following are graphs of MA(2), and you can find graphs of other MA models in Jupyter Notebook files.