Week2 assignment

Problem 1

a. **calculate the first four moments values by using normalized formula in the "Week1 - Univariate Stats".**

Moment:

1. The mean. The expected value of the distribution
2. Variance. This measures the dispersion of the distribution around the mean.
3. Skewness. Measures the asymmetry of the distribution. Positive skew distributions skew to the right with the mean > median. Negative skew distributions skew to the left with mean < median.
4. Kurtosis. Measures the fattness of the tails.

$$\hat{\mu}_3 = E[(\frac{X-\mu}{\sigma})^3] = \frac{u_3}{\sigma^3}$$

$$\hat{\mu}_4 = E[(\frac{X-\mu}{\sigma})^4] - 3 = \frac{u_4}{\sigma^4} - 3$$

We subtract 3 from the Kurtosis to report "Excess" Kurtosis. The kurtosis of the Normal Distribution is 3 – excess kurtosis is kurtosis in excess of the normal distribution.

Thus, the first four moments values are as follows.

	x
Mean	1.0490
Variance	5.4272
Skewness	0.8793
Kurtosis	23.0700

b. calculate the first four moments values again by using your chosen statistical package.

Utilizing functions from the scipy.stats module, the first four moments values are as follows.

	x
Mean	1.0490
Variance	5.4272
Skewness	0.8806
Kurtosis	23.1222

c. Is your statistical package functions biased? Prove or disprove your hypothesis respectively.

The function for mean and skewness is unbiased, while the function for variance and kurtosis is not unbiased.

First, I set null hypotheses to be: H0: functions for mean, variance, skewness and kurtosis are unbiased and threshold, or alpha, to be 5%.
Then, I go through:
1. Sample 100 standardized random normal values
2. Calculate 4 moments values for the sample
3. Sample the mean, variance, skewness and kurtosis by repeating steps 1 and 2 1000 times
4. Calculate mean and standard deviation of 4 moments values respectively
5. Calculate T statistics with null hypotheses being all functions are unbiased
6. Use the CDF function to calculate the p-value corresponding to a two-sided test

The outcome is as follows:
Mean T-statistic: 0.3535520851234843, p-value: 0.7237491845625073
Variance T-statistic: 223.0129292783607, p-value: 0.0
Skewness T-statistic: -0.1448558966634165, p-value: 0.8848538669185118
Kurtosis T-statistic: -2.9320560616231655, p-value: 0.003444055974008542
Fail to reject the null hypothesis for mean - function might be unbiased.
Reject the null hypothesis for variance - function might be biased.
Fail to reject the null hypothesis for skewness - function might be unbiased.
Reject the null hypothesis for kurtosis - function might be biased.

Problem2

a. Fit the data in problem2.csv using OLS. Then, fit the data using MLE given the assumption of normality. Compare their beta and standard deviation of the OLS error to the fitted MLE σ. What's your finding? Explain any differences.

To fit the data using OLS, I construct the regressor matrix and calculate the intercept and coefficient.

$$X = [1 \ X]$$

From a linear algebra perspective:

$$Y = X\beta$$

Multiply each side by $(X'X)^{-1}X'$

To fit the data using MLE given the assumption of normality, I use the normal_log_likelihood function to define the log-likelihood of the model. I then use an optimization algorithm (scipy.optimize.minimize) to maximize the log-likelihood. Since most optimization algorithms are designed to minimize functions, I minimize the negative of the normal_log_likelihood function. This function evaluates the fit of a set of parameters(β,σ) to the observed data under the assumption that the model errors follow a normal distribution:

$$\ell(\beta, \sigma) = -\frac{1}{2} \sum \left[\log(2\pi\sigma^2) + \frac{(y_i - \hat{y}_i)^2}{\sigma^2} \right]$$

The outcome is as follows:

	beta - const	beta - x	sigma
OLS	-0.087384	0.775274	1.003756319417732
MLE	-0.087384	0.775274	1.003756313795847

The coefficients (beta) and the standard deviation (sigma) obtained from fitting the data using Ordinary Least Squares (OLS) and Maximum Likelihood Estimation (MLE) under the assumption of normality are almost identical. The minuscule difference in the sigma estimates is a result of computational precision and the inherent differences in the optimization processes of OLS and MLE, but it does not significantly impact the model's interpretation or conclusions.

b. 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 MLE under normality assumption and T distribution assumption. Which is the best of fit?

Unlike the normal distribution, a T-distribution includes an additional parameter, the degrees of freedom, which influences the shape of the distribution.
The outcome is as follows:

	beta - const	beta - x	sigma	R^2
MLE under normality assumption	-0.08738446	0.77527409	1.0037563137958476	0.3456068835648126
MLE under T distribution assumption	-0.08743013	0.7750945	1.1913429591339617	0.3456068629024218

An R^2 value closer to 1 indicates that the model explains a greater proportion of the variability, suggesting a better fit. So MLE under normality assumption is the best of fit as it has a smaller sigma and bigger R^2.

c. Fit the data in problem2_x.csv using MLE given $X = [X_1, X_2]$ follows the multivariate normal distribution. Assume X as a random variable, follows the fitted gaussian distribution, X_1 (problem2_x1.csv) are a part of observed value of X, What's the distribution of X_2 given each observed value? Plot the expected value along with the 95% confidence interval.

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Given $X_1 = a$

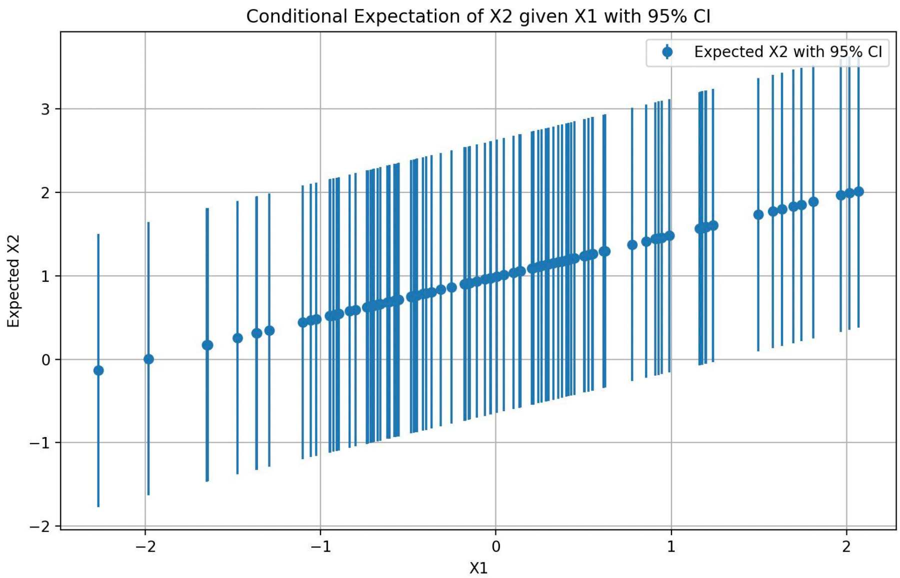
$$X_2 \sim N(\bar{\mu}, \bar{\Sigma})$$

where $\bar{\mu} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(a - \mu_2)$

$$\bar{\Sigma} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}$$

So I calculate the mean and variance of X1 to fit a Gaussian distribution and fits a multivariate normal distribution to X = [X1, X2] and computes the mean vector and covariance matrix. Then I derive the conditional distribution parameters of X2 given X1. For each observed value of X1, it calculates the expected value of X2 and the 95% confidence interval.

The plot is as follows:



d. (Extra Credit: 1 point) Assume $\epsilon \sim N(0, \sigma^2 I_n)$, using Maximum Likelihood Estimation (MLE), derive the estimator for β and σ^2. Show your detailed proof.

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

$$Y = X\beta + \varepsilon \qquad f(\varepsilon_i \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2}$$

$$\varepsilon_i = y_i - x_i^T\beta \qquad L(\beta, \sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i \mid \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right)}$$

So the log-likelihood : $\ell(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\underbrace{(y_i - x_i^T\beta)^2}$

take the derivative of ℓ with respect to σ^2 and set it to zero :

$$\frac{\partial}{\partial\beta}(y_i - x_i^T\beta)^T(y_i - x_i^T\beta) = -2x_i(y_i - x_i^T\beta)$$

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - x_i^T\beta) = 0$$

$$\therefore \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i x_i^T \beta$$

$$X^T Y = X^T X \beta$$

① $\hat{\beta} = (X^T X)^{-1} X^T Y$

② $\dfrac{\partial \ell}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{(Y - X\hat{\beta})^T(Y - X\hat{\beta})}{2(\sigma^2)^2} = 0$

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

To make it unbiased :

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

Problem3
Fit the data in problem3.csv using AR (1) through AR (3) and MA (1) through MA (3), respectively. Which is the best of fit?

AR models rely on a linear combination of their own past values, whereas MA models depend on a linear combination of past error terms. The ARIMA model, which combines these approaches, can be specialized into pure AR or MA models by adjusting its parameters (p,d,q). I used the ARIMA model from the statsmodels library to fit the data and calculated the Akaike Information Criterion (AIC) for each model. The AIC is a criterion for evaluating the fit of a model, intended to quantify the balance between the fit of the model to the data and the complexity of the model. Generally, the lower the AIC, the better the model.
The outcome is as follows:
ARIMA(1, 0, 0) AIC: 1644.6555047688475
ARIMA(2, 0, 0) AIC: 1581.079265904978
ARIMA(3, 0, 0) AIC: 1436.6598066945867

ARIMA(0, 0, 1) AIC: 1567.4036263707874
ARIMA(0, 0, 2) AIC: 1537.9412063807388
ARIMA(0, 0, 3) AIC: 1536.8677087350309
So the best model is AR(3) with AIC: 1436.6598