WEEK2 PROJECT

Peiyi Xu

2022.1.14

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.

Solution 1

Case 1 Conditional Distribution of the Multivariate Normal

- We assume that the (x, y) datapoints in problem1.csv are independent variables under the same bivariate normal distribution.

$$\mu = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x)$$

$$\sigma = \sigma_y - \frac{\sigma_{xy}^2}{\sigma_x}$$

Case 2 OLS

- We assume that y can be explained by x using regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  and all assumptions of OLS are satisfied.

- Then

$$E(y|x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$Var(y|x) = Var(\beta_1 x + \epsilon|x) = Var(\epsilon)$$

Conclusion: Same mean, Different variance

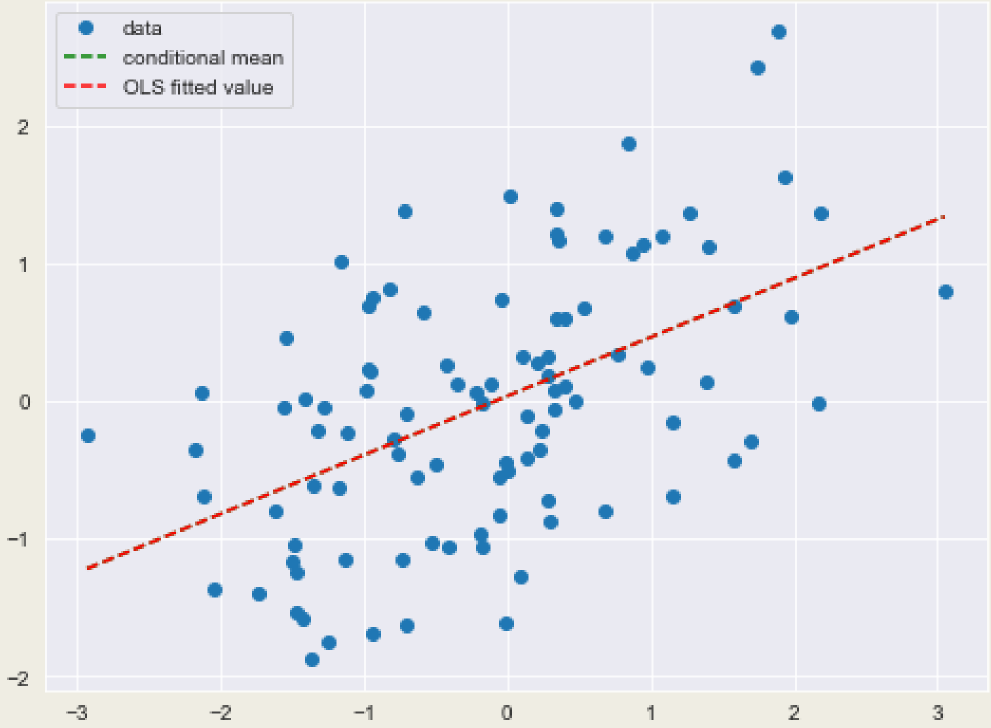


Figure 1. fitted y value given x as condition

sigma_1 = 0.657, sigma_2 = 0.664

Problem 2

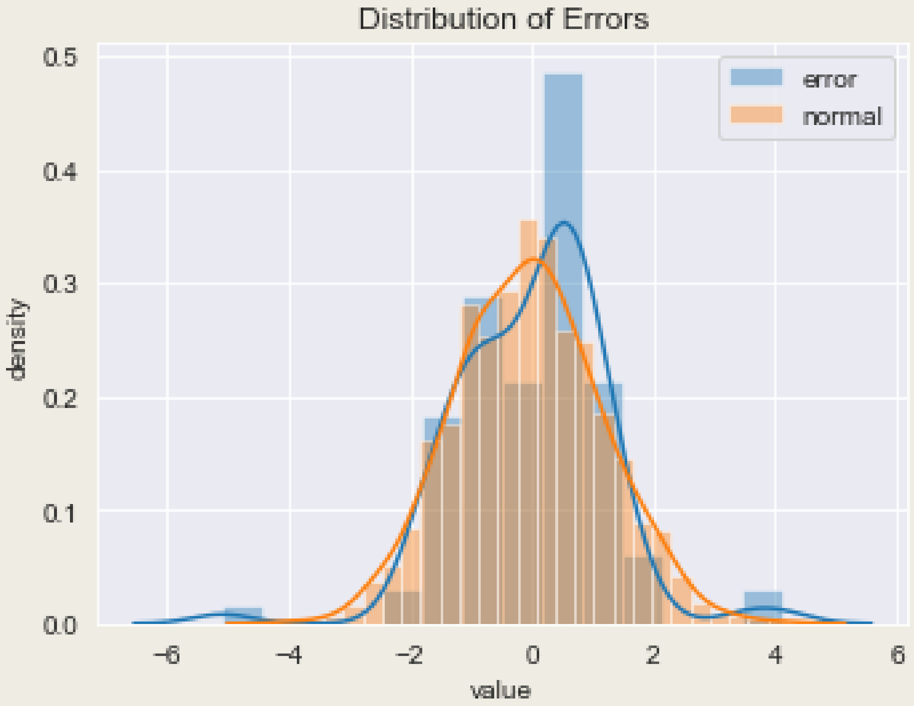
- Fit the data in problem2.csv using OLS and calculate the error vector. Look at it's distribution. How well does it fit the assumption of normally distributed errors?

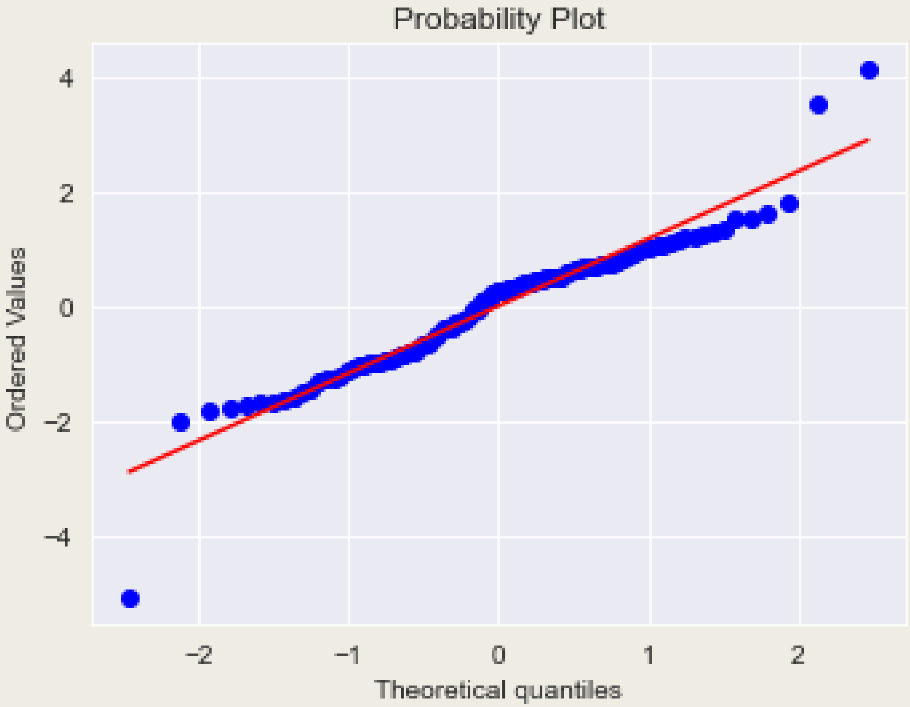
- Fit the data using MLE given the assumption of normality. Then fit the MLE using the assumption of a T distribution of the errors. Which is the best fit?

- What are the fitted parameters of each 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?

Solution2

1. Distribution of Error





Sharpiro-Wilk Test

- Sharpiro-Wilk test examines whether a variable is normally distributed.
- Null Hypothesis: errors are normally distributed, $\alpha = 0.5$
- In our case, p-value of Sharpiro-Wilk test is 0.00015
- The result suggests that errors may not be normally distributed.

2. Fit the data using MLE

For regression $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ we assume errors follow:

Normal Distribution (a better fit)

- $\beta_0 = 0.1198$
- $\beta_1 = 0.6051$

- $SSE = 143.61$
- $AIC = 325.98$
- $BIC = 333.79$

T Distribution

- $\beta_0 = 0.1232$
- $\beta_1 = 0.5951$

- $SSE = 143.62$
- $AIC = 318.94$
- $BIC = 329.37$

Indications

- Though errors may not be normally distributed, the normal distribution is still a better choice than t distribution

- Breaking of the normality assumption will result in different estimated values of parameters and thus a difference in the expected value of y

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?

Solution 3

1. AR processes

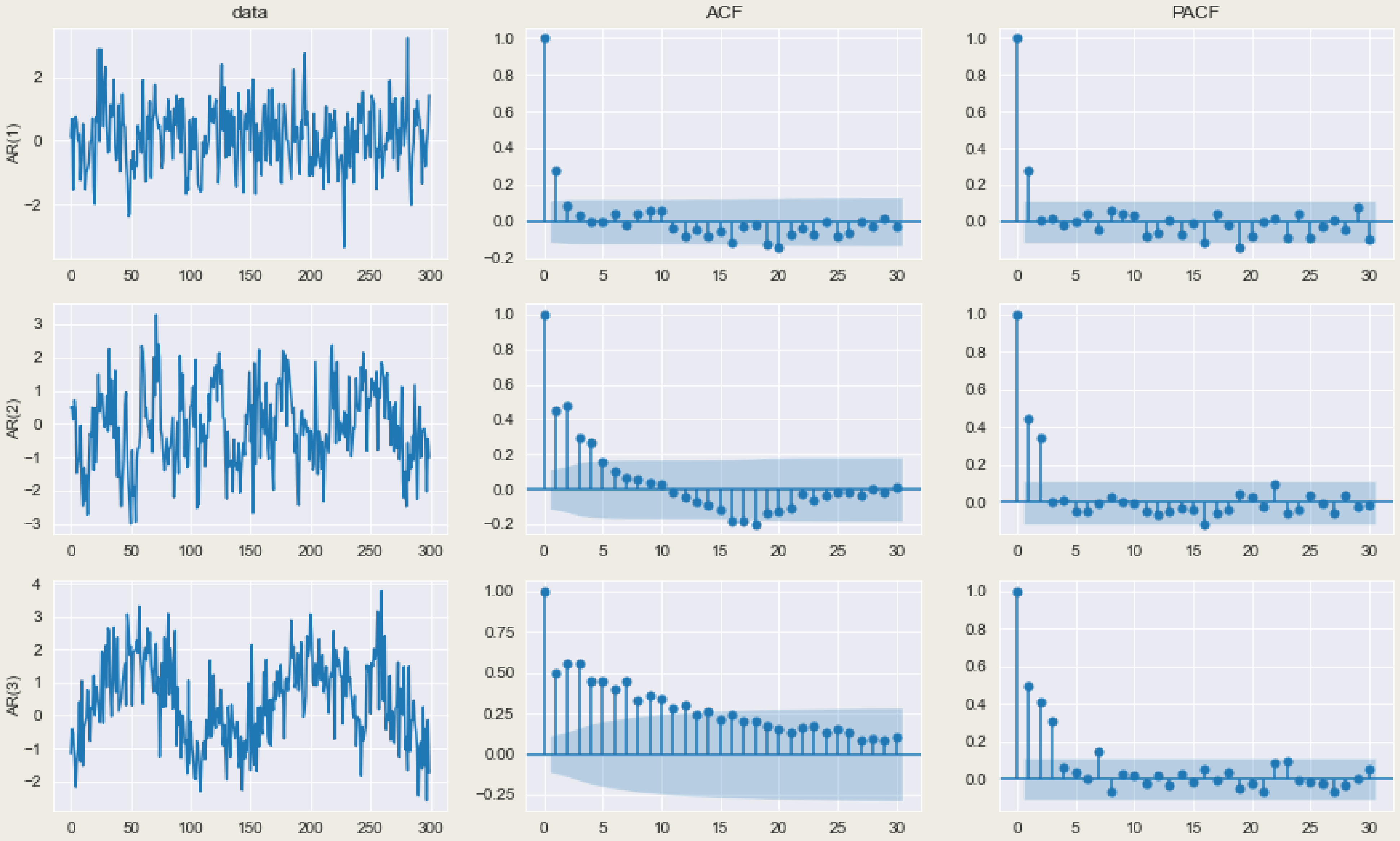
- We simulate the following AR processes

$$y_t = 0.3y_{t-1} + \epsilon_t$$
$$y_t = 0.3y_{t-1} + 0.3y_{t-2} + \epsilon_t$$
$$y_t = 0.3y_{t-1} + 0.3y_{t-2} + 0.3y_{t-3} + \epsilon_t$$

  and plot their ACF and PACF values.

- For an AR(n) process, lag-1 to lag-n PACF is likely to be significantly different from 0.

  However, there is no obvious pattern with ACFs.



Simulation of AR(1), AR(2), AR(3) process

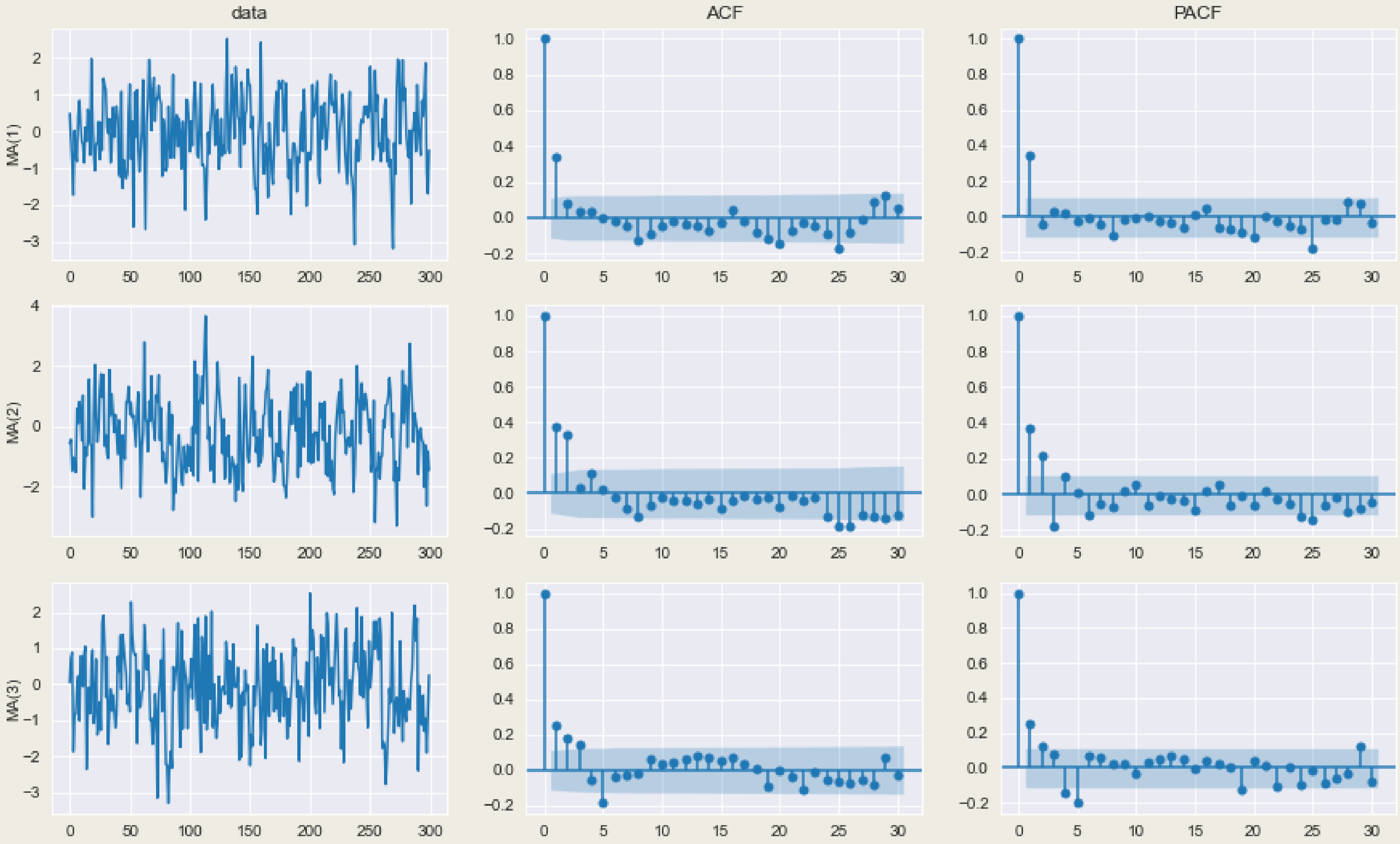
2. MA processes

- We simulate the following MA processes

$$y_t = \epsilon_t + 0.3\epsilon_{t-1}$$
$$y_t = \epsilon_t + 0.3\epsilon_{t-1} + 0.3\epsilon_{t-2}$$
$$y_t = \epsilon_t + 0.3\epsilon_{t-1} + 0.3\epsilon_{t-2} + 0.3\epsilon_{t-3}$$

  and plot their ACF and PACF values.

- for a MA(n) process, lag-1, to lag-n ACF is significantly different from 0.



Simulation of MA(1), MA(2), MA(3) process

3.How to identify type and order

- We can use ACF graphs and PACF graphs to identify the type and order of each process by observing the number of lags they remain significant.

- For example, if we find that only lag-1 to lag-4 PACF of a process is significantly different from 0,but the ACF remains significant for a large number of lags, we may consider it to be an AR(4).

- For another, if we find that only lag-1 to lag-3 ACF of a process is significantly different from 0, but no obvious pattern in PACF, we may think it to be a MA(3).