Homework – Week 2

Yiting Song

September 15, 2024

Problem 1 (3pts)

Given the dataset in problem1.csv

1. Calculate the first 4 moment values using the normalized formulas in the Week 1 notes.
2. Calculate the first 4 moment values using your chosen statistical package.
3. Are your statistical package functions biased? Prove or disprove your hypotheses. Explain your conclusion.

```
 Mean:  1.0489703904839585
Variance:  5.4217934611998455
Skewness:  0.8806086425277363
Kurtosis:  23.122200789898972
```

These data are the 4 moments I calculated using the formula on the note. Here I did not do anything unbiased with the variance, skewness and kurtosis, so they are all biased estimates.

```
 Mean_lib:  1.0489703904839585
Variance_lib:  5.4217934611998455
Skewness_lib:  0.8806086425277364
Kurtosis_lib:  23.12220078989723
```

These data are the result of my calculations using python's math functions. I calculated the mean and variance using the `numpy` library and the skewness and kurtosis using the `scipy.stats` library.

```
 mean_difference:  0.0
variance_difference:  0.0
skewness_difference:  -1.1102230246251565e-16
kurtosis_difference:  -3.552713678800501e-15
```

These figures are the differences I obtained by subtracting the data in the two graphs as above.

In conclusion, the Moments obtained by these two methods can be seen as equal after ignoring the small error introduced by the precision of the floating point numbers. So, we can see that these python libraries are biased estimates for the calculation of variance, skewness and kurtosis.

Problem 2 (5pts)

Assume the multiple linear regression model $Y = X\beta + \epsilon$

1. Fit the data in `problem2.csv` using OLS. Then fit the data using MLE given the assumption of normality. Compare the beta values and the standard deviation of the OLS errors to the fitted MLE σ. What is your finding? Explain any differences.

```
 OLS Intercept:-0.08738446427005074
OLS Beta:0.7752740987226111
OLS Residual Std:1.008813058320225
MLE Intercept:-0.08738446427005074
MLE Beta:0.7752740987226111
MLE s:1.0037564910465389
```

OLS: Use `LinearRegression()` to fit the regression model, get the slope and intercept, and calculate the standard error from the residuals.

MLE: Define a maximized log-likelihood function containing the slope of the regression model, the intercept, and the standard deviation of the error, s. Minimize the negative log-likelihood function by `scipy.optimize.minimize` to find the optimal MLE parameters.

The beta values obtained from OLS and MLE are the same. The standard error values are different because the MLE estimator is biased while OLS is unbiased.

2. Fit the data in `problem2.csv` using MLE given the assumption of a T distribution of errors. Show the fitted parameters. Compare the fitted parameters among the MLE under the normality assumption and T distribution assumption. Which is the best fit?

```
 MLE Intercept:-0.08738445444737096
MLE Beta:0.7752740987226111
MLE s:1.0037564910465198
AIC:575.0751261088672
T distribution MLE Intercept:-0.09726670821914582
T distribution MLE Beta:0.6749772607216558
T distribution MLE df:7.158538103286912
T distribution MLE s:0.8550715107877596
T distribution AIC:570.5868066278038
```

Use MLE to fit the data, assume that the error terms conformed to normal and t-distributions. Calculate the AIC to compare the fit of the two models.

Normal distribution assumption: find the optimal intercept, slope, and standard deviation of the residuals to maximize the log-likelihood function.

T-distribution assumption: find the optimal intercept, slope, degrees of freedom, and standard deviation of the residuals to maximize the log-likelihood function.

The AIC values and the standard deviation under the T-distribution assumption are smaller than the normal distribution. Thus, the T-distribution assumption is a better fitting model

3. Fit a multivariate distribution to the data in `problem2.csv` Given the values of what X1 are the conditional distributions for X2 for each observation. Plot the expected value X2 along with the 95% confidence interval and the observed value.

```
 Mean of X1:  0.0010226951490000004
Mean of X2:  0.9902438191000001
Covariance Matrix:
```

$$\begin{bmatrix} 1.06977464 & 0.53068455 \\ 0.53068455 & 0.96147329 \end{bmatrix}$$

```
Cov(X1, X1):  1.0697746428027173
Cov(X1, X2):  0.5306845547134215
Cov(X2, X2):  0.9614732933624854
```

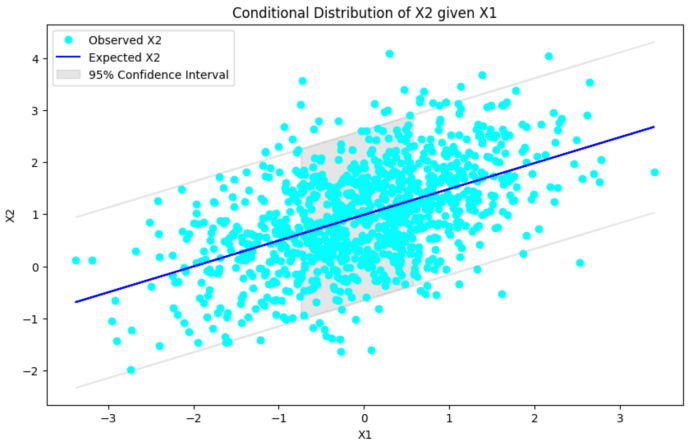

Figure 1: conditional distribution of X2 given X1

First, calculate the mean of X1 and X2, and estimate the relationship between them by the covariance matrix. Then, use the formulas for conditional mean and variance to refer the conditional mean and variance of X2 given X1. By the conditional mean and variance, calculate the 95% confidence interval for X2 at each X1 value. Finally, the observed X2 values, the expected X2 values, and the 95% confidence intervals are visualized.

The observed X2 values are scattered around the conditional expected value, with most falling within the 95% confidence interval, indicating that the model using the covariance matrix and mean is a reasonable fit for the data.

4. (1 point Extra Credit). $Y = X\beta + \epsilon$ and $\epsilon \sim N(0, \sigma^2 I)$. Derive the maximum likelihood estimators for β and σ^2.

$$\epsilon \sim N(0, \sigma^2)$$
$$Y \sim N(X\beta, \sigma^2 I)$$

The probability density function (PDF) of Y is:

$$f(Y|X, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{(Y - X\beta)^\top (Y - X\beta)}{2\sigma^2}\right)$$

The log-likelihood function is:

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^\top (Y - X\beta)$$

To find the maximum likelihood estimator for β, we take the derivative of the log-likelihood with respect to β:

$$\frac{\partial \ell(\beta, \sigma^2)}{\partial \beta} = \frac{1}{\sigma^2} X^\top (Y - X\beta) = 0$$

Solving for β, we get the MLE for β:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

Next, we find the MLE for σ^2 by taking the derivative of the log-likelihood with respect to σ^2:

$$\frac{\partial \ell(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)^\top (Y - X\beta) = 0$$

Solving for σ^2, we get:

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\beta})^\top (Y - X\hat{\beta})}{n}$$

Thus, the MLE for β and σ^2 are:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y, \quad \hat{\sigma}^2 = \frac{(Y - X\hat{\beta})^\top (Y - X\hat{\beta})}{n}$$

Problem 2 (3pts)

1. Examine the data in `problem3.csv`; which AR(n) or MA(n) model do you expect to fit this data best? Fit the data using AR(1) - AR(3) and MA(1) - MA(3) models. Which is the best fit and does this confirm your hypothesis?

```
 ARIMA(1, 0, 0) - AIC: 1644.6555047688475
ARIMA(2, 0, 0) - AIC: 1581.0792659049775
ARIMA(3, 0, 0) - AIC: 1436.6598066945826
ARIMA(0, 0, 1) - AIC: 1567.4036263707872
ARIMA(0, 0, 2) - AIC: 1537.941206380739
ARIMA(0, 0, 3) - AIC: 1536.8677087350306
```

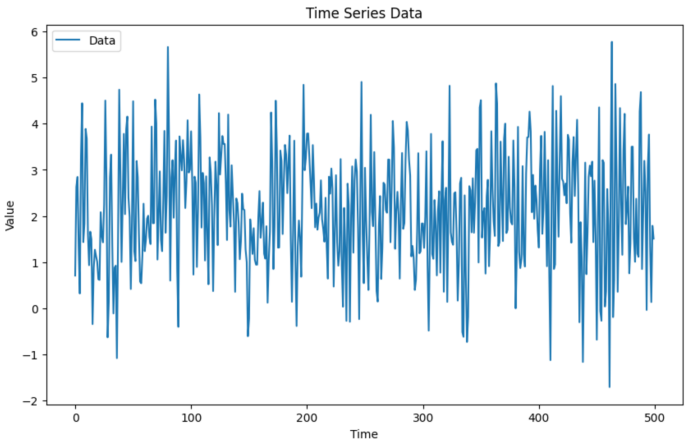

Figure 2: Time Series

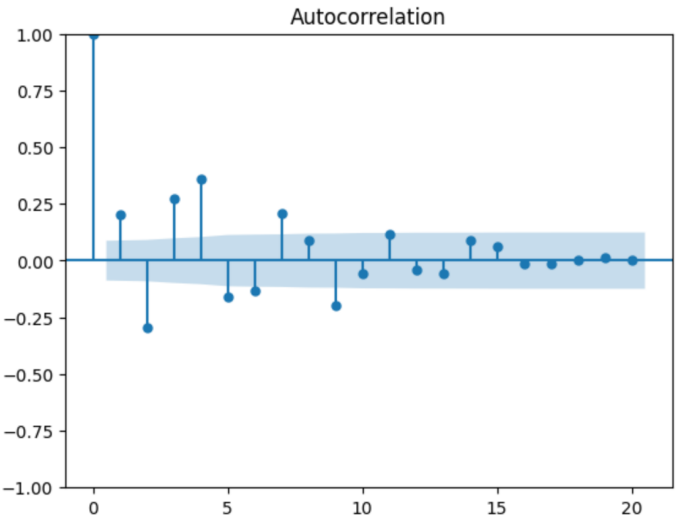

Figure 3: Autocorrelation

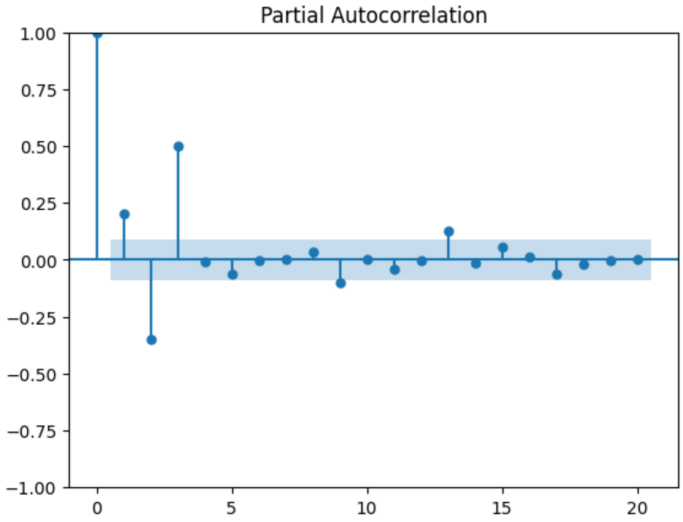

Figure 4: Partial Autocorrelation

The ACF shows a strong initial autocorrelation followed by a rapid decline. Indicates that the data may be amenable to a AR model, such as the AR(1) model, capturing the relationship between the current value and the previous period.

The PACF shows a strong correlation at lag 1 and lag 2, indicating that we may need a multi-lag AR model such as AR(2) or AR(3).

The AR(3) model has the lowest AIC value, suggesting that the AR(3) model may be more appropriate for this data set. The MA(1) to MA(3) models have higher AIC values, indicating that the AR model captures the characteristics of the data better than the MA model.

Based on the ACF and PACF, and the AIC values, AR(3) is the most appropriate model for this data.