Problem 1 (3pts)

Given the dataset in problem1.csv:

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

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

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

Answer:

Moments	Biased	Python	Unbiased (Calculated by complex formula)
Mean	1.0490	1.0490	1.0490
Variance	5.4218	5.4218	5.4272
Skewness	11.1173	11.1173	11.1507
Kurtosis	767.8841	767.8841	767.8886

The data calculated in the first column is based on the formulas: kurtosis=$\sum(x-\bar{x})^4/n$, skewness=$\sum(x-\bar{x})^3/n$, and Variance=$\sum(x-\bar{x})^2/n$, where these are biased estimates.

The second column represents the results obtained through Python using the built-in moments package. It can be observed that the results match exactly with the first column, indicating that the moment's package in Python produces biased estimates. However, the error caused by bias is very small and can almost be ignored.

The last column displays the unbiased results calculated using the unbiased formulas provided in the lecture's PDF as follows. The calculation process can be viewed in Excel.

$$\mathcal{K}(\mathbf{x}) = \sum_i (x_i - \mu)^4/n.$$

$$\begin{aligned}
\sigma^2(\bar{\mathbf{x}}) &= \frac{\sigma^2(x)}{n} \\
\mathcal{S}(\bar{\mathbf{x}}) &= \frac{\mathcal{S}(x)}{n^2} \\
\mathcal{K}(\bar{\mathbf{x}}) &= \frac{\mathcal{K}(x)+3(n-1)\sigma^4(x)}{n^3}
\end{aligned}$$

$$\begin{aligned}
\sigma^2_{\bar{\mathbf{x}}}(\mathbf{x}) &= \tfrac{n-1}{n}\sigma^2(\mathbf{x}) \\
&= \sigma^2(\mathbf{x}) - \sigma^2(\bar{\mathbf{x}}) \\
\mathcal{S}_{\bar{\mathbf{x}}}(\mathbf{x}) &= \tfrac{(n-2)(n-1)}{n^2}\mathcal{S}(\mathbf{x}) \\
&= \mathcal{S}(\mathbf{x}) - (3n-2)\mathcal{S}(\bar{\mathbf{x}}) \\
\mathcal{K}_{\bar{\mathbf{x}}}(\mathbf{x}) &= \tfrac{(n-1)}{n^3}\left((n^2-3n+3)\mathcal{K}(\mathbf{x}) + (6n-9)\sigma^4(\mathbf{x})\right) \\
&= \tfrac{(n-2)(n-2)}{n^2}\mathcal{K}(\mathbf{x}) + (2n-3)\mathcal{K}(\bar{\mathbf{x}})
\end{aligned}$$

$$\hat{\mathcal{K}}(\mathbf{x}) = \frac{n^2}{(n-1)^3(n^2-3n+3)}\left[[(n(n-1)^2 + (6n-9))]\,\mathcal{K}_{\bar{\mathbf{x}}}(\mathbf{x}) - n(6n-9)\sigma^4_{\bar{\mathbf{x}}}(\mathbf{x})\right]$$

Problem 2 (5pts)

Assume the multiple linear regression model $YY = XXXX + \epsilon\epsilon$, *wwheeeeee* $YY \in RRnn$,$XX \in RRnn \times pp$, $\beta\beta \in RRpp$, $\epsilon\epsilon \in RRNN$:

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.

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?

c. Fit the data in problem2_x.csv using MLE given $XX = [XX1, XX2]$ follows the multivariate normal distribution. Assume X as a random variable, follows the fitted gaussian distribution, $XX1$ (problem2_x1.csv) are a part of observed value of X, What's the distribution of $XX2$ given each observed value? Plot the expected value along with the 95% confidence interval.

$$\mathcal{K}(\mathbf{x}) = \sum_i (x_i - \mu)^4/n.$$

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

Answer:

a.

0	1	2	3	4	5	6
	coef	std err	t	P>|t|	[0.025	0.975]
constant	-0.0874	0.071	-1.222	0.223	-0.228	0.054
x	0.7753	0.076	10.226	0.000	0.626	0.925

Parameter	MLE Estimate
Intercept	-0.08738448066024189
Slope	0.7752741349919541
Standard Deviation	1.0037563087151677

The two tables above represent the results of fitting the given data using OLS and MLE, respectively. It can be observed that the β values obtained from both methods are very close, if not identical. However, there is a significant difference in the standard deviation of errors between OLS (0.076) and MLE (approximately 1). The standard deviation from MLE is notably larger than that from OLS.

b.

T-distribution Table:

Metric	Value
MLE estimate of mean	0.003196397390137874
MLE estimate of standard deviation	0.9300936862556104
MLE estimate of degrees of freedom	7.050913044430027
Log Likelihood	-597.1258716868359
AIC	1200.2517433736718
BIC	1210.146695473316

The plot above shows the fitting results for the T-distribution, where the estimated mean is 0.003, the standard deviation is 0.93, and the AIC is around 1200.

Norm Table:

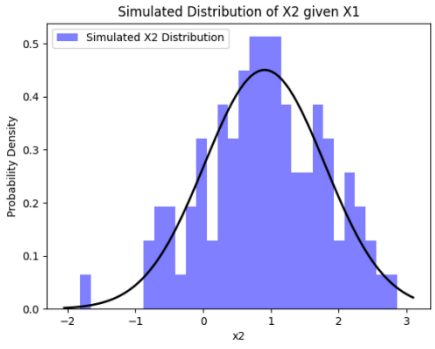
Metric	Value
MLE estimated standard deviation	1.2408149341330212
Log Likelihood	-326.94226434283473
AIC	657.8845286856695
BIC	664.4811634187655

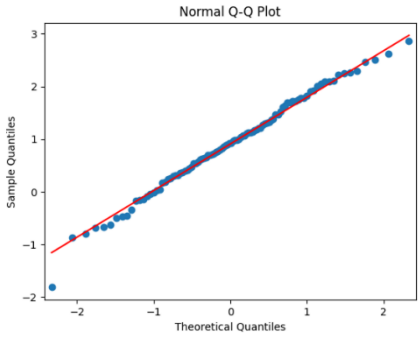
T-distribution Table:

Metric	Value
MLE estimate of standard deviation	0.9300936862556104
Log Likelihood	-597.1258716868359
AIC	1200.2517433736718
BIC	1210.146695473316

The Normal distribution exhibits lower absolute values of Log Likelihood, AIC, and BIC. Therefore, I believe that Maximum Likelihood Estimation (MLE) under the normality assumption provides a better fit.

c.





In the model, I noticed that X1 and X2 are interchangeable, and they are mathematically symmetric. Therefore, I believe that X1 and X2 should follow the same distribution. Thus, I argue that X2 also follows a Gaussian distribution, and the proof is as follows:

Firstly, for each observed value of x1 in x1_observed, calculate the conditional mean of x2 given x1: Then, compute the conditional mean_x2, representing the expected value of x2 given x1, using the mean vector and covariance matrix information. Generate simulated values of x2 from a normal distribution using the calculated conditional mean and conditional standard deviation (square root of variance). Lastly, perform hypothesis tests to examine whether x2 follows a Gaussian distribution.

From the above plot, it can be observed that X2 approximately follows a normal distribution, and this observation is further confirmed by the QQ plot.

Shapiro-Wilk Test Statistic value is 0.99, which supports the assertion that X2 should follow a Gaussian distribution. What's more, the p-value is 0.83, which is much more than 0.05, telling us that we cannot reject the null hypothesis that X2 follows a Gaussian distribution. As a result, the distribution of $XX2$ given each observed value is Gaussian distribution.

d.

I believe the purpose of this question is to solve the formula mentioned in class.

$$ll = -\frac{n}{2}ln\left(\sigma^2 2\pi\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\epsilon_i - 0\right)^2$$

The regression model is defined by the equation:

$Y = X\beta + \varepsilon$

The likelihood function $L(\beta, \sigma^2)$ for a linear regression model with normally distributed errors is the product of individual probability densities:

$L(\beta, \sigma^2) = \prod (1/\sqrt{(2\pi\sigma^2)}) * \exp(-(y_i - x_i'\beta)^2 / (2\sigma^2))$

The Log-Likelihood Function

The log-likelihood function $\ell(\beta, \sigma^2)$ simplifies to:

$$\ell(\beta, \sigma^2; y, X) = \ln(L(\beta, \sigma^2; y, X))$$

$$= \ln((2\pi\sigma^2)^{(-N/2)} \exp(-1/(2\sigma^2) \Sigma (y_i - x_i'\beta)^2))$$

$$= \ln((2\pi\sigma^2)^{(-N/2)}) + \ln(\exp(-1/(2\sigma^2) \Sigma (y_i - x_i'\beta)^2))$$

$$= -N/2 \ln(2\pi\sigma^2) - 1/(2\sigma^2) \Sigma (y_i - x_i'\beta)^2$$

$$= -N/2 \ln(2\pi) - N/2 \ln(\sigma^2) - 1/(2\sigma^2) \Sigma (y_i - x_i'\beta)^2$$

$$\ell(\beta, \sigma^2) = -n/2 \log(2\pi) - n \log(\sigma) - (1/2\sigma^2) \Sigma (y_i - x_i'\beta)^2$$

The Maximum Likelihood Estimators

The MLEs for β and σ^2 are obtained by maximizing the log-likelihood function. They are:

$\hat{\beta}_{MLE} = (X'X)^{-1}X'Y$ (Same as OLS as stated in class PDF)

$\hat{\sigma^2}_{MLE} = (1/n) * \Sigma (y_i - x_i'\hat{\beta}_{MLE})^2$

proof:

$\max_{\beta, \sigma 2} \ell(\beta, \sigma 2; y, X)$

Let the partial derivatives of the above equation for β and variance be 0 respectively.

$$\nabla\beta\ \ell(\beta,\sigma^2;y,X)$$

$$= \nabla\beta\ (-N/2 \ln(2\pi) - N/2 \ln(\sigma^2) - 1/(2\sigma^2) \sum (y_i - x_i'\beta)^2)$$

$$= -1/\sigma^2 \sum x_i(y_i - x_i'\beta)$$

$$= -1/\sigma^2 (\sum x_i y_i - \sum x_i x_i'\beta)$$

which is equal to zero only if

$$\sum x_i y_i - \sum x_i x_i' \beta = 0$$

Therefore, the first of the two equations is satisfied if

$$\beta = (\sum x_i x_i')^{-1} \sum x_i y_i = (X'X)^{-1} X'y$$

$$\partial/\partial\sigma^2 \, \ell(\beta,\sigma^2;y,X)$$

$$= \partial/\partial\sigma^2 \, (-N/2 \ln(2\pi) - N/2 \ln(\sigma^2) - 1/(2\sigma^2) \sum (y_i - x_i'\beta)^2)$$

$$= -N/2\sigma^2 \, [\partial/\partial\sigma^2 \, (1/\sigma^2 \sum (y_i - x_i'\beta)^2)]$$

$$= -N/2\sigma^2 \, [\sum (y_i - x_i'\beta)^2 \, (-1/(\sigma^2)^2)]$$

$$= -N/2\sigma^2 + 1/(2\sigma^4) \sum (y_i - x_i'\beta)^2$$

which, if we assume $\sigma^2 \neq 0$, is equal to zero only if

$$\sigma^2 = 1/N \sum (y_i - x_i'\beta)^2$$

Problem 3 (2pts)

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

Answer:

For fitting the data into problem3.csv using ARIMA analysis in statsmodels.API package in Python, we can get the result of the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) for the six simulations, respectively.

Models	AIC	BIC

AR (1)	1644.66	1657.30
AR (2)	1581.08	1597.94
AR (3)	**1436.66**	**1457.73**
MA (1)	1567.40	1580.05
MA (2)	1537.94	1554.80
MA (3)	1536.87	1557.94

As a result, AR (3) has the least AIC and BIC, so AR (3) is the most fitting model for the given data.