Dean&Francis
ISSN 2959-6157

Simple Linear Regression Analysis with Non-Normal Errors

Yinhao Wang

Nanjing Foreign Language School, Nanjing, Jiangsu 210006, China.

Abstract:

We consider a simple linear regression with non-normal errors; specifically, errors follow the double exponential distribution and uniform distribution. Maximum likelihood estimates for parameters (intercept and slope) are investigated. Although they do not have closed mathematical forms, they can be derived uniquely by numerical methods. Through Monte Carlo simulation, ordinary least square estimates (LSE) and maximum likelihood estimates (MLE) are derived and compared through their biases, variances and mean squared errors. Simulation studies show that both MLE and LSE follow normal distributions and MLE has smaller mean squared errors compared to LSE. This paper suggests or confirms that one prefers to use MLE to estimate unknown parameters if there exists strong evidence that errors do not follow a normal distribution.

Keywords: Linear regression, MLE, LSE, double exponential distribution, Monte-Carlo simulation, MSE (Mean Squared Error)

1 Introduction

1.1 Background

Linear regression analysis is widely used to help predict and interpret data and support scientific research and decision making by modeling linear relationships between independent and dependent variables.

Linear regression usually assumes that the error term follows a normal distribution, which may simplify statistical inference. However, parameter estimates and predictions may be biased when the error term distribution deviates from normal. Therefore, to improve model accuracy appropriate error term distribution assumptions are crucial.

Traditionally, Linear Least Squares Estimation is used for estimating parameters for the fit line of datasets, it has evident advantages over other estimation devices due to its simplicity and practicality. Thompson and Zeckhauser (1970) praised its "unassailable advantage" in saving computer cost when dealing with complex regression problems. However, computer cost is no longer the dominant factor in evaluating regression methods. This makes the problem of least squares estimation more prominent, that is, it can only obtain more accurate results when the error term is close to the normal distribution. However, in real-world situations, error terms do not always follow normal distributions; for instance, when processing datasets where there is a suspicion

ISSN 2959-6157

of heterogeneity within the population, and the data indicate substantial discrepancies--for example, Hsu (1979) mentioned a situation of aircraft mid-air collision risk -- Laplace distribution provides a better fit for the tail region of the data (Reed, 2006); the uniform distribution is used to model the randomness of service requests over a fixed interval (Jerrum & Valiant & Vazirani, 1986), while the Laplace distribution can be applied in financial risk management due to its heavy tails (Kozubowski & Podgórski, 2001).

1.2 Literature Review

Let $k: R \times R \to R$ be a binary function. Zeckhauser and Thompson (1970) considers

$$y_i = a + bx_i + \epsilon_i, \epsilon_i \ f(z, \mu, \sigma, \theta) = k(\sigma, \theta) \bullet e^{-\left|\frac{z-\mu}{\sigma}\right|^{\theta}}, i = 1, \ldots, n, \quad (1)$$

where $a, b \in R$ are multiplier and bias, while μ and σ are mean value and variance, respectively. Although the error distribution in their research is flexible, the estimation of a, b heavily depends on the estimation of θ; the MLE algorithm relies on a recursive approach, which means that the determination of the parameter θ could potentially fail to converge or may not achieve a sufficiently precise outcome.

Dean and King (2009) further explore the possibility that the error distribution adheres to a highly adaptable distribution known as the Generalized Lambda Distribution, characterized by four distinct parameters. Once more, the applicability of these distributions is contingent upon the specific ranges of the parameters involved. Moreover, the process of deriving MLE is influenced by the initial parameter values and is susceptible to issues of non-convergence.

In this project, we consider errors follow two specific distributions, i.e. double exponential and uniform. In these two particular cases, we are able to derive MLE respectively. Through Monte Carlo simulation, we are able to approximate the asymptotic normal distribution of MLE; compare variance, bias, and mean square errors of two estimators (LSE & MLE).

2 Method

When $\theta = 2$, this section will give the proof that MLE and LSE yield the same parameter estimations under the following assumption of normally distributed error terms.

Assumption Given the sample dataset x_i, y_i, $i = 1, 2, 3, \ldots, n$, and assuming the linear relationship

$$y_i = a + bx_i + \epsilon_i, i = 1, \ldots, n,$$

where a and b are parameters to be estimated and $\epsilon_i \ N(0, \sigma^2)$, $i = 1, \ldots, n$, are error terms.

2.1 Maximum Likelihood Estimation

Under the assumption of normality, the probability density function of the error terms ϵ_i, $i = 1, \ldots, n$, are given by

$$\epsilon_i \sim f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{x^2}{2\sigma^2}\right), i = 1, \ldots, n. \quad (2)$$

From (1) and (2), we have

$$\epsilon_i = y_i - (a + bx_i), i = 1, \ldots, n,$$

and

$$Lik(a, b, \sigma^2) ? \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y_i - a - bx_i)^2}{2\sigma^2}\right),$$

where the likelihood function $Lik: R \times R \times R_+ \to R$ quantifies the observed data.

Taking the natural logarithm of the likelihood function, we get the log-likelihood function

$$l(a, b, \sigma^2) ? logLik(a, b, \sigma^2) = \sum_{i=1}^{n} -\frac{1}{2} log(2\pi\sigma^2) - \frac{(y_i - a - bx_i)^2}{2\sigma^2}.$$

Next, we are able to derive the following optimal conditions

$$\begin{cases} \frac{\partial logLik(a, b, \sigma^2)}{\partial a} = \frac{1}{\sigma^2} \sum_{i=1}^{n} y_i - a - bx_i = 0 \\ \frac{\partial logLik(a, b, \sigma^2)}{\partial b} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0 \end{cases} \quad (3)$$

Thus, we get $\hat{a} = \bar{y} - \hat{b}\bar{x}$, $\hat{b} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$, where $\bar{x}$ and $\bar{y}$, respectively, are sample means from the data.

2.2 Least Squares Estimation

We have to minimize the sum of squared residuals: $\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$. Taking partial derivatives with respect to a and b and setting them to zero, we have

$$\begin{cases} \frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)^2}{\partial a} = -2 \sum_{i=1}^{n} y_i - a - bx_i = 0 \\ \frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)^2}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0 \end{cases}$$

$$\begin{cases} \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0 \end{cases}$$

according to (3). Thus, we again get $\hat{a} = \bar{y} - \hat{b}\bar{x}$,

$$\hat{b} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.$$

Whatever error distribution it is dealing with, the formula of Linear LSE does not change, and thus, does not reflect the information of error distribution.

When facing with Laplace distributed error terms $\epsilon_i\ f(x) = \frac{1}{2\sigma}exp\left(-\frac{|x|}{\sigma}\right)$, MLE considers the problem of distribution type. We give the likelihood function

$$Lik(a,b,\sigma)\prod_{i=1}^{n}\frac{1}{2\sigma}exp\left(-\frac{|y_i - a - bx_i|}{\sigma}\right) = \frac{1}{2^n\sigma^n}exp\left(-\frac{\sum_{i=1}^{n}|y_i - a - bx_i|}{\sigma}\right).$$

To minimize $\sum_{i=1}^{n}|y_i - a - bx_i|$, we calculate the partial derivatives of the Log-Likelihood function and its derivatives as

$$\begin{cases} \frac{\partial logLik(a,b,\sigma^2)}{\partial a} = \sum_{i=1}^{n} -sgn(y_i - a - bx_i) = 0 \\ \frac{\partial logLik(a,b,\sigma^2)}{\partial b} = \sum_{i=1}^{n} -x_i sgn(y_i - a - bx_i) = 0 \end{cases}, \quad (4)$$

where $sgn: R \to \{\pm 1\}$ is the sign function.

Denoting the indicator function by 1 and using equations in (4), we derive the expression of a as

$$\tilde{a} := \min\left\{a \in R : \sum_{i=1}^{n} -\frac{1(\{y_i - a - bx_i \leqslant 0\})}{n} \geqslant \frac{1}{2}\right\}$$

which means that a is the median of $\{y_1 - bx_1, y_2 - bx_2, \ldots, y_n - bx_n\}$. Next, try a series of values $b_j, j = 1,\ldots,k,$ and obtain a_j for each. Ultimately, the pair of (a_j, b_j) that makes $Lik(\sigma)$ be the minimal, which is equivalent to making $\sum_{i=1}^{n}|y_i - a - bx_i|$ be the minimal. That is the required (a,b).

Likewise, in the cases when

$$\epsilon_i \sim f(\theta) = \begin{cases} \frac{1}{2\theta}, if\ -\theta \leqslant x \leqslant \theta \\ 0, else \end{cases}, i = 1,\ldots,n,$$

we derive the likelihood function

$$Lik(\theta) = \frac{1}{2^n\theta^n}1\left(\theta \geqslant \max_{a,b}\{|y_i - a - bx_i|\}\right)$$

and subsequently get

$$\frac{\partial logLik(\theta)}{\partial\theta} = -\frac{n}{\theta}\left(\theta \geqslant \max_{a,b}\{|y_i - a - bx_i|\}\right) \quad (5)$$

To force (5) to be negative, the estimation will be

$$\hat{\theta}(a,b)? \max_{a,b}\{|y_i - a - bx_i|\}$$

To minimize it, we only need to focus on $y_i - a - bx_i$ and the conclusion that $\hat{a}$ is the middle point of $y_i - bx_i$, which is discovered with ease. As a result,

$$\hat{a} = \frac{\max_i\{y_i - bx_i\} + \min_i\{y_i - bx_i\}}{2}.$$

Again, a series of values $b_j, j = 1,\ldots,k$, are tried and each element has a correspondent $\hat{a}_j$; ultimately, a pair of best-fit (a_j, b_j) is obtained.

Remark It is not difficult to observe from the proof that to estimate the parameters, MLE provides a more flexible alternative to LSE by maximizing the likelihood function for each error distribution situation investigated, and thus, is more robust in dealing with non-normal error terms.

Specifically, MLE may outperform LSE when the error term follows a continuous uniform distribution or a Laplace distribution, which we will explore further in this paper.

2.3 Experimental Preparation

We use R language to obtain the required datasets, and then compute biases, variances, and mean squared errors for these datasets respectively. In this paper, the scale parameter for the error distributions (σ) and sample size number are set as variables. They are compared so as to observe how the two estimators perform relatively in different cases.

Here, we use computed statistics (in cases of $samplesize = 20,1000,5000$; $log_2\sigma = -2,-1,0,1,2$) to make a chart to evaluate the efficacy of the two estimators. Bias(A) and MSE(A) changes are specially graphed using Mathematica.

To better explore the performance of the two estimators

ISSN 2959-6157

under the demanded circumstances, we use R as the tool to perform the Monte-Carlo Simulation so as to generate the error term of each sample point. At the inception, an initial line was set (in this project $y = 2 + 5x$)

To precisely monitor the performance of the two estimators across various scenarios, the simulation is conducted 10,000 times. A simulated independent variable x is created, which is a sequence ranging from 1 to n, normalized by division with n. The number of simulation, K, is then defined, and arrays to store the results from LSE and MLE are initialized.

During each simulation, the response variable y is generated by incorporating random error terms that follow a double exponential distribution. The lm function in R is utilized to fit the linear model y~x, yielding the estimated coefficients which are stored in A_lse and B_lse. This process estimates model parameters by minimizing the squared sum of the residuals.

For the MLE part, the code initially takes the slope b0 derived from LSE as a starting point, computes its standard error se, and based on this, creates a series of potential values for both the intercept and the slope. For each possible slope value, the corresponding intercept is determined, and the model's fit is assessed by evaluating the likelihood function. The slope and intercept that result in the smallest objective function are chosen as the MLE outcomes and are stored in B_mle and A_mle.

3 Experiment Result

3.1 Uniform cases

σ	BIAS				VARIANCE				MSE			
	A		B		A		B		A		B	
	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE
2^{-2}	-0.0001419889	-0.0004941876	0.000640312	0.000482701	0.0044984	0.0026597	0.0125032	0.007693	0.0044984	0.0026599	0.0125037	0.0076929
2^{-1}	-0.0007140949	-0.0007429484	0.000848528	0.000911043	0.0176947	0.0106625	0.0493356	0.030794	0.0176952	0.010663	0.0493363	0.030795
2^{0}	0.0004520305	-0.001815484	-0.000707107	0.00221359	0.071284	0.042648	0.1959159	0.124596	0.0712842	0.0426513	0.1959164	0.1246009
2^{1}	0.003020884	0.008021094	-0.00209762	-0.0156493	0.2883108	0.1775068	0.7971727	0.496911	0.2883199	0.1775711	0.7971771	0.4971561
2^{2}	-0.002350213	-0.009402379	0.00979796	0.019105	1.169844	0.6864107	3.245199	1.980249	1.169849	0.6864991	3.245295	1.980614
2^{3}	0.05137556	-0.00572429	-0.0834865	0.0113578	4.628126	2.888415	12.8474	8.328704	4.630765	2.888447	12.85437	8.328833

Fig 1. Comparison of the two estimators in uniform cases (sample size: 20).

σ	BIAS				VARIANCE				MSE			
	A		B		A		B		A		B	
	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE
2^{-2}	-5.04E-05	4.27E-06	0.00002	-0.000034641	8.18373E-05	1.266936E-06	0.000245262	4.019629E-06	8.183985E-05	1.266954E-06	0.000245262	4.019641E-06
2^{-1}	0.0001029	-1.34E-07	-5.47723E-05	1.78885E-05	0.000338459	5.517151E-06	0.000997389	1.728660E-05	0.00033847	5.517151E-06	0.000997392	1.728692E-05
2^{0}	0.0006318	2.60E-05	-0.00156301	-1.14018E-05	0.00133721	2.167501E-05	0.003990476	6.716280E-05	0.00133761	2.167569E-05	0.003992919	6.716293E-05
2^{1}	0.0006591	7.4528E-05	-0.000173205	-0.00015748	0.00536117	8.913126E-05	0.01600327	0.0002855865	0.005361606	8.913681E-05	0.0160033	0.0002856113
2^{2}	0.0010058	0.0002557	-0.000244949	-0.000244949	0.02112921	0.000348317	0.06368872	0.0010934560	0.02113022	0.000348382	0.06368878	0.001093476
2^{3}	-0.004030	-0.0003460	0.00704273	0.000884308	0.08509121	0.001350766	0.2553991	0.0043536080	0.08510745	0.001350886	0.2554487	0.00435439

Fig 2. Comparison of the two estimators in uniform cases (sample size: 1000).

σ	BIAS				VARIANCE				MSE			
	A		B		A		B		A		B	
	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE
2^{-2}	7.307784E-05	5.422667E-06	-0.000137004	-8.11172E-06	1.668399E-05	5.390501E-08	4.943747E-05	1.70407E-07	1.66893E-05	5.393441E-08	4.945624E-05	1.704727E-07
2^{-1}	-0.000154686	4.05465E-06	0.000237697	-8.01873E-06	6.527369E-05	2.126983E-07	0.0001977569	6.89185E-07	6.52976E-05	2.127147E-07	0.0001978134	6.892489E-07
2^{0}	0.000719785	-1.574001E-05	-0.000211896	1.17473E-05	0.000274347	9.070323E-07	0.0008206387	2.9387E-06	0.000274348	9.072801E-07	0.0008206836	2.938834E-06
2^{1}	1.573897E-05	1.241525E-05	-0.00093755	-2.19089E-05	0.001059623	4.050647E-06	0.003181416	1.31811E-05	0.001060141	4.050801E-06	0.003182295	1.318154E-05
2^{2}	-4.050147E-05	-5.048681E-05	0.000316228	7.13442E-05	0.004178617	1.479810E-05	0.01275556	4.70116E-05	0.004178618	1.480065E-05	0.01275566	4.701672E-05
2^{3}	-0.0005266664	6.193088E-05	0.00151658	-9.21954E-05	0.017286950	5.527038E-05	0.0521024	0.000180417	0.01728723	5.527422E-05	0.05210470	0.0001804255

Fig3. Comparison of the two estimators in uniform cases (sample size: 5000).

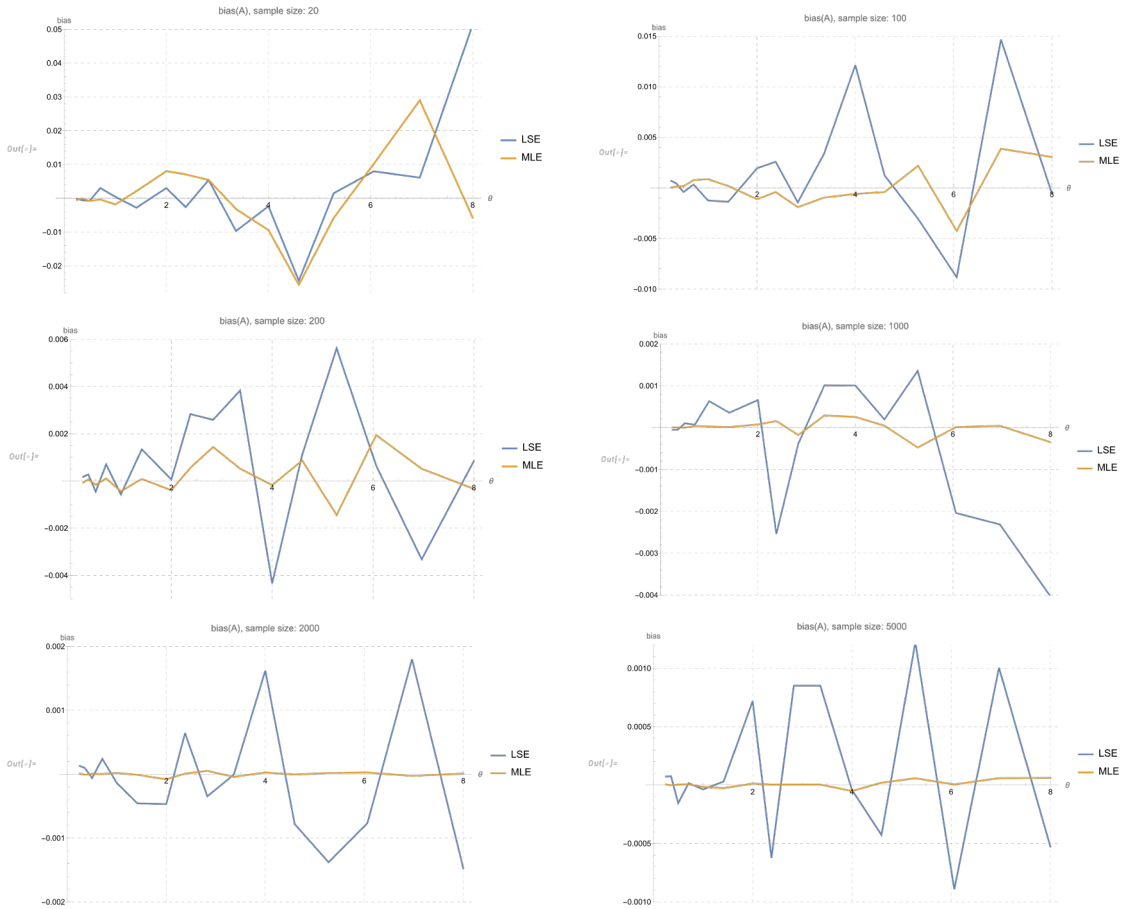

Fig 4. "Bias(A) of estimator– σ " plot in uniform cases.

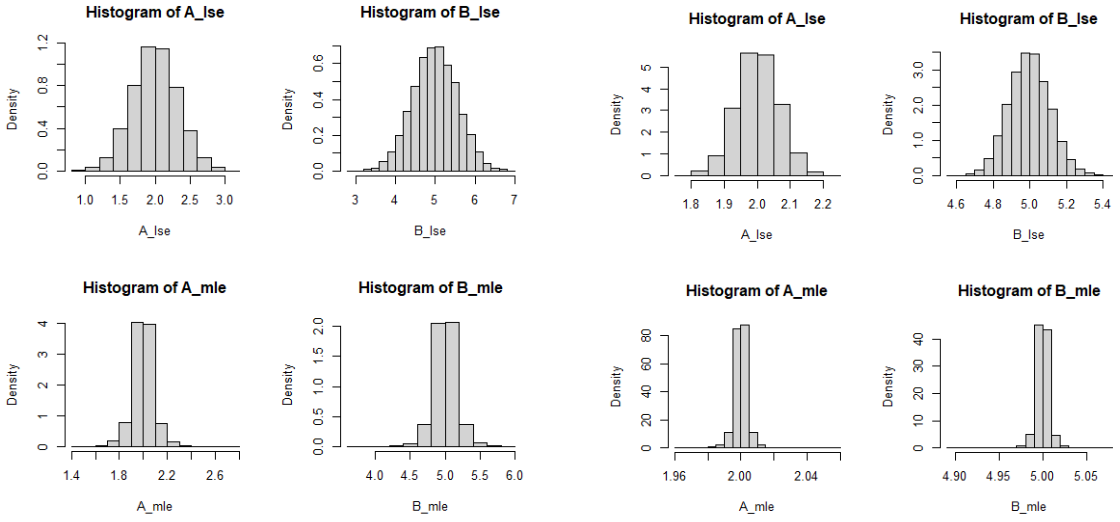

Fig 5. Histogram for the estimated parameters for uniform cases obtained each time ($\theta = 4$). Left: sample size=200, Right: sample size= 5000.

In the case of small sample sizes (e.g. 20), we observe that the bias of LSE is usually smaller, but the variance is larger. This means that the LSE may be closer to the true parameter value on average, but there is a higher uncertainty in individual estimates. In contrast, MLE may show a larger bias but a smaller variance, indicating that although the systematic bias of MLE is relatively larger when facing a limited object point to be investigated, its stability of

ISSN 2959-6157

accuracy for different sets of statistics is prominently high. As the number of sample points increases, the bias of both LSE and MLE decreases. However, MLE shows a more significant reduction, indicating that its accuracy is improving more rapidly; in fact, when the sample size is between 20~100, MLE starts to outperform LSE, thus indicating that when dealing with cases with a sample size bigger than 100, MLE can definitely be the better choice, no matter viewing by the criteria of accuracy or stability. The variance of LSE and MLE both decrease when sample size increases, but at a slightly different rate (MLE is faster than LSE). Given that in all the cases investigated, MLE derives smaller variances of estimated parameters, it is indicated that MLE has better stability. At the same time, since MSE is the sum of the square of the bias and the variance, and the square of the bias is usually much smaller than the variance, the main factor affecting the MSE is the variance, which makes the MSE of MLE more prominent, indicating that it performs better in terms of overall performance.

Overall, the accuracy of LSE seems to outperform that of MLE when sample size number is small, but when dealing with a set of data with a large number of sample points, MLE excels over LSE prominently. However, due to the minimal $bias^2$, the MSE of the two estimators mainly depends on the variance—as shown previously in the text and charts—making stability of one particular estimator the determinative indicator of evaluation.

3.2 Laplace Cases

σ	Bias				Variance				MSE			
	A		B		A		B		A		B	
	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE
2^{-2}	7.85765E-05	0.001689843	-0.001115053	-0.003969314	0.02642013	0.02222235	0.0740465	0.06337974	0.02642013	0.0222252	0.07404775	0.0633955
2^{-1}	-0.006589996	-0.001847484	0.007524676	-0.0004061793	0.1060807	0.08843186	0.2940259	0.2502695	0.1061241	0.0884353	0.2940825	0.2502697
2^{0}	0.005343677	0.01588941	-0.00924476	-0.02712896	0.4322481	0.3653148	1.207559	1.058717	0.4322767	0.3655672	1.207644	1.059453
2^{1}	-0.005133515	0.03152136	-0.01572188	-0.06343546	1.752605	1.485159	4.824152	4.321616	1.752631	1.486153	4.824399	4.32564
2^{2}	-0.01280649	0.05164715	0.04982093	-0.07529043	6.875537	5.687932	19.18482	16.44741	6.875701	5.690599	19.1873	16.45308
2^{3}	0.02347491	0.1578303	-0.1331458	-0.3232366	26.76347	23.14805	75.40058	66.88787	26.76403	23.17297	75.41831	66.99235

Fig 6. Comparison of the two estimators in Laplace cases (sample size: 20).

σ	Bias				Variance				MSE			
	A		B		A		B		A		B	
	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE
2^{-2}	-0.0002032391	0.0001874241	0.0002418601	-0.0001827662	0.0005031517	0.0002763357	0.001518850	0.0008329396	0.0005031930	0.0002763708	0.001518909	0.0008329731
2^{-1}	0.0009351535	0.0002543145	-0.001228191	-0.0001168355	0.001946199	0.001068071	0.005801307	0.003170071	0.001947074	0.001068136	0.005802815	0.003170084
2^{0}	0.00128793	0.0002170219	-0.002142019	-0.00135207	0.007994147	0.004358322	0.02420829	0.0129625	0.007995806	0.004358369	0.02421288	0.01296433
2^{1}	-0.002222316	-0.001523082	0.002624755	0.00220201	0.03179685	0.01744064	0.09467763	0.05189007	0.03180179	0.01744296	0.09468452	0.05189492
2^{2}	0.002046905	0.001102511	-0.006799861	-0.004212966	0.1284845	0.06786932	0.3836157	0.2033033	0.1284887	0.06787054	0.3836619	0.2033210
2^{3}	-0.003114378	-0.00625726	0.004421381	0.0099712	0.5221659	0.2766615	1.57593	0.8402628	0.5221756	0.2767006	1.575949	0.8403623

Fig 7. Comparison of the two estimators in Laplace cases (sample size: 1000).

σ	Bias				Variance				MSE			
	A		B		A		B		A		B	
	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE	LSE	MLE
2^{-2}	0.0001269946	0.0001016524	-0.0001910655	-0.0001792166	0.0001023544	5.237414E-05	0.0003054025	0.0001564762	0.0001023706	5.23845E-05	0.000305439	0.0001565083
2^{-1}	-0.000176447	4.271938E-05	0.0002120853	-0.0001627974	0.00040199	0.0002071532	0.001213748	0.0006298134	0.0004020211	0.000207155	0.001213793	0.0006298399
2^{0}	0.000355838	0.0002512507	-0.0006756817	-0.0003436128	0.001582766	0.0008258649	0.004722178	0.002486714	0.001582893	0.000825928	0.004722634	0.002486832
2^{1}	0.000305322	7.869827E-05	-0.0003719587	-0.0005141806	0.006339498	0.00321994	0.01894005	0.009674486	0.006339591	0.003219947	0.01894019	0.009674751
2^{2}	-0.000453416	-0.000236440	0.001996322	0.0008086193	0.02582131	0.01364339	0.07776869	0.04064625	0.02582152	0.01364345	0.07777268	0.0406469
2^{3}	-0.0003742723	-0.00205192	-0.003117458	0.003248653	0.1030338	0.05389161	0.3088786	0.1637885	0.1030339	0.05389582	0.3088883	0.1637991

Fig 8. Comparison of the two estimators in Laplace cases (sample size: 5000).

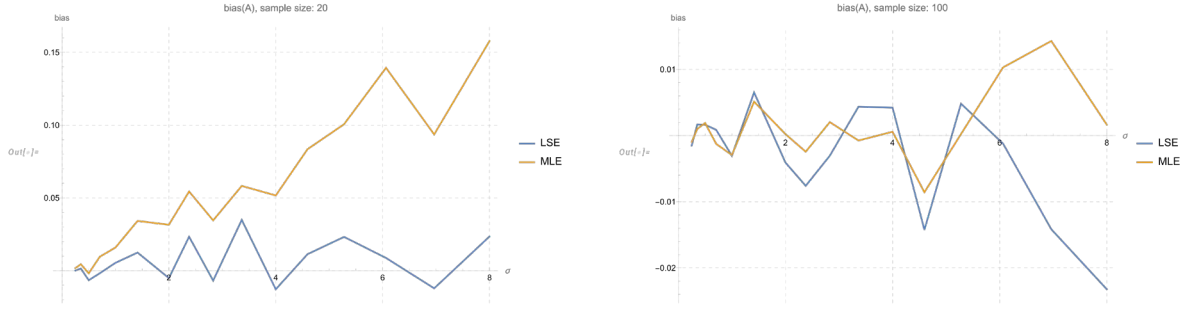

Fig 9. "Bias(A) of estimator – σ" plot in uniform cases.

Fig 10. "Bias(A) of estimator – σ" 3D plot in uniform cases from two different aspects. The second aspect reflects the tendency of the points to gather towards 0 when sample size turns larger.

ISSN 2959-6157

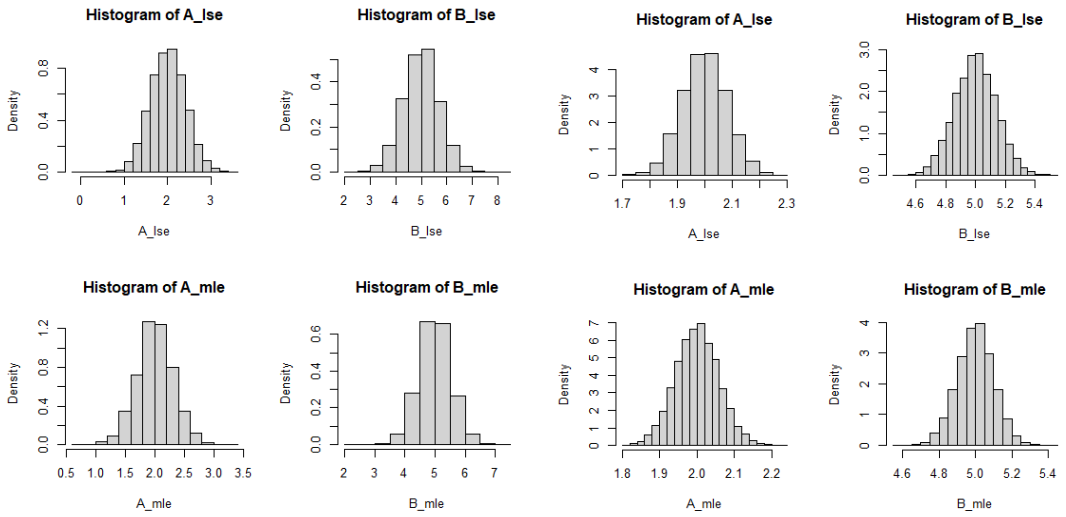

Fig 11. Histogram for the estimated parameters for Laplace cases obtained each time ($\sigma = 2$). Left: sample size=200, Right: sample size= 5000.

Firstly, considering bias, MLE began to outperform LSE when the sample size was between 500 and 1000. However, with smaller sample sizes (e.g., 20), MLE can be considerably more biased than LSE. Nonetheless, as with uniform cases, the final MSE is largely determined by the variance, as $bias^2$ is negligible compared to variance.

For Variance, MLE generally performs better than LSE, especially when sample sizes increase (e.g., 1000), the MLE becomes significantly smaller.

But as the case of uniform distributions, the final MSE is mostly determined by the size of the variance – the bias is negligible compared to the variance. So regardless of whether sample sizes are large or small, the MSE of MLE is less than that of LSE, indicating that MLE, in general, is superior in terms of overall performance.

Furthermore, in Fig. 5 and Fig. 11, the output parameters of both distribution cases follow a normal distribution approximately. In Laplace cases, it is easy to observe that the output obtained by MLE converges much better than that derived by LSE. In uniform cases, this phenomenon seems even more prominent.

4 Conclusion

In this study, we have undertaken an in-depth exploration of the performance of Maximum Likelihood Estimation (MLE) and Linear Least Squares Estimation (LSE) within the framework of linear regression analysis. Our focus has been on scenarios where the error terms do not conform to the traditional assumption of normal distribution, a common challenge in real-world data analysis.

Through rigorous Monte-Carlo simulations, we have demonstrated that while LSE offers computational simplicity and has been historically preferred due to its "unassailable advantage" in saving computational resources, it falls short when the error terms deviate from normality. Similarly, the uniform distribution's applicability in modeling service requests over a fixed interval, as highlighted by Elandt (1961), and the Laplace distribution's utility in financial risk management, as noted by Kotz et al. (2001), underscore the importance of considering the error distribution.

Our results indicate that MLE, despite its computational complexity, provides a more nuanced approach to parameter estimation. It accounts for the specific characteristics of the error distribution, offering a significant advantage when dealing with non-normal distributions such as the Laplace or uniform. This was empirically validated through our simulations, which showed that MLE not only adapts more effectively to various distributions but also converges towards more accurate estimates with increasing sample size.

A critical finding from our analysis is that MLE begins to outperform LSE in terms of bias and mean square error as the sample size grows, particularly in the range of 500 to 1000 observations, when Laplace (or double exponential) distribution is investigated while the same scenario happens in uniform cases when sample size is between 20 and 100. This suggests that for larger datasets, MLE is likely to yield more reliable and precise estimates. The reduction in bias and the lower variance observed with MLE signify its enhanced stability and overall improved performance compared to LSE.

In conclusion, our study reaffirms the importance of selecting an estimation method that aligns with the underlying data distribution. While LSE remains a valuable

tool for smaller datasets or when data closely adhere to normality, MLE presents a compelling alternative for scenarios involving non-normal error distributions.

References

Zeckhauser, R. and Thompson, M. (1970) 'Linear regression with non-normal error terms', *The Review of Economics and Statistics*, 52(3), p. 280. doi:10.2307/1926296.

Mandelbrot, B. (1971) 'Linear regression with non-normal error terms: A comment', *The Review of Economics and Statistics*, 53(2), p. 205. doi:10.2307/1925721.

Hsu, D.A. (1979) 'Long-tailed distributions for position errors in navigation', *Applied Statistics*, 28(1), p. 62. doi:10.2307/2346812.

Kozubowski, T.J. and Podgórski, K. (2001) 'Asymmetric Laplace laws and modeling financial data', *Mathematical and Computer Modelling*, 34(9–11), pp. 1003–1021. doi:10.1016/s0895-7177(01)00114-5.

Jerrum, M.R., Valiant, L.G. and Vazirani, V.V. (1986) 'Random generation of combinatorial structures from a uniform distribution', *Theoretical Computer Science*, 43, pp. 169–188. doi:10.1016/0304-3975(86)90174-x.

Dean, B. and King, R.A.R. (2009) *Versatile regression: Simple regression with a non-normal error distribution*. Applied Statistics Education and Research Collaboration (ASEARC). Available at: (Accessed: 7 September 2024).

Appendix

Appendix 1 Mathematica code

```
In[•]:= dataset1 = {{1/4, -0.0002032391}, {2^(-1.5), 0.0003388565}, {1/2, 0.0009351535}, {2^(-0.5), 0.0005340668}, {1, 0.00128793},
          {2^0.5, 0.0005551289}, {2, -0.002222316}, {2^1.25, 0.0001692939}, {2^1.5, -0.005235728}, {2^1.75, 0.0006050632}, {4, 0.002046905},
          {2^2.2, -0.009532271}, {2^2.4, 0.0001726605}, {2^2.6, -0.007174284}, {2^2.8, 0.002599233}, {8, -0.003114378}};
        dataset2 = {{1/4, 0.0001874241}, {2^(-1.5), 0.0002625227}, {1/2, 0.0002543145}, {2^(-0.5), 0.0007725386}, {1, 0.0002170219},
          {2^0.5, -0.001132008}, {2, -0.001523082}, {2^1.25, -0.001254115}, {2^1.5, -0.001056174}, {2^1.75, 0.0003592985}, {4, 0.001102511},
          {2^2.2, -0.004760071}, {2^2.4, 0.0003246675}, {2^2.6, -0.002080041}, {2^2.8, -0.001762455}, {8, -0.00625726}};
        ListLinePlot[{dataset1, dataset2}, PlotLegends → {"LSE", "MLE"}, PlotLabel → "a_bias, sample size: 1000", AxesLabel → {"σ", "bias"},
         GridLines → Automatic, GridLinesStyle → Directive[Gray, Dashed], ImageSize → Large]
```

Fig 12. A sample Mathematica code graphing the bias changes in double exponential cases (sample size: 1000).

Appendix 2 R code

The following code is for the double exponential distributed error term $\sigma = 8$ and $samplesize = 5000$.

```
a<-2
b<-5
sigma<-2^3
n<-5000
x<-1:n/n
K<-10000
A_lse<-rep(0, K)
B_lse<-rep(0, K)
A_mle<-rep(0, K)
B_mle<-rep(0, K)
for (i in 1:K){
y<-a+b*x+rexp(n, 1/sigma)-rexp(n, 1/sigma)
mod<-lm(y~x)
A_lse[i]<-mod$coef[1]
B_lse[i]<-mod$coef[2]
b0<-mod$coef[2] ### used for constructing a series of values of b
se<-sqrt(diag(vcov(mod)))[2]  ###. standard error of b0 estimate
low<-b0-3*se
up<-b0+3*se
L<-199  ### try L+1=200 values of b
b_L<-seq(low, up, (up-low)/L)
a_L<-rep(0, L+1)
obj<-rep(0, L+1)
for ( j in 1:(L+1)){
a_L[j]<-median(y-b_L[j]*x)
obj[j]<-sum(abs(y-a_L[j]-b_L[j]*x))/n
}
B_mle[i]<-b_L[which.min(obj)]
A_mle[i]<-median((y-B_mle[i]*x))
}
par(mfrow=c(2,2))
hist(A_lse, prob=T, nclass=15)
hist(B_lse, prob=T, nclass=15)
hist(A_mle, prob=T, nclass=15)
hist(B_mle, prob=T, nclass=15)
cbind(mean(A_lse)-a, mean(A_mle)-a, mean(B_lse)-b, mean(B_mle)-b)
cbind((sd(A_lse))^2, (sd(A_mle))^2, (sd(B_lse))^2, (sd(B_mle))^2)
cbind((sd(A_lse))^2+(mean(A_lse)-a)^2,(sd(A_mle))^2+(mean(A_mle)-a)^2, (sd(B_lse))^2+(mean(B_lse)-b)^2, (sd(B_mle))^2+(mean(B_mle)-b)^2)
```