June Lee

Project Week 2

Problem 1)

The results from manually calculating the mean, variance, skew, and kurtosis in Python is shown under "Part 1." The results from using Pandas DataFrame is shown under "Part 2."

```
Part 1
Mean = 1.0489703904839585
Variance = 5.427220681881726
Skew = 0.8819320922598405
Kurtosis = 23.244253469616186
Part 2
Mean = 1.0489703904839582
Variance = 5.4272206818817255
Skew = 0.8819320922598395
Kurtosis = 23.2442534696162
```

The documentation for Pandas DataFrame states that it returns unbiased estimates of variance, skew, and kurtosis. Given the formulas for unbiased estimators of variance, skew, and kurtosis, we can confirm this by generating a simple dataset, with the number of observations equal to 4.

Consider the dataset [0, 0, 2, 6]. Then, the sample mean is equal to 2.

$x = [0, 0, 2, 6] \qquad \bar{x} = 2$

unbiased $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{4} (x_i - 2)^2}{4-1} = \dfrac{4+4+0+16}{3} = \dfrac{24}{3} = \boxed{8}$

unbiased skew $= \dfrac{n}{(n-1)(n-2)} \dfrac{\sum_{i=1}^{4} (x_i - 2)^3}{(\hat{\sigma}^2)^{\frac{3}{2}}}$

$= \left(\dfrac{4}{(3)(2)}\right)\left(\dfrac{-8 + -8 + 0 + 64}{(8)^{\frac{3}{2}}}\right)$

$= \left(\dfrac{2}{3}\right)\left(\dfrac{48}{(8)^{3/2}}\right) = \boxed{\sqrt{2}} \approx 1.4142$

unbiased kurt. $= \left(\dfrac{(n+1)n}{(n-1)(n-2)(n-3)}\right) \dfrac{\sum_{i=1}^{4} (x_i - 2)^4}{(\hat{\sigma}^2)^2} - 3\left(\dfrac{(n-1)^2}{(n-2)(n-3)}\right)$

$= \left(\dfrac{(5)(4)}{(3)(2)(1)}\right)\left(\dfrac{16+16+0+256}{64}\right) - \dfrac{3(3^2)}{(2)(1)}$

$= \left(\dfrac{10}{3}\right)\left(\dfrac{9}{2}\right) - \dfrac{27}{2} = \boxed{\dfrac{3}{2}}$

The results obtained using the package confirms this, and thus are unbiased.

```
----------------------------------------
Part 3
Mean = 2.0
Variance = 8.0
Skew = 1.414213562373095
Kurtosis = 1.5
```

Problem 2)

1) Fitting the data in problem2.csv using OLS, the following results are obtained.

```
LinregressResult(slope=0.7752740987226117, intercept=-0.08738446427005078,
rvalue=0.5878833928295756, pvalue=5.594765259448202e-20,
stderr=0.07581425102261538, intercept_stderr=0.07149629062481402)
```

The "stderr" that is reported by stats.linregress is the standard error of the estimate of the slope. To obtain the standard error of the OLS error terms, set

$$\varepsilon_i = y_i - (intercept + slope * x_i)$$

and calculate the biased sample variance of the error terms, which yields 1.003756.

Using MLE (assuming errors are normally distributed), the following results are obtained

```
Estimated Beta coefficients MLE: [-0.08738448  0.77527452]
Estimated std err MLE: 1.0037562314196935
```

where beta_0 = -0.08738 and beta_1 = 0.77527. This is as expected since the MLE of the beta is equal to that under OLS, and the MLE of the standard error of the residuals is equal to the biased sample standard error of the residuals (as proven in extra credit).

2) Using MLE (assuming a T-distibution of errors), the following results are obtained.

```
Estimated coefficients: [-0.09726859  0.6750099 ]
Estimated degrees of freedom (nu): 7.159860005381045
Estimated scale parameter (sigma): 0.8551054189259828
```

Comparing the fitted values, we observe that the intercept is slightly more negative (-0.0972 vs -0.0874) and the slope is slightly less positive (0.675 vs 0.775). It is difficult to conclude whether the MLE under normal distribution or the MLE under T-distribution better fits the data. While the standard deviation of the errors is lower under the T-distribution than the normal distribution (0.855 vs 1.003), we have also estimated for the degrees of freedom under the T-distribution. Thus, we would need some context behind the data to determine whether the assumption that the errors follow a T-distribution is valid before declaring it to be a better fit.

3) Given problem2_x.csv, we seek to fit the data to a multivariate normal distribution such that $X \sim N(\mu, \Sigma)$, where X is a (2x1) random vector, μ is a (2x1) vector of the expected values, and Σ is the (2x2) covariance matrix. Using Pandas DataFrame, we obtain the following results for sample mean and covariance matrix.

```
x1    0.001023
x2    0.990244
dtype: float64
          x1        x2
x1  1.069775  0.530685
x2  0.530685  0.961473
```

From our notes, we know that given $X_1 = a, X_2 \sim N(\mu', \Sigma')$ for each observation where

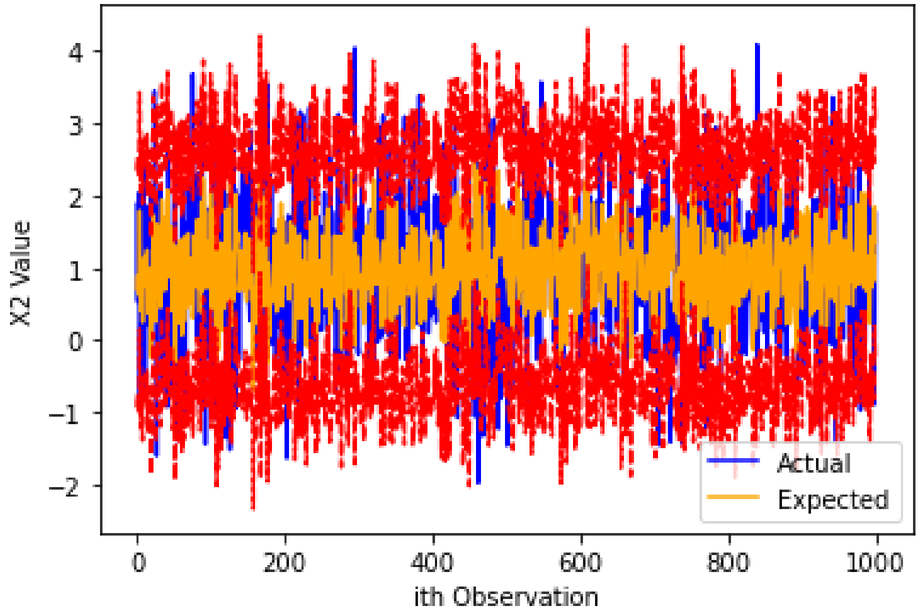
$$\mu' = \mu_2 + \Sigma_{12}\Sigma_{11}^{-1}(a - \mu_1)$$

$$\Sigma' = \Sigma_{22} - \Sigma_{12}\Sigma_{11}^{-1}\Sigma_{12}$$

Given our observed values, we arrive at

$$E(X_2) = \mu_2 = \overline{X_2} + \frac{0.540685}{1.069775}(a - \overline{X_1})$$

$$var(X_2) = 0.961473 - \frac{0.530685^2}{1.069775}$$



4)

4) $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$.

~~$\varepsilon = Y - X\beta$~~.

Since $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$,

$$f_Y(Y_i \mid x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_i - x_i\beta)^2}{2\sigma^2}\right).$$

$$\prod_{i=1}^{n} f_Y(Y_i \mid x_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - x_i\beta)^2\right)$$

Taking log-likelihood,

$$\ell = \left(\frac{-n}{2}\right) \ln(2\pi\sigma^2) + \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - x_i\beta)^2$$

differentiating w/ respect to β,

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

assuming $\frac{1}{\sigma^2} \neq 0$.

$$0 = \sum_{i=1}^{n} x_i^T Y_i - \sum_{i=1}^{n} x_i^T x_i \beta$$

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

$$\hat{\beta}_{MLE} = \left(\sum_{i=1}^{n} x_i^T x_i\right)^{-1} \sum_{i=1}^{n} x_i^T Y_i$$

$$= \boxed{(X^T X)^{-1} X^T Y}.$$

$$\frac{\partial l}{\partial \sigma^2} = \frac{-n}{2\sigma^2} - \left(\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\beta)^2\right)\left(\frac{-1}{(\sigma^2)^2}\right)$$

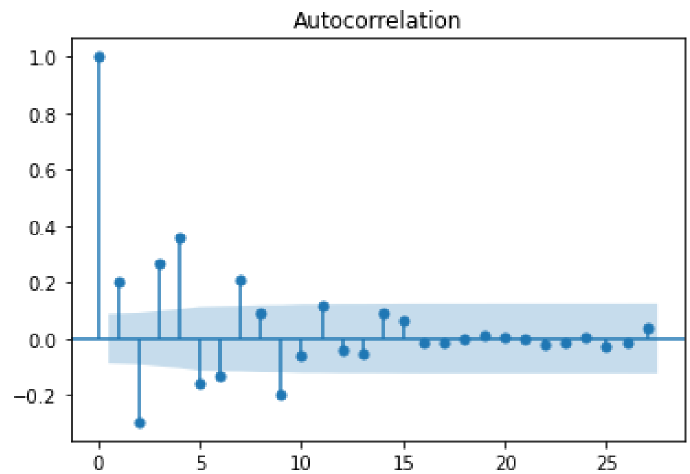
$$= \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2$$

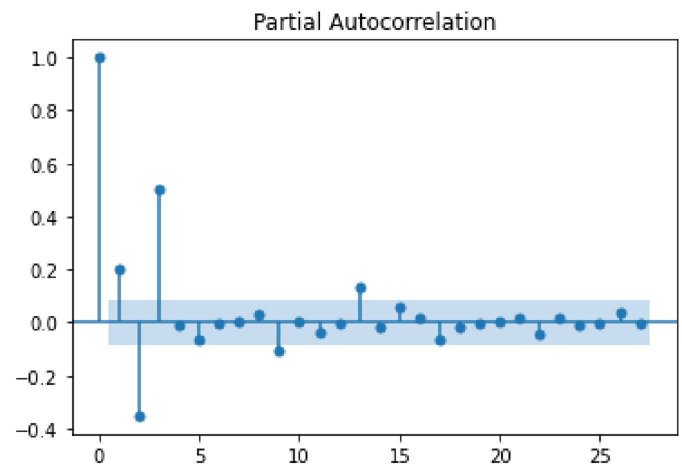
$$= \left(\frac{1}{2\sigma^2}\right)\left(-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2\right)$$

setting $\frac{\partial l}{\partial \sigma^2} = 0$ and assuming $\frac{1}{2\sigma^2} \neq 0$,

$$\boxed{\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}.$$

Problem 3)





The ACF and PACF plots seem to show autocorrelation decreasing rapidly following the first timestep, while the partial autocorrelation converges to zero over a longer time in a “cone” shape. Also seeing that the absolute value of the partial autocorrelation is quite strong for the first 3 timesteps, I would hypothesize that a MA(3) model would fit this data best.

```
AR(1)
                               SARIMAX Results
==============================================================================
Dep. Variable:                      x   No. Observations:                  500
Model:                 ARIMA(1, 0, 0)   Log Likelihood                -819.328
Date:                Sat, 14 Sep 2024   AIC                           1644.656
Time:                        11:47:08   BIC                           1657.299
Sample:                             0   HQIC                          1649.617
                                - 500
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1258      0.070     30.473      0.000       1.989       2.263
ar.L1          0.2019      0.045      4.512      0.000       0.114       0.290
sigma2         1.5517      0.105     14.743      0.000       1.345       1.758
===================================================================================
Ljung-Box (L1) (Q):                   2.51   Jarque-Bera (JB):                 1.42
Prob(Q):                              0.11   Prob(JB):                         0.49
Heteroskedasticity (H):               1.37   Skew:                            -0.00
Prob(H) (two-sided):                  0.04   Kurtosis:                         2.74
===================================================================================
```

```
AR(2)
                               SARIMAX Results
==============================================================================
Dep. Variable:                      x   No. Observations:                  500
Model:                 ARIMA(2, 0, 0)   Log Likelihood                -786.540
Date:                Sat, 14 Sep 2024   AIC                           1581.079
Time:                        11:47:08   BIC                           1597.938
Sample:                             0   HQIC                          1587.694
                                - 500
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1270      0.049     43.663      0.000       2.032       2.222
ar.L1          0.2732      0.042      6.486      0.000       0.191       0.356
ar.L2         -0.3505      0.043     -8.068      0.000      -0.436      -0.265
sigma2         1.3603      0.094     14.455      0.000       1.176       1.545
===================================================================================
Ljung-Box (L1) (Q):                  15.51   Jarque-Bera (JB):                 3.12
Prob(Q):                              0.00   Prob(JB):                         0.21
Heteroskedasticity (H):               1.20   Skew:                            -0.11
Prob(H) (two-sided):                  0.24   Kurtosis:                         2.68
===================================================================================
```

```
AR(3)
                               SARIMAX Results
==============================================================================
Dep. Variable:                      x   No. Observations:                  500
Model:                 ARIMA(3, 0, 0)   Log Likelihood                -713.330
Date:                Sat, 14 Sep 2024   AIC                           1436.660
Time:                        11:47:08   BIC                           1457.733
Sample:                             0   HQIC                          1444.929
                                - 500
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1209      0.085     24.990      0.000       1.955       2.287
ar.L1          0.4515      0.040     11.179      0.000       0.372       0.531
ar.L2         -0.4887      0.037    -13.104      0.000      -0.562      -0.416
ar.L3          0.5047      0.040     12.769      0.000       0.427       0.582
sigma2         1.0132      0.068     14.939      0.000       0.880       1.146
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 0.84
Prob(Q):                              0.90   Prob(JB):                         0.66
Heteroskedasticity (H):               1.04   Skew:                            -0.03
Prob(H) (two-sided):                  0.81   Kurtosis:                         2.81
===================================================================================
```

```
MA(1)
                               SARIMAX Results
==============================================================================
Dep. Variable:                      x   No. Observations:                  500
Model:                 ARIMA(0, 0, 1)   Log Likelihood                -780.702
Date:                Sat, 14 Sep 2024   AIC                           1567.404
Time:                        11:47:08   BIC                           1580.047
Sample:                             0   HQIC                          1572.365
                                - 500
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1236      0.085     25.028      0.000       1.957       2.290
ma.L1          0.6434      0.034     18.847      0.000       0.577       0.710
sigma2         1.3282      0.090     14.782      0.000       1.152       1.504
===================================================================================
Ljung-Box (L1) (Q):                  11.73   Jarque-Bera (JB):                 1.18
Prob(Q):                              0.00   Prob(JB):                         0.55
Heteroskedasticity (H):               1.39   Skew:                            -0.02
Prob(H) (two-sided):                  0.04   Kurtosis:                         2.77
===================================================================================
```

```
MA(2)
                               SARIMAX Results
==============================================================================
Dep. Variable:                      x   No. Observations:                  500
Model:                 ARIMA(0, 0, 2)   Log Likelihood                -764.971
Date:                Sat, 14 Sep 2024   AIC                           1537.941
Time:                        11:47:08   BIC                           1554.800
Sample:                             0   HQIC                          1544.556
                                - 500
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1255      0.060     35.199      0.000       2.007       2.244
ma.L1          0.4344      0.044      9.775      0.000       0.347       0.522
ma.L2         -0.2306      0.047     -4.949      0.000      -0.322      -0.139
sigma2         1.2473      0.086     14.558      0.000       1.079       1.415
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 1.67
Prob(Q):                              0.88   Prob(JB):                         0.43
Heteroskedasticity (H):               1.28   Skew:                            -0.03
Prob(H) (two-sided):                  0.11   Kurtosis:                         2.72
===================================================================================
```

```
MA(3)
                               SARIMAX Results
==============================================================================
Dep. Variable:                      x   No. Observations:                  500
Model:                 ARIMA(0, 0, 3)   Log Likelihood                -763.434
Date:                Sat, 14 Sep 2024   AIC                           1536.868
Time:                        11:47:08   BIC                           1557.941
Sample:                             0   HQIC                          1545.137
                                - 500
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1259      0.059     35.880      0.000       2.010       2.242
ma.L1          0.5582      0.045     12.333      0.000       0.469       0.647
ma.L2         -0.2286      0.053     -4.308      0.000      -0.333      -0.125
ma.L3         -0.1531      0.048     -3.216      0.001      -0.246      -0.060
sigma2         1.2394      0.085     14.592      0.000       1.073       1.406
===================================================================================
Ljung-Box (L1) (Q):                   1.60   Jarque-Bera (JB):                 1.75
Prob(Q):                              0.21   Prob(JB):                         0.42
Heteroskedasticity (H):               1.25   Skew:                            -0.06
Prob(H) (two-sided):                  0.15   Kurtosis:                         2.73
===================================================================================
```

However, the results seem to indicate that the AR(3) model is the best fit, with lower AIC and BIC scores of around 1436 and 1457, compared to that of the other models (which are greater than 1500).