Week02 - Project Report

Ruchen Shi

Problem1

a.

The answer is:
Mean: 1.0489703904839585
Variance: 5.427220681881727
Skewness: 0.8792880598472443
Kurtosis: 23.069982510610544

b.

I use scipy.stats package in Python to calculate, the answer is:
Mean: 1.0489703904839585
Variance: 5.427220681881727
Skewness: 0.8819320922598395
Kurtosis: 23.24425346961619

c.

Considering that there is only a dataset provided, I can choose to simulate multiple sub-datasets from the original dataset by using bootstrapping to perform a t-test for assessing the potential bias in the statistical package functions. Bootstrapping is a resampling method that involves randomly selecting a sample of the data, with replacement, to create 'new' datasets.
Here's the general approach I take:

1. Perform resampling with replacement to create a number of bootstrap samples from the original dataset.
2. For each bootstrap sample, calculate the moments manually and using the statistical package.
3. For each moment, perform a paired t-test across all bootstrap samples to compare the manual and package calculations.

This method assumes that the original dataset is representative of the population. The bootstrap samples are treated as new datasets, and the paired t-test assesses whether there is a consistent difference between the manual and package calculations across these samples.
The output is below:

```
Paired t-test results:
Mean: statistic=nan, p-value=nan
Variance: statistic=nan, p-value=nan
Skewness: statistic=-18.65461424523332, p-value=7.234695977072955e-67
Kurtosis: statistic=-102.69203889394862, p-value=0.0
```

For mean and variance:
The means and variances calculated manually and by the package are identical. This is expected, as these are straightforward calculations that should not differ between methods. So, the output is nan.

For kurtosis:

Assume the kurtosis function is not biased.
That is, the null hypothesis H0 is: the kurtosis function is not biased.
And the alternative hypothesis is: the kurtosis function is biased.
We set the threshold as 0.05 and conduct the student t test: When the sample size is the same as the original dataset, and the number of bootstrap samples is 1000, The p-value is approximately 0, which is smaller than 0.05. So, we reject the null hypothesis, and the conclusion is that the kurtosis function is biased, and the result is statistically significant.

For skewness:

Assume the skewness function is not biased.
That is, the null hypothesis H0 is: the skewness function is not biased.
And the alternative hypothesis is: the skewness function is biased.
We set the threshold as 0.05 and conduct the student t test: When the sample size is the same as the original dataset, and the number of bootstrap samples is 1000, The p-value is approximately 3.17e-84, which is smaller than 0.05. So, we reject the null hypothesis, and the conclusion is that the skewness function is biased, and the result is statistically significant.
As a conclusion, I can say that my statistical package functions are biased.

Problem2

a.

The Ordinary Least Squares (OLS) regression yields the following results:

- The intercept (constant) coefficient β_0 is approximately -0.087.
- The slope coefficient β_1 for variable **x** is approximately 0.775.
- The standard deviation of the OLS residuals is approximately 1.006.

The Maximum Likelihood Estimation (MLE) under the assumption of normality yields the following results:

- The estimated intercept (constant) coefficient β_0 is approximately -0.087.
- The estimated slope coefficient β_1 for variable **x** is approximately 0.775.
- The estimated standard deviation (sigma) of the errors is approximately 1.004.

The similarity in the coefficients and standard deviations suggests that the assumptions of the OLS model are close to those of the normal distribution assumed in the MLE. The slight differences in the standard deviations (1.006 for OLS vs 1.004 for MLE) might be due to the MLE's explicit optimization to maximize the likelihood of the observed data under the assumption of normality, which can lead to a slightly tighter fit of the data as indicated by a slightly smaller estimated standard deviation.

b.

The MLE with T-distributed errors yields the following results:

- The estimated intercept (constant) coefficient β_0 is approximately -0.092.
- The estimated slope coefficient β_1 for variable x is approximately 0.615.
- The estimated standard deviation (sigma) of the errors is approximately 0.741.
- The log-likelihood for the T-distribution is approximately -284.055.

For comparison, the previous MLE under the normality assumption had:

- A log-likelihood for the normal distribution of approximately -284.538.

Comparing the log-likelihood values, the model with T-distributed errors has a higher log-likelihood (less negative) than the one with normally distributed errors, indicating a better fit to the data since the higher the log-likelihood, the better the model fits the data.

c.

For each observed value of $X1$, the conditional distribution of $X2$ is a normal distribution with the mean and variance specified by the conditional formulas you've used:

- The mean of $X2$ given $X1$ is calculated as:

$$\text{mean}(X_2|X_1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1)$$

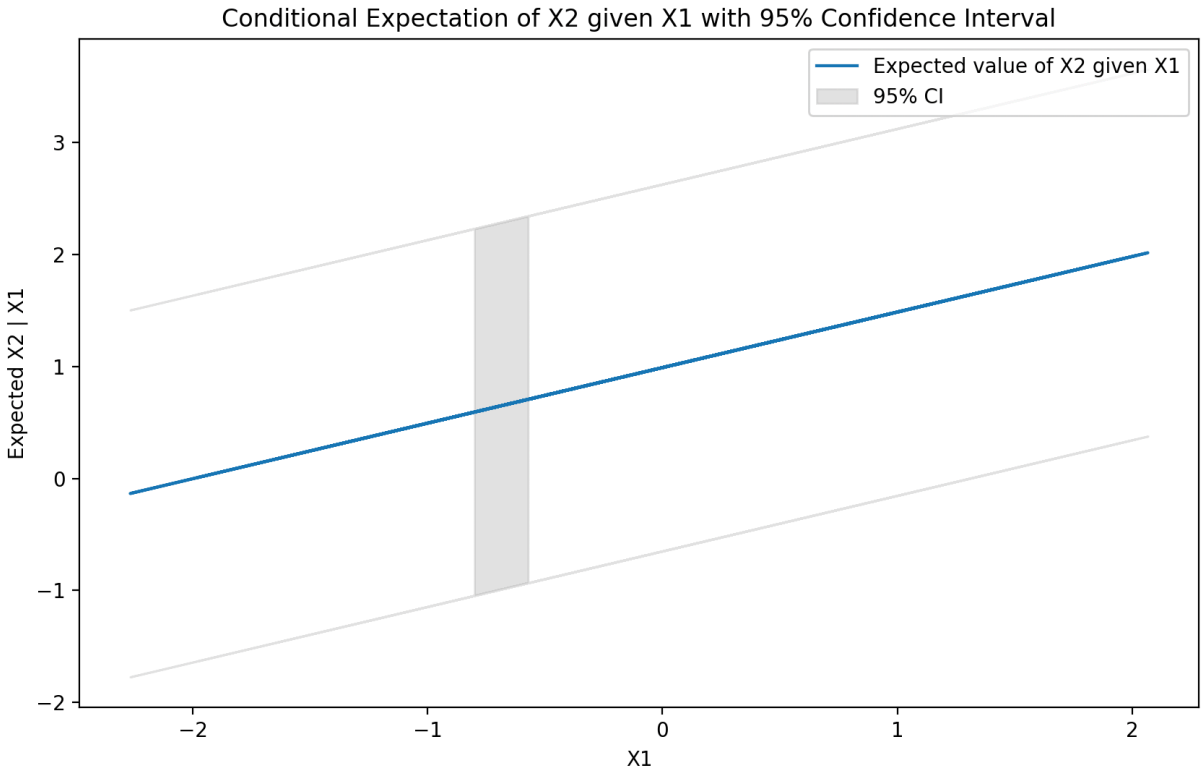
- The variance of $X2$ given $X1$ is:

$$\text{var}(X_2|X_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

Where:

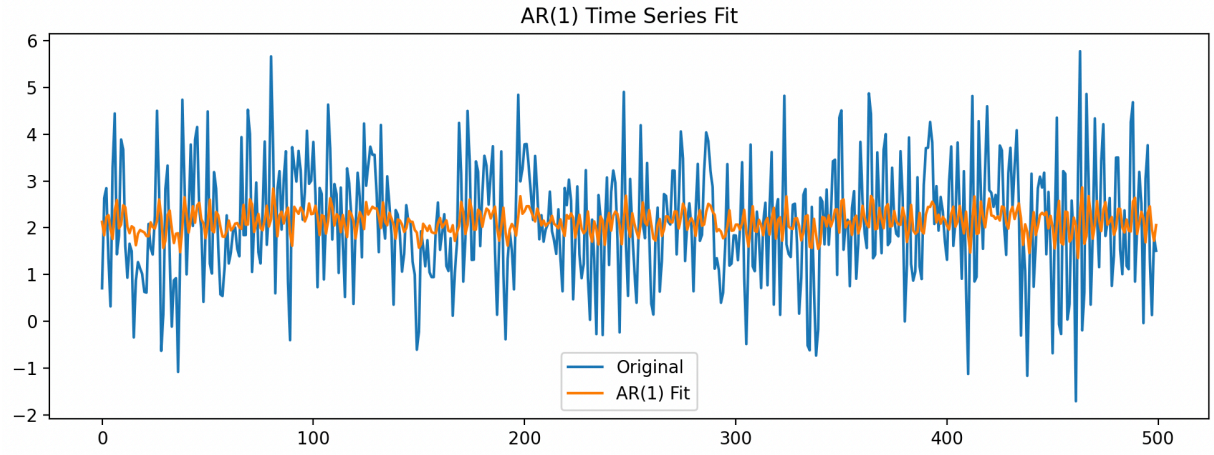
- $\mu 1$ and $\mu 2$ are the sample means of $X1$ and $X2$, respectively.
- $\Sigma 11$, $\Sigma 22$, $\Sigma 12$, and $\Sigma 21$ are the elements of the sample covariance matrix.

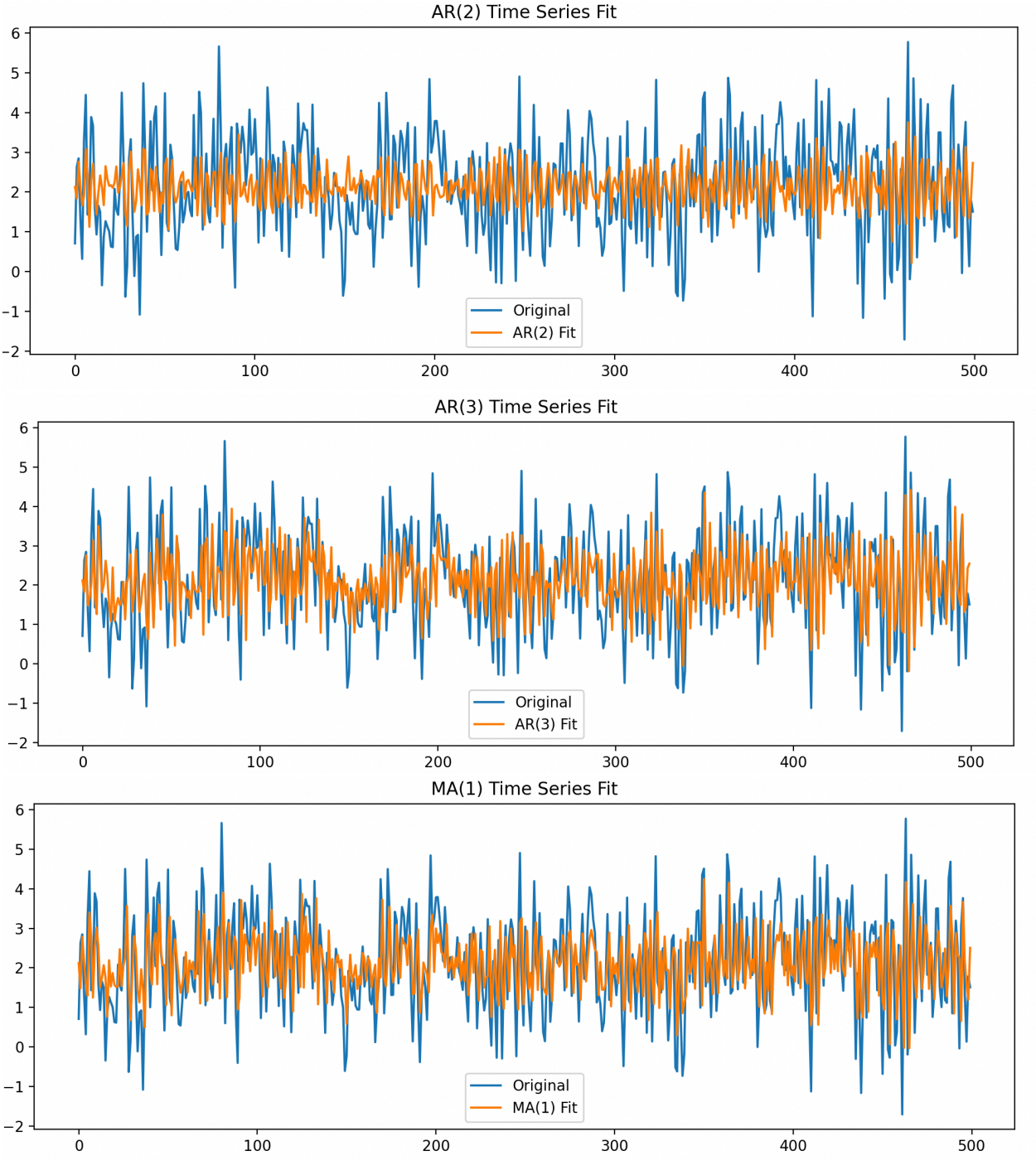
For any particularly observed value $x1$ of $X1$, $X2$ follows a normal distribution with the mean and variance computed using the above formulas. This results in a unique normal distribution for $X2$ for each value of $X1$, characterized by its own mean and variance.

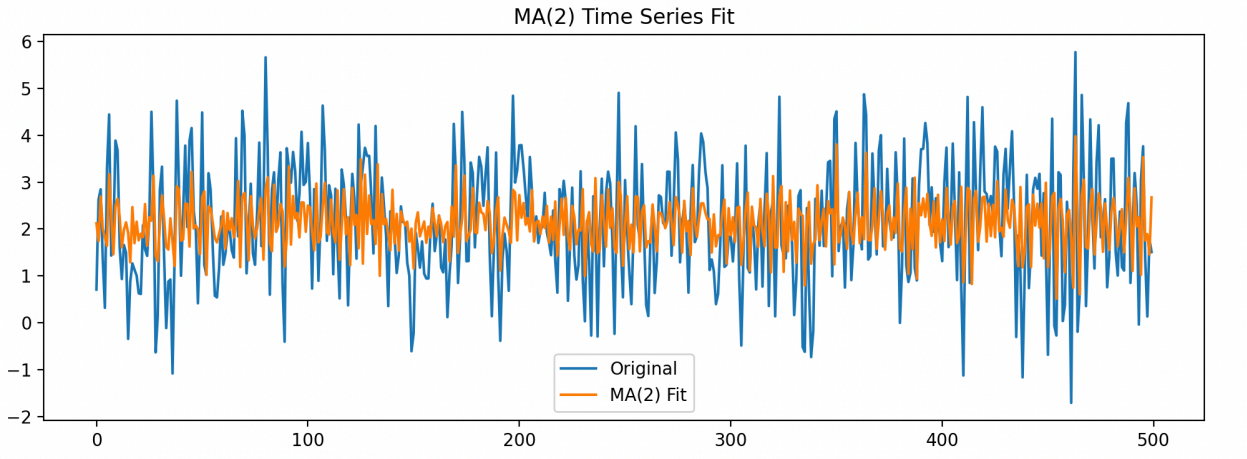


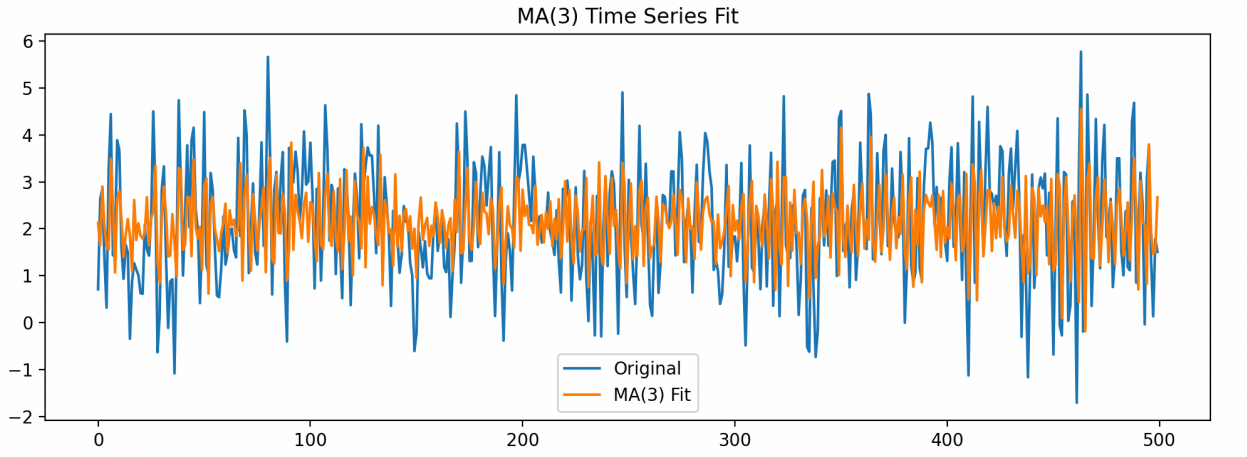
Problem3

I use Python statsmodels package to fit the data. The fitting graphs are as follows:




AR(2) Time Series Fit
Original
AR(2) Fit
AR(3) Time Series Fit
Original
AR(3) Fit
MA(1) Time Series Fit
Original
MA(1) Fit





From the graphs, we can intuitively see that AR(3) seems to be the best fit. However, we need a more precise and quantitative analysis or criteria to determine. I choose the AIC and BIC to help me determine which one is the best fit. If the AIC/BIC value of a model is smaller, the fitting of this model is better.

The AIC and BIC value for each model is as below:

```
AIC and BIC for each model:
AR(1): AIC=1644.6555047688475, BIC=1657.299329064114
AR(2): AIC=1581.079265904978, BIC=1597.9376982986669
AR(3): AIC=1436.6598066945867, BIC=1457.7328471866977
MA(1): AIC=1567.4036263707874, BIC=1580.047450666054
MA(2): AIC=1537.9412063807388, BIC=1554.7996387744276
MA(3): AIC=1536.8677087350309, BIC=1557.9407492271419
```

After comparation, we can conclude that based on AIC or BIC, the AR(3) is the best fit.