FinTech545 - hw1

Samuel Fuller

January 2024

I have created a Github repository. The link to my repo is here. Various folders are named for navigation as well as filenames (specifically for code) corresponding to each question.

1 Problem 1

1.1 Part a

To solve part a, I implemented the summation formulas associated with the first four moments as outlined below:

$$\hat{\mu}_1 = E[X] = \frac{1}{n}\sum_{i}^{n} x_i = 1.05002041 \tag{1}$$

$$\hat{\mu}_2 = E\left[(X-\hat{\mu}_1)^2\right] = \frac{1}{n-1}\sum_{i}^{n}(x_i-\hat{\mu}_1)^2 = 5.42722179 \tag{2}$$

$$\hat{\mu}_3 = E\left[(X-\hat{\mu}_1)^3\right] = \frac{n}{(n-1)(n-2)(\sigma^3)}\sum_{i}^{n}(x_i-\hat{\mu}_1)^3 = 0.88057695 \tag{3}$$

$$\hat{\mu}_4 = E\left[(\frac{X-\hat{\mu}_1}{\sigma})^4\right] = \frac{1}{n}\sum_{i}^{n}(\frac{x_i-\hat{\mu}_1}{\sigma})^4 = 26.06838787 \tag{4}$$

1.2 Part b

The moment function I used standardized the moments was from the scipy stats package. Setting the center to zero

$$\hat{\mu}_1 = 1.04897039 \tag{5}$$
$$\hat{\mu}_2 = 5.42179346 \tag{6}$$
$$\hat{\mu}_3 = 11.11725066 \tag{7}$$
$$\hat{\mu}_4 = 767.88414813 \tag{8}$$

Since the function uses the following formula, its values differ significantly from the manually calculated values where c = 0 for $\hat{\mu}_1$ and c = μ_1 for all other $\hat{\mu}$:

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(x_i-c)^k$$

1.3 Part c

See the above detail on how the scipy stats moment function calculates each respective moment. Given the differences in the formula used by the statistical python package and the formulas I used, for example

$$\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_1)^3 \neq \frac{n}{(n-1)(n-2)(\sigma^3)}\sum_{i}^{n}(x_i-\hat{\mu}_1)^3$$

The package calculated moment divides simply by n, leading to a biased estimator (which will eventually converge to the unbiased estimator), whereas the manual calculation correctly rectifies the bias and normalizes skewness. The same is true for variance, mean and kurtosis.

2 Problem 2

2.1 Part a

After fitting OLS, we find the MLE parameters β and σ to be equal to:
$\beta = 0.7690$ $\sigma = 1.003775$

```
OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared (uncentered):          0.341
Model:                            OLS   Adj. R-squared (uncentered):     0.338
Method:                 Least Squares   F-statistic:                     103.1
Date:                Fri, 26 Jan 2024   Prob (F-statistic):           8.74e-20
Time:                        17:57:36   Log-Likelihood:                -285.29
No. Observations:                 200   AIC:                             572.6
Df Residuals:                     199   BIC:                             575.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.7690      0.076     10.154      0.000       0.620       0.918
==============================================================================
Omnibus:                       12.156   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               17.262
Skew:                           0.389   Prob(JB):                     0.000178
Kurtosis:                       4.211   Cond. No.                         1.00
==============================================================================
```

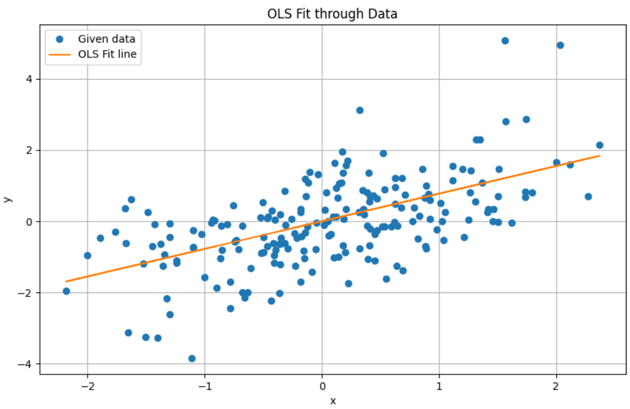

Figure 1: OLS Fit

We then define a function to calculate the log-likelihood assuming a normal distribution of errors:

$$\mathcal{L}(\sigma, \beta) = -\frac{n}{2}\ln(\sigma^2 2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\varepsilon_i - 0)^2$$

Where $\varepsilon_i = y_i - \beta x_i$

After fitting and minimizing the negative of the above formula, we find the MLE parameters β, $\mathcal{L}$, and σ to be equal to

$$\beta = 0.7690 \ \sigma = 1.008 \ \mathcal{L} = -285.289$$

These are the exact same as the OLS parameters proving the success of the log-likelihood function and optimization algorithm.

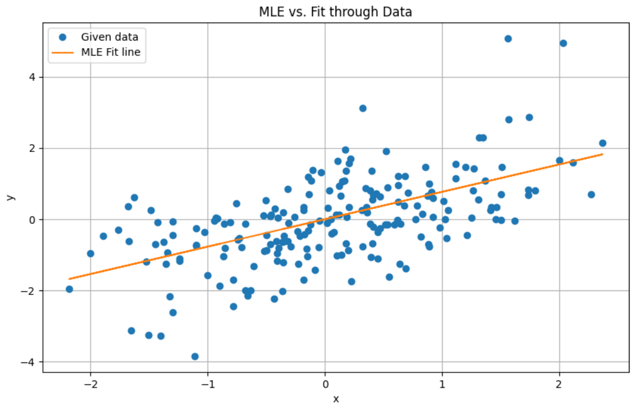

Figure 2: Normal MLE Fit

2.2 Part b

Using the same optimization function as in parts a) and c) to minimize the negative log-likelihood, we minimize the negative of the log-likelihood function defined below for a T distribution of errors.

$$\mathcal{L}(\beta) = -\sum_{i=1}^{n} \ln\left(k + (y_i - \beta x_i)^2\right).$$

The fitted $\beta = 0.7631$, $\mathcal{L}(\beta) = -1059.671$

In order to compare the fitted parameters from the MLE with normal assumption vs the MLE with T distribution assumption, we calculate the $R_s q$ measures for both using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2_{normal} = 0.346 \; R^2_t = 0.346$$

Evidently, the R^2 values for both the normal and T distribution assumptions are the same, thus in theory both fit the data equivalently well. The β values from both MLE fits are also incredibly similar to the β attained from the OLS regression fit, proving that both are good fits for the data. This makes sense given the size of the sample size of the data in question given it is above 30 and the T distribution is similar to the normal distribution of errors above this sample size threshold.

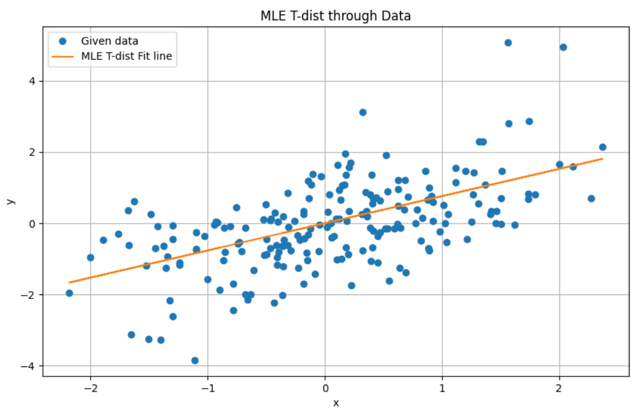

Figure 3: T distribution MLE

2.3 Part c

Using the following formula for log-likelihood for a multivariate normal distribution, we can first, estimate the parameters μ and Σ and then use these MLE parameters to calculate the conditional distribution of X2 given the observed X1 values we are given from the file problem2_x1. We can then use the conditional variance to calculate the standard deviation and then the subsequent 95% confidence interval of X2 given the conditional normal distribution for each value of X1. The initial log-likelihood formula is as follows:

$$l(\mu, \Sigma) = -\frac{mp}{2}\log(2\pi) - \frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x^{(i)} - \mu)^T \Sigma^{-1}(x^{(i)} - \mu)$$

Where μ and Σ are the mean vector and covariance matrix for the distribution respectively.

We then minimize the negative of this log-likelihood to solve for estimated mu and Σ using the `scipy.optimize` library from which we then save the parameters and then use the following formulas and observed X1 values to solve for the conditional distribution of X2 given observed X1.

$$\text{Fitted } \mu = [0.00102567 \quad 0.99022151]$$

$$\text{Fitted}\Sigma = \begin{bmatrix} 0.99693215 & 0.2735571 \\ 0.27354289 & 0.94110685 \end{bmatrix}$$

$$\text{Neg log likelihood} = 2690.987$$

Using the following formulas to calculate the conditional X2 distribution given the observed X2 values:

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

$$\sigma_{X_2|X_1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

The normal distributions of X2 given each observed X1 are as follows (please run Problem3.py for the full output):

X1	X2 Mean	X2 Variance
-0.799934199	0.7704389992334721	0.8660430799029785
0.447928241	1.1128510910720246	0.8660430799029785
1.195901414	1.3180941135369442	0.8660430799029785
0.28913008	1.0692770489779897	0.8660430799029785
0.134273328	1.0267845252125944	0.8660430799029785
-0.619003921	0.8200860700812098	0.8660430799029785
-0.028256951	0.9821863940119362	0.8660430799029785
0.140796589	1.0285745009126555	0.8660430799029785
1.62987122	1.4371749547450752	0.8660430799029785
-0.560802656	0.8360564336865924	0.8660430799029785
-1.648942809	0.5374719631080467	0.8660430799029785
0.045220933	1.0023486452007055	0.8660430799029785
-0.149366638	0.9489540280569337	0.8660430799029785
-1.291858637	0.6354554702460269	0.8660430799029785
0.910159753	1.2396869135667035	0.8660430799029785
-1.982287724	0.44600248218601934	0.8660430799029785
0.621478689	1.1604731446208099	0.8660430799029785
0.006004675	0.9915877467992276	0.8660430799029785
0.240275753	1.0558714749602038	0.8660430799029785
-0.895251137	0.7442841354584029	0.8660430799029785
	...	
2.017166308	1.5434483042665788	0.8660430799029785
-0.57220435	0.8329278212947946	0.8660430799029785

Table 1: X1 and conditional X2 Means and Variances

The plot of the expected value of X2 given observed X1 on the X-axis and the 95% confidence interval of X2 is below:

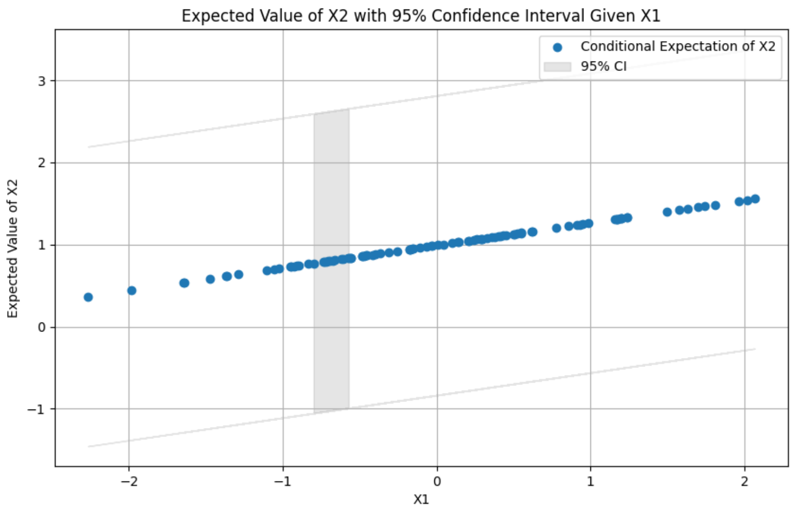

Figure 4: X2 distribution conditional on observed X1

3 Problem 3

Using the `statsmodels.api` package SARIMAX function to compute AR and MA for AR(1) through AR(3) as well as MA(1) through MA(3). The functions calculate MA and AR as follows:

$$MA(1) = y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$$
$$MA(2) = y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$$
$$MA(3) = y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \theta_3\varepsilon_{t-3}$$

$$AR(1) = y_t = c + \phi_1 y_{t-1} + \varepsilon_t$$
$$AR(2) = y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$
$$AR(3) = y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \varepsilon_t$$

The function outputs are as follows.
AR 1

Dep. Variable:	x	**No. Observations:**	500
Model:	SARIMAX(1, 0, 0)	**Log Likelihood**	-819.328
Date:	Fri, 26 Jan 2024	**AIC**	1644.656
Time:	19:22:16	**BIC**	1657.299
Sample:	0	**HQIC**	1649.617
	- 500		
Covariance Type:	opg		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
intercept	1.6965	0.110	15.461	0.000	1.481	1.912
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

MA 1

Dep. Variable:	x	**No. Observations:**	500
Model:	SARIMAX(0, 0, 1)	**Log Likelihood**	-780.702
Date:	Fri, 26 Jan 2024	**AIC**	1567.404
Time:	19:22:16	**BIC**	1580.047
Sample:	0	**HQIC**	1572.365
	- 500		
Covariance Type:	opg		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
intercept	2.1236	0.085	25.028	0.000	1.957	2.290
ma.L1	0.6434	0.034	18.846	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

AR 2

Dep. Variable:	x	**No. Observations:**	500
Model:	SARIMAX(2, 0, 0)	**Log Likelihood**	-786.540
Date:	Fri, 26 Jan 2024	**AIC**	1581.079
Time:	19:22:16	**BIC**	1597.938
Sample:	0	**HQIC**	1587.694
	- 500		
Covariance Type:	opg		

	coef	std err	z	P> |z|	[0.025	0.975]
intercept	2.2914	0.126	18.249	0.000	2.045	2.537
ar.L1	0.2732	0.042	6.487	0.000	0.191	0.356
ar.L2	-0.3505	0.043	-8.067	0.000	-0.436	-0.265
sigma2	1.3602	0.094	14.455	0.000	1.176	1.545

Ljung-Box (L1) (Q):	15.50	**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

MA 2

Dep. Variable:	x	**No. Observations:**	500
Model:	SARIMAX(0, 0, 2)	**Log Likelihood**	-764.971
Date:	Fri, 26 Jan 2024	**AIC**	1537.941
Time:	19:22:16	**BIC**	1554.800
Sample:	0	**HQIC**	1544.556
	- 500		
Covariance Type:	opg		

	coef	std err	z	P> |z|	[0.025	0.975]
intercept	2.1255	0.060	35.199	0.000	2.007	2.244
ma.L1	0.4345	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

AR 3

Dep. Variable:	x	**No. Observations:**	500
Model:	SARIMAX(3, 0, 0)	**Log Likelihood**	-713.330
Date:	Fri, 26 Jan 2024	**AIC**	1436.660
Time:	19:22:16	**BIC**	1457.733
Sample:	0	**HQIC**	1444.929
	- 500		
Covariance Type:	opg		

	coef	std err	z	P> |z|	[0.025	0.975]
intercept	1.1291	0.142	7.963	0.000	0.851	1.407
ar.L1	0.4515	0.040	11.180	0.000	0.372	0.531
ar.L2	-0.4887	0.037	-13.104	0.000	-0.562	-0.416
ar.L3	0.5048	0.040	12.770	0.000	0.427	0.582
sigma2	1.0131	0.068	14.940	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 3

Dep. Variable:	x	**No. Observations:**	500
Model:	SARIMAX(0, 0, 3)	**Log Likelihood**	-763.434
Date:	Fri, 26 Jan 2024	**AIC**	1536.868
Time:	19:22:16	**BIC**	1557.941
Sample:	0	**HQIC**	1545.137
	- 500		
Covariance Type:	opg		

	coef	std err	z	P> |z|	[0.025	0.975]
intercept	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

In order to assess the best fit out of AR and MA, we can use AIC and BIC of each model. The corresponding AICs and BICs are below. Each array begins with AR(1) value, then MA(1), then AR(2), then MA(2), etc.

AIC values:

1644.6555051096989
1567.4036263758758
1581.0792678834373
1537.9412062214383
1436.6598065779694
1536.8677087301505

BIC values:

1657.2993294049654
1580.0474506711423
1597.9377002771262
1554.7996386151272
1457.7328470700804
1557.9407492222615

Clearly AR3 is best of fit as it has both the lowest AIC and BIC.