Thailand Statistician
April 2024; 22(2): 348-362
http://statassoc.or.th
Contributed paper

# Intercept-only Model under Non-normality

**Bouchafaa Asma, Djeddour-Djaballah Khedidja* and Benjrada Mohammed Essalih**
Laboratoire MSTD, Faculty of Mathematics USTHB, Algiers, Algeria
*Corresponding author; e-mail: kdjaballah@usthb.dz

Received: 20 October 2021
Revised: 14 April 2022
Accepted: 14 April 2022

## Abstract

In this paper, we consider a linear regression intercept-only model under the hypothesis of non-normality. Generally, the errors are independent and normally distributed. In our case, we assume the errors are independent and follow an exponential law. We prove the consistency and establish the asymptotic distribution of the maximum likelihood estimator for the parameter of the intercept-only model. Numerical simulations confirm the accuracy of this estimator. We notably exhibit the advantages of the maximum likelihood estimator compared to the classical ordinary least square estimator. Finally, we applied the approach to a data of a real-life example, namely the Canadian lynx data.

**Keywords:** Regression, estimation, exponential distribution, intercept-only, convergence.

## 1. Introduction

The main use of regression is to illuminate a supposed linear relationship between predictor variables and an outcome variable Azais (2005). Regression is an old and established statistical method, with a background that is more relevant for its role in traditional explanatory modeling than for prediction. With the advent of big data, regression is widely used to train a model for predicting outcomes, rather than explaining the data. In this case, the main items of interest are the fitted outcome values. If the main point of a model is prediction, we might not care too much about which independent variables are included, as long as the model fits well. But if the purpose of our model is to see which variables are significant, then much attention needs to be paid to this issue. The goodness of fit is closely related to model selection. Usually, the first step is to determine if there is a relationship between the outcome and the predictors. The null hypothesis refers that there is no relationship between any of the predictors and the response. If the null hypothesis is accepted, we retain the intercept-only model. In these case, it should be noted that there is no slope then the intercept is estimated by the mean response, in Gaussian case. For the sequel, the mean model is linked with the gaussian case else it is the intercept-only model. The mean model may seem overly simplistic, but the sample average is a simple but very powerful descriptor. It counts among the most basic ways to describe, analyze and summarize information about a phenomenon. In the absence of explanatory variables, the mean can be a model by itself. At first glance, it does not seem that studying regression without predictors would be very useful. We are not suggesting that using regression without predictors is a major data analysis tool. Often a model with intercept and predictors is compared to an intercept-only model to test whether the predictors add over and above the intercept-only. There are many phenomena that, when graphed, do not have an incline. Points of

the cloud are almost around a horizontal line. If we are trying to model those phenomena, thus we want our model not to contain a slope.

The mean model is often the starting point for constructing forecasting models for time series data, including random walk models. For example, a brief look at the intercept-only model, consider a time series presenting the daily closing price of the Dow Jones Industrial Average over a some period. Suppose we wish to create a regression model for this time series. But we don't know what factors influence the Closing Price. Neither do we want to assume any inflation, trend or seasonality in the data set. In the absence of any assumptions about inflation, trend, seasonality or the presence of explanatory variables, the best we can do is the intercept-only model. In the intercept-only model, all forecasts take the value of the intercept. Some mathematical transformation (e.g., differencing, logging, deflating, etc.) converts the original time series into a sequence of values, that are independent and identically distributed. Then we can use the mean model to obtain forecasts and confidence limits for the transformed series. We can use the mean model to obtain forecasts and confidence limits for the transformed series. Then, we reverse the transformation to get previsions and confidence limits for the original series. We get the mean of the outcome, e.g. expected value of the outcome, when we do not control for anything. Another reason for doing this is that some packages require the user to define a base model.

We do think that it is worthwhile to look at regression models without predictors to see what they can tell us about the nature of the constant. Understanding the regression constant parameter in these simpler models will help us to understand both the constant and the other regression coefficients in later more complex models. In the case where there are no predictors, the equation reduces to,

$$y_i = a + u_i \qquad i = 1, ..., n. \tag{1}$$

The disadvantage of parametric modeling is the requirement that the structural model and error distribution be correctly specified. In some cases, for example, if the observations come from a discrete distribution or the deviations from the mean have a strong dissymmetry, the hypothesis of normality is no longer tenable. Violation of the normality assumption sometimes may be attributed to the skewed nature of the dependent variable. The data distribution may deviate from a Gaussian distribution in multiple ways. Rather than being continuous, data may be discrete, such as integer counts or even binomial character states. Continuous variables may deviate from normality in terms of skewness (showing a long tail on one side), kurtosis (curvature leading to light or heavy tails), and even higher-order moments. For example, measures of actuating asymmetry are distributed half-normally. Many applications have positive response variables. Such variables usually have distributions right-skewed. The boundary at zero limits the left tail of the distribution.

We suppose that residues $u_i$ are i.i.d.having exponential distribution $\mathcal{E}(\theta^{-1}), \theta > 0$ with $E(u_1) = \theta$, $Var(u_1) = \theta^2$. The exponential probability distribution describes the probabilities with which a random variable $u$ takes on values, where $u$ can only be positive. More precisely, the probability density function of the law exponential for value $u$ is given by

$$f(u) = \theta^{-1} \exp\left(-\theta^{-1} u\right). \tag{2}$$

For example, package of R contains a data set on the number of lynx caught per year in Canada between 1821 and 1934; see also (K.S. Lim, 2020). The shape of the histogram has a decreasing tendency, the values of the observations are all positive; it makes us think of an exponential law. We can suppose that the observations $(x_1, ..., x_n)$ with $n = 114$ are the realizations of random variables $X_1, ..., X_n$ independent of exponential law $\mathcal{E}(\theta)$ with $\theta > 0$.

We propose to estimate the constant parameter in the intercept-only model (1), using maximum likelihood and ordinary least square methods. We would study and compare the properties of the obtained estimators.

Departures from normality may have several causes. For example, they may be due to outlying values in the responses. For this, several researchers proposed to perform regression analysis using a

model that assumes a non-Gaussian distribution for the error terms. Bangdiwala (2018) discuss the statistical and geometric interpretation of simple linear regression models. Gaso et al. (2019) compare the accuracy of two methods. Namely a simple regression method between different vegetation indices and a time series method based on optimization. In Huber (1981) and Tiku et al. (1986), it has been mentioned that the underlying distribution is, in most situations, basically non-normal. Tiku et al.(1986) construct a model with a variety of bivariate non-normal distributions by using the conditional method. Qamarul Islam et al. (2001) consider the simple linear regression model and considered several non-normal distributions for the random error, both symmetric and skew. They obtain the modified maximum likelihood estimators of parameters. Qamarul Islam and Tiku (2004) derive the modified maximum likelihood estimators of the parameters in multiple linear regression models and compare them with the least-squares estimators and the Huber (1981) $M$ -estimators. Nguyen (2015) emphasizes problems where fuzzy data appear naturally and need to be used and analyzed within applied statistics. In Djaballah and Tazerouti (2022) the problem of checking the linearity of a regression relationship is addressed.

Many research focused on the broad class of elliptic distributions, particularly on the multivariate $t$-distribution. For example, Zellner (1976); Sutradhar and Ali (1986); Galea et al. (2002); Diaz-Garcia et al. (2003) investigate cases performed within the elliptic distribution family to analyze more complex situations, such as data with missing values in the response variables, and with monotone missing response variables. The same problem was also approached, within a Bayesian framework, by assuming a multivariate skewed and heavy-tailed distribution for the error terms see Ferreira and Steel (2007, 2012).

The paper is structured as follows: the estimates are presented in Section 2 as well as the finite and asymptotic properties of the ML estimator. Section 3 provides simulations and application to real data. Section 4 is devoted to the proofs.

## 2. Main Results

We will consider two estimators for $a$, the maximum likelihood estimator and the ordinary least squares estimator. The least squares technique has traditionally been justified by two assumptive arguments, it provides the maximal likelihood regression coefficients, if the errors are Gaussian and of all unbiased linear estimators, least squares have a minimal variance. Both of these properties have at times been adduced to call least squares the best of regression techniques. Because least squares possess in addition, the attribute of the computational facility, this method long has reigned as the foremost tool in reducing data to mathematically descriptive relationships. The first argument above assumes a normal distribution of the error terms. We argue that this supposition is often unwarranted, and we can show that significant gains in likelihood may be achieved when the regression technique allows for the more general class of error distribution.

When the probability distribution of errors is assumed, it is possible, to obtain consistent and efficient estimates (minimum variance) of the parameters using the maximum likelihood method. This technique could be widely applied to non-Gaussian noise problems.

### 2.1. Maximum likelihood estimator

We consider likelihood based inference for the parameter $a$. For that, let us calculate $\log L(y_1, ..., y_n; a)$ and look for the solution that maximizes this quantity.

**Proposition 1** The log-likelihood of $y$ is given by

$$L(y_1, ..., y_n, a, \theta) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n}(y_i - a)\right) \mathbf{1}_{\{\inf y_i - a \geq 0\}}. \tag{3}$$

Suppose $\theta$ known, then the estimate of $a$ is given by

$$\widehat{a}_{mle} = \inf_{1 \leq i \leq n} y_i. \tag{4}$$

We have

**Corollary 1** If $\theta$ is unknown, it can be estimate by

$$\widehat{\theta} = \overline{y} - \inf_{1\leq i\leq n} y_i. \tag{5}$$

The distribution of $\widehat{a}_{mle}$: it is helpful to know the law of an estimator. It allows us to calculate the characteristics of the estimator and construct a confidence interval for the parameter. In some cases, that distribution can be determined directly from observed random variables law. The distribution function of $\widehat{a}_{mle}$ is:

$$\begin{aligned} F_{\widehat{a}_{mle}}(t) &= P(\inf_{1\leq i\leq n} y_i \leq t) = 1 - P(\inf_{1\leq i\leq n} y_i \geq t) \\ &= 1 - \prod_{i=1}^{n} P(y_i \geq t) \\ &= 1 - [1 - F_Y(t)]^n = 1 - \left(e^{-\frac{(t-a)}{\theta}} 1_{\{(t-a)\geq 0\}})\right)^n \\ &= 1 - e^{-n\frac{(t-a)}{\theta}} 1_{\{t\geq a\}}. \end{aligned}$$

From where

$$f_{\widehat{a}_{mle}}(t) = \frac{n}{\theta} \exp\left(-\frac{n(t-a)}{\theta}\right) \mathbf{1}_{\{t>a\}}. \tag{6}$$

It's the exponential distribution $\mathcal{E}xp(\frac{n}{\theta})$.

An unknown parameter can have more than one estimator. When we use point estimates, we want them to have certain properties. These properties are important in choosing the best estimator for the parameter, that is, the one that comes closest to the true parameter.

The bias of $\widehat{a}_{mle}$ is defined as $E(\widehat{a}_{mle}) - a$. It is the distance between the average of collection estimates and the single parameter being estimated. The bias also is the expected value of the error, since $E(\widehat{a}_{mle}) - a = E(\widehat{a}_{mle} - a)$. The ideal situation is to have an estimate, unbiased, with low variance, and with few outliers.

### 2.2. Expectation and variance of $\widehat{a}_{mle}$

We calculate expectation and variance of $\widehat{a}_{mle}$.

**Mean:** knowing that $\widehat{a}_{mle}$ can be written as $\widehat{a}_{mle} = Z + a$, where $Z$ follows the law $\mathcal{E}xp(\frac{n}{\theta})$, we deduce that

$$E(\widehat{a}_{mle}) = EZ + a = \frac{\theta}{n} + a. \tag{7}$$

It is therefore, obvious that $\widehat{a}_{mle}$ is a biased estimator of $a$. The bias of $\widehat{a}_{mle}$ is $\frac{\theta}{n}$ tends to 0 when $n \to \infty$, we deduce that $\widehat{a}_{mle}$ is asymptotically unbiased.

**Variance:** the variance of $\widehat{a}_{mle}$ is simply the expected value of the squared sampling deviations; that is,

$$Var(\widehat{a}_{mle}) = E(\widehat{a}_{mle} - E(\widehat{a}_{mle}))^2 = \frac{\theta^2}{n^2}. \tag{8}$$

The variance of $\widehat{a}_{mle}$ is $\frac{\theta^2}{n^2}$. The variance $Var(\widehat{a}_{mle})$ tends to 0 when $n$ tends to $\infty$.

Every time a sample is taken, we lose some part of the information about the population. That inevitably results in an error in the estimate. Therefore, if we want a very high level of precision, we must take a sample of a size sufficient to extract sufficient information from the population to estimate with the desired accuracy.

### 2.3. The Mean squared error ($MSE$)

It is used to indicate how far, on average, the collection of estimates is from the single parameter being estimated. If the MSE is relatively low, then the estimators are likely more highly clustered (than highly dispersed) around the $a$.

$$\begin{aligned} MSE(\widehat{a}_{mle}) &= E\left(\widehat{a}_{mle}-a\right)^2 \\ &= \left(biais\left(\widehat{a}_{mle}\right)\right)^2 + Var\left(\widehat{a}_{mle}\right) \\ &= \left(\frac{\theta}{n}\right)^2 + \frac{\theta^2}{n^2} \\ &= \frac{2\theta^2}{n^2} \to 0 \text{ when } n \to \infty. \end{aligned}$$

Note the difference between $MSE$ and variance.

A consistent sequence of estimators is a sequence of estimators that converge in probability to the quantity being estimated as the sample size grows without bound. In other words, increasing the sample size increases the likelihood that the estimator is close to the population parameter. Mathematically, a sequence of estimators $\{t_n; n \geq 0\}$ is a consistent estimator for parameter $a$ if and only if, for all $\varepsilon > 0$, no matter how small we have

$$\lim_{n\to\infty} P(|t_n - a| > \varepsilon) = 0. \tag{9}$$

The consistency defined above may be called weak consistency. The sequence is strongly consistent, if it converges almost surely to the true value.

**Proposition 2** Let

$$y_{(1)} = \inf_{1\leq i\leq n}(y_i).$$

We have $\sqrt{n}(y_{(1)} - a) \to 0$ in probability when $n \to \infty$.

This shows the consistency of the estimator with

$$y_{(1)} = \widehat{a}_{mle} = t_n$$

and $|t_n - a| = y_{(1)} - a$ because $y_{(1)} > a$. A convergent estimator deviates from the parameter with a low probability, if the sample size is large enough.

**Lemma 1** We prove that $\frac{1}{\log n} y_{(1)} \to \theta$ in probability, $n \to \infty$.

Let $\delta > 0$, this amounts to showing that

$$P\left(\left|\frac{1}{\log n} y_{(1)} - \theta\right| > \delta\right) \to 0 \text{ when } \quad n \to \infty. \tag{10}$$

One of the principal uses of the idea of an asymptotic distribution is in providing approximations to the cumulative distribution functions of statistical estimators. An asymptotical distribution estimator is a consistent estimator whose distribution around the real parameter $\widehat{a}_{mle}$ approaches some distribution with standard deviation shrinking in proportion to $n$ as the sample size $n$ grows. In the asymptotic analysis of the estimators, it is the main challenge is to find the asymptotic distribution of the estimator. We obtain the asymptotic distribution of the maximum likelihood estimator.

The asymptotic distribution of $\widehat{a}_{mle}$ is given in the theorem below.

**Theorem 1** *The asymptotical distribution of $\widehat{a}_{mle}$ is given in terms of Gumbel's distribution. We have*

$$\left(y_{(1)} - \theta \log n\right) \to Z \quad \textit{in distribution} \tag{11}$$

*when $n \to \infty$, where the distrution of $Z$ is defined by*

$$F_Z(t) = 1 - \exp\left(-e^{-\frac{t-a}{\theta}}\right).$$

Thus the limiting distribution of the maximum likelihood estimator is linked with Gumbel's distribution. Thus the limiting distribution of the maximum likelihood estimator is linked with Gumbel's distribution. We can use this distribution to understand the asymptotic behavior of the a. $\widehat{a}_{mle}$.

In hydrology, Gumbel's law is used to analyze variables, such as monthly and annual, maximum values of daily precipitation and river flow volumes, and also to describe droughts.

A second estimator is developed for comparison.

### 2.4. Ordinary least squares

In statistics, ordinary least squares (OLS) is a linear least-squares method for estimating the unknown parameters in a linear regression model. OLS method chooses the parameters by minimizing the sum of the squares of the differences between the observed dependent variable in the given data set and those predicted by the linear function of the independent variable. The resulting estimator can be expressed by a simple formula, especially in the case of simple linear regression. There are several methods for constructing an estimator; among them, the least-squares method (OLS) and the maximum likelihood method are the most used.

**Proposition 3** Assume that $\theta$ is known, the OLS estimator $\widehat{a}_{ols}$ of $a$ is $\overline{y} - \theta$, where $\overline{y}$ is the empirical mean of $y_i$.

**Remark 1** Note that the OLS estimator of $a$ is different from the previously determined maximum likelihood estimator.

There is a reason why we should not use OLS, because there is a violation of the usual assumptions. Indeed, OLS assumes that the mismatch between what is expected and observed is $E(u_i) = 0$. Alas in our case $E(u_i) = \theta$.

The error term accounts for the variation in the dependent variable that the independent variables do not explain. For the model to be unbiased, the average value of the error term must equal zero. The OLS estimator is identical to the maximum likelihood estimator (MLE) under the normality assumption for the error terms.

That assumption is not necessaries for the validity of the OLS method. However, if we assume that the normality assumption does not hold, then some properties must have to be added. In that case, we can get an OLS estimator.

The least-squares estimators are point estimates of the linear regression model parameters. However, generally, we also want to know how close those estimates might be to the reals values of parameters.

The expectation and variance of $\widehat{a}_{ols}$ are

- $E\left(\widehat{a}_{ols}\right) = a$ and
- $Var(\widehat{a}_{ols}) = \frac{\theta^2}{n}$.

It is therefore obvious that $\widehat{a}_{ols}$ is an unbiased estimator of $a$. The bias being equal to zero, we deduce the $MSE$:

$$MSE(\widehat{a}_{ols}) = Var\left(\widehat{a}_{ols}\right) = \frac{\theta^2}{n}. \tag{12}$$

The variance $Var(\widehat{a}_{ols})$ tends to 0 when $n$ tends to infinity. We conclude that $\widehat{a}_{ols}$ is an efficient estimator of $a$. Finally, the $MSE$ obtained for $\widehat{a}_{mle}$ tends to 0 more fast than that obtained with the estimator $\widehat{a}_{ols}$, we conclude that the estimator $\widehat{a}_{mle}$ is better than $\widehat{a}_{ols}$.

We have established that the constant in an OLS regression model has something to do with the mean of the response variable. In particular, in intercept-only models, the intercept is almost equal

to the average of the response variable. If the data errors are Gaussian and independent, the OLS estimators will be maximum likelihood estimators and will be unbiased and of minimal variance. However, if the noise is not Gaussian, the OLS adjustment will give parameter estimates that may be biased. Even when normality does not hold, the Gauss-Markov theorem states that the best linear unbiased estimator of regression coefficients is still yielded by OLS estimation, so long as the errors have expectation zero, are uncorrelated, and have equal variance.

## 3. Simulation Study

This section is established with the intention of examining the performance quality of the estimators under study over a finite sample size $n$.

The simulation is conducted for a certain value of the parameter to be estimated, namely $a$.

### 3.1. Design of simulation

We generate data that fits our model as follows:

- Generate $n$ iid random variables $\{\varepsilon_i\}_{i=1}^{n}$ from $\mathcal{E}xp(\theta^{-1})$
- $y_i = a + \theta + \varepsilon_i$ for all $i = 1, ..., n$,

where the intercept $a$ is chosen arbitrarily.

**Remark 2** The intercept $a$ and the scale parameter $\theta$ have been appropriately chosen in order to have a good model. Actually, the intercept would be better to be moderately and the exponential parameter small. This helps to have a better comparison.

The Mean Square Error ($MSE$) is chosen to be a criterion for quantifying the performance of our estimators. The $MSE$ of an estimator $\hat{\theta}$ with respect to an unknown parameter $\theta$ is defined as

$$MSE\left(\hat{\theta}\right) = E_\theta\left(\hat{\theta} - \theta\right)^2.$$

### 3.2. Consistency results

To give an overview of the influence of the sample size $n$ on the quality of fit, the least square (OLS) and maximum likelihood estimators (MLE) $\hat{a}_{OLS}$ and $\hat{a}_{MLE}$ respectively was implemented from Model 1 which contaminated by exponential errors with $\theta = 0.25$. See Figure 1.

To exhibit more comparison of the influence of the sample size $n$ on the estimation fit, the values of $MSE$ and $Bais$ are computed from Model 1 and summarized in Table 1. The first column displayed for the different values of $n$, the second column displayed for the results(MSE and Bias) of $\hat{a}_{ols}$. While the last column is for $\hat{a}_{mle}$.

**Table 1** Simulation results: the values of the MSE (for both $OLS$ and $MLE$ estimators) with the corresponding Bias

| Estimators | OLS | | MLE | |
|---|---|---|---|---|
| sample size $n$ | $MSE$ | $Bias$ | $MSE$ | $Bias$ |
| 50 | $4.1301\times10^{-2}$ | 0.0132 | $3.3325\times10^{-4}$ | 0.1454 |
| 200 | $2.4581\times10^{-3}$ | 0.0059 | $6.7712\times10^{-5}$ | 0.0131 |
| 500 | $2.2041\times10^{-4}$ | 0.0011 | $2.3865\times10^{-6}$ | 0.0027 |
| 1000 | $2.4891\times10^{-4}$ | 0.0042 | $3.6711\times10^{-7}$ | -0.0085 |

**Remark 3** From the simulation results in Table 1 and Figure 1, we can see that the quality of fit depends directly on the estimation method and the sample size $n$. Actually, the larger the sample size, the better the quality of performance will be. Furthermore, the quality of fit declines substantially for the Ordinary Least Squares method compared to the Maximum likelihood method but it increases with a sufficiently large sample size.

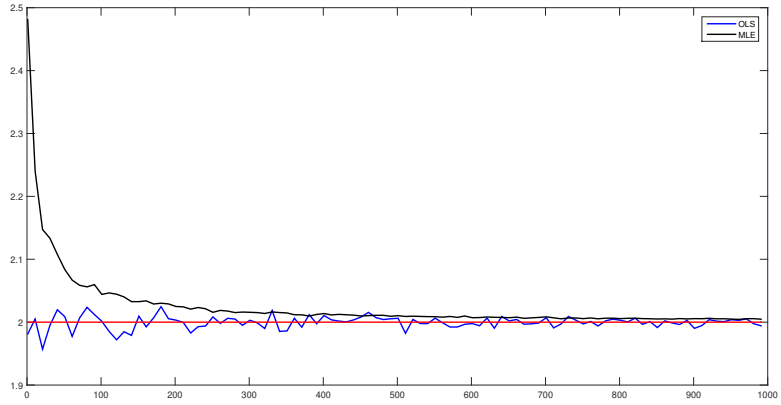
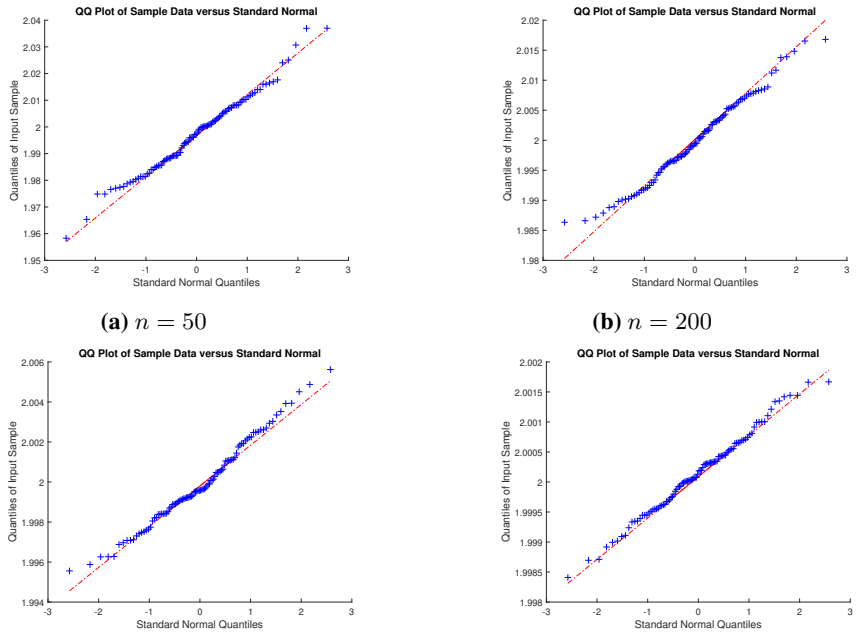



**Figure 1** Simulation results of estimation from Model 1 for $\theta = 0.25$ and $a = 2$. The red line corresponds to the true intercept, the blue line corresponds to estimation by the Last Square Error method, the black line corresponds to the estimation by the Maximum Likelihood Method

### 3.3. Asymptotic normality

In this subsection, we examine the asymptotic normality of the understudy estimators throughout normal-probability plots. For this aim, we only consider the estimation given by the Ordinary Least Squares method. And for better comparison this estimator was implemented here for $\theta = 0.25$, $m = 100$ iterations, and $n = 50$, 200, 500 and 1000. The results of this numerical implementation are summarized in Figure 2.




**(a)** $n = 50$ **(b)** $n = 200$

**(c)** $n = 500$ **(d)** $n = 1000$

**Figure 2** The normal-probability plots of the ordinary least squares estimator for $n = 50$ $n = 200$, 500 and 1000, $\theta = 0.25\%$

**Remark 4** From graphs summarised in Figure 2, we can see that for the asymptotic normality, the estimator provides good performance for a large sample size. That indicates that the convergence in distribution becomes better more and more along with $n$.

## 4. Implementation to The Number of Lynx in Canada

We re-examine the annual trappings of the Canadian lynx over the years 1821-1934, which have been reported and analyzed extensively. The "data set" R package contains those data. Figure 1 presents the corresponding histogram; from which we think that we have an exponential distribution. In fact, it has a decreasing tendency, furthermore, the observations are all positive.

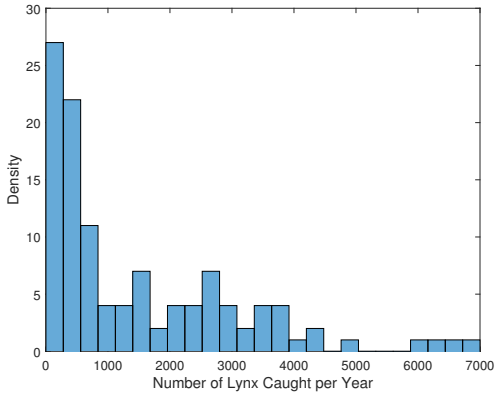



**Figure 3** Histogram for data of the number of lynx caught per year in Canada from 1821 to 1934

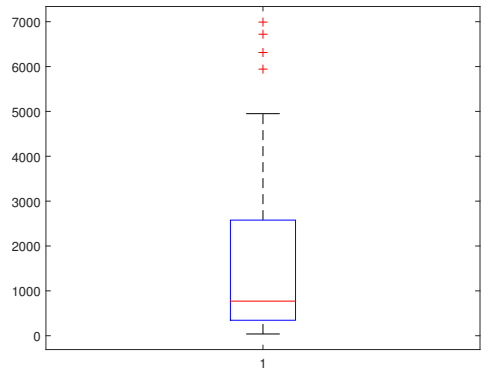



**Figure 4** Boxplot of the number of lynx caught per year in Canada from 1821 to 1934

From the graph in Figure 4, we can see that the data might contain an invariable intercept. Hence, our goal here is to compare the adjustment of our data under the two possible models, namely:

1. **The presence of the intercept**: In this situation, we suppose that the data are coming from the following model
$$lynx_i = a + \varepsilon_i \quad \text{for all } i = 1, ..., 114 \tag{13}$$
2. **Ignoring the intercept**: We assume that the data are from
$$lynx_i = \varepsilon_i \quad \text{for all } i = 1, ..., 114 \tag{14}$$

where $lynx_i$ stands for a number of lynx caught in the $i$-th year between 1821 and 1934. The random variables $\{\varepsilon_i\}_{i=1}^{114}$ are supposed to be iid and follow an exponential distribution of shape parameter $\mu$ to be estimated.

At this level, we are interested to estimate the exponential parameter $\theta$ and use this latter to conclude the intercept estimator. In the situation where the intercept is considered, and from the result in Corollary 1, we have

$$\hat{\theta} := \left(\overline{lynx} - \inf\left(lynx_i\right)\right) = 0.00066710. \tag{15}$$

While for the second model, we have

$$\hat{\theta} := \overline{lynx} = 0.00065018. \tag{16}$$

Using the result in previous sections 2.1 and 2.2, we get

$$\hat{a}_{ols} = 37.9919 \tag{17}$$

and

$$\hat{a}_{mle} = 39. \tag{18}$$

To provide a more clear comparison, we consider the following data

$$lynx'_i = lynx_i - \hat{a}_{mle} \quad \text{for all } i = 1, ..., 114. \tag{19}$$

So, the aim now is to compare the adjustment of our new data and see which model explains more clearly the number of lynx caught per year in Canada.

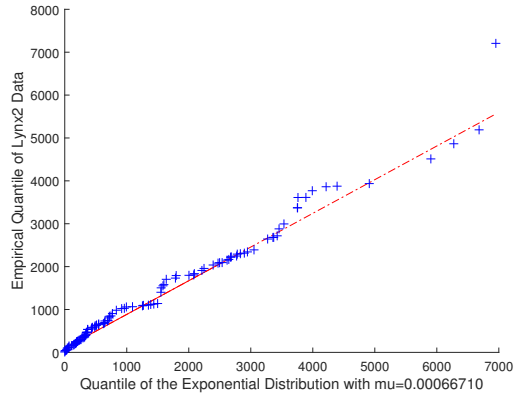



**Figure 5** The quantile-quantile plot of the quantiles of the sample data lynx2 versus the theoretical quantile values from an exponential distribution with $\theta = 0.00066710$

**Remark 5** Figures 5-6 reveal that the data $lynx'_i$ coincides better with an exponential model with an intercept compared to a free-exponential model (ie without intercept). Hence, we can conclude that the data set on the number of lynx caught per year in Canada on the period between 1821 and 1934 refers to an exponential model with intercept-only.

## 5. Conclusion

In regression, the intercept-only model has no independent variables. Thus, in the Gaussian case, it predicts that the best estimate of the dependent variable is the overall mean. In this study, we have considered that the distribution is not Gaussian but exponential. We have determined two

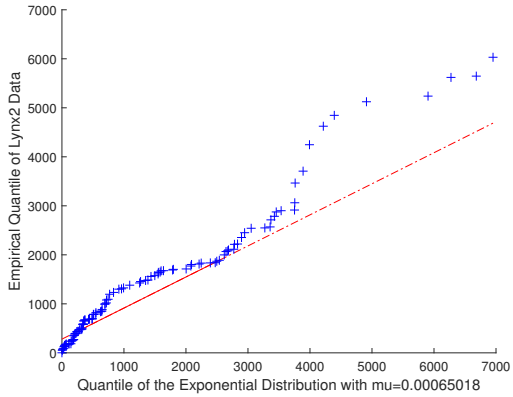



**Figure 6** The quantile-quantile plot of the quantiles of the sample data lynx2 versus the theoretical quantile values from an exponential distribution with $\theta = 0.00065018$

estimators, namely the maximum likelihood estimator (MLE) and the ordinary least squares estimator (OLS). Both techniques can be employed to construct consistent estimators. We have proved the consistency of the two estimators and the asymptotic normality of the maximum likelihood estimator. The second estimator was developed essentially for comparison. The estimated parameter differs for the two methods. The advantage of the likelihood method proposed is significant, regarding its rate of convergence. Estimation using ML in conjunction with an intercept-only model shows higher accuracy than using an OLS estimate based on the same model. However, the implementation of OLS method allows ease of calculation of the estimators. Even though both methods have limited utility to estimate the parameter with high accuracy, they could be suitable for predictions. We have illustrated the performance of the two estimators via numerical studies. The approach is implemented to estimate the model of the number of lynx caught per year in Canada from a real data set.

## References

Azais JM, Bardet JM. Le modle linaire par l'exemple: rgression, analyse de la variance et plans d'expriences illustres avec R, SAS et Splus. Dunod; 2005.

Bangdiwala SI. Regression: simple linear. Int J Inj Control Sa. 2018; 25(1):113-115.

Djaballah-Djeddour K, Tazerouti M. Test for Linearity in Non-Parametric Regression Models. Aust J Stat. 2015; 51(1):16-34.

Diaz-Garcia JA, Galea Rojas M, Leiva-Sanchez V. Influence diagnostics for elliptical multivariate linear regression models. Commun Stat - Theor M. 2003; 32(3):625-641.

Ferreira JT, Steel FJ. A new class of skewed multivariate distributions with applications to regression analysis. Stat Sinica. 2007; 505-529.

Galea M, Paula GA, Bolfarine H. Local influence in elliptical linear regression models. The Statistician. 1997; 46(1):71-79.

Gaso DV, Berger AG, Ciganda VS. Predicting wheat grain yield and spatial variability at field scale using a simple regression or a crop model in conjunction with Landsat images. Comput Electron Agr. 2019; 159(1): 75-83

Huber PJ. Robust Statistics. Wiley, New York. 1981.

Qamarul Islam M, Tiku ML. Multiple linear regression model under nonnormality. Commun Stat - Theory M. 2004; 33(10):2443-2467.

Sazak HS. Regression analysis with a stochastic design variable. Int Stat Rev. 2006; 74(1): 77-88.

Sutradhar BC, Mir MA. Estimation of the parameters of a regression model with a multivariate t error variable. Commun Stat - Theor M. 1986; 15(2):429-450.

Tiku ML, Tan WY, Balakrishnan N. Robust Inference. New York Marcel Dekker. 1986.
Tiku ML, Qamarul Islam M, Sazak HS. Estimation in bivariate nonnormal distributions with stochastic variance functions. Comput Stat Data An. 2008; 52(3): 1728-1745.
Tiku ML, Islam MQ, Selcuk AS. Non-normal regression.II.Symmetric distributions. Commun Stat - Theor M. 2001; 30(1):1021-1045.
Zellner A. Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. J Am Stat Assoc. 1976; 71(354):400-405.
Lim KS. Canadian lynx data. Encyclopedia of Mathematics. Available from: http://encyclopediaofmath.org/index.php?title=Canadian_lynx_data&oldid=46191.

## Appendix

**Proof of Proposition 1**

Given the assumed structural model (1) and the known error distribution, the conditional distribution of $y$ can be derived. We have $u_i \sim \mathcal{E}xp(\theta^{-1})$ which gives by definition $f_u(z) = \frac{1}{\theta}\exp\left(-\frac{z}{\theta}\right)\mathbf{1}_{\{z\geq 0\}}$ and $F_u(z) = 1-\exp\left(-\frac{z}{\theta}\right)$. The distribution function of $y$ is thus deduced as follows:

$$\begin{aligned} F_y(t) &= P(y\leq t) = P(a+u\leq t) \\ &= 1-\exp\left(-\frac{t-a}{\theta}\right) \end{aligned}$$

with $t-a\geq 0$. Therefore

$$f_y(t) = \frac{1}{\theta}\exp\left(-\frac{t-a}{\theta}\right)\mathbf{1}_{\{t\geq a\}}. \tag{20}$$

The log-likelihood is then written

$$\begin{aligned} L(y_1,...,y_n,a,\theta) &= \prod f_y(y_i) \\ &= \frac{1}{\theta^n}\exp\left(-\frac{1}{\theta}\sum_{i=1}^{n}(y_i-a)\right)\mathbf{1}_{\{\inf y_i\geq a\}}. \end{aligned}$$

Let us note

$$g(a) = \tfrac{1}{\theta^n}\exp\left(-\tfrac{1}{\theta}\textstyle\sum_{i=1}^{n}(y_i-a)\right)$$

for every fixed $\theta$ and with $a\in\,]-\infty, \inf_{1\leq i\leq n} y_i]$. The derivative of $g$ is

$$g'(a) = \tfrac{1}{\theta^{n+1}}\exp\left(-\tfrac{1}{\theta}\textstyle\sum_{i=1}^{n}(y_i-a)\right).$$

The derivative $g'(a)$ is always positive $\forall\theta>0$. This implies that $g$ is increasing from $-\infty$ to $\inf y_i$. Consequently, the maximum likelihood estimator of $a$ is reached in:

$$\widehat{a}_{mle} = \inf_{1\leq i\leq n} y_i. \tag{21}$$

**Proof of Corollary 1**

To search for the global max, it suffices to maximize the function $L(y_1,...,y_n,a,\theta)$ or $\log L(y_1,...,y_n,a,\theta)$ with respect to $\theta$. We place ourselves outside the case (which is of zero probability whatever the value of the parameter) where $\sum_{i=1}^{n}(x_i-\inf_{1\leq i\leq n}(x_i))=0$ (which means that all $x_i$ are equal). Suppose that $a$ is fixed

$$\log L(y_1,...,y_n,a,\theta) = -n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n}(y_i-a). \tag{22}$$

Condition necessary: $\frac{d\log L}{d\theta} = 0$. We have

$$\frac{d\log L}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}(y_i - a) \implies n\widehat{\theta} = \sum_{i=1}^{n} y_i - na \implies \widehat{\theta} = \overline{y} - a$$

$a$ is unknown, we can replace it by its estimator:

$$\widehat{\theta} = \overline{y} - \inf_{1\le i\le n} y_i \tag{23}$$

Then we check that the second derivative at this point is negative, which ensures that the critical point is indeed a maximum. By calculating the second derivative, we get

$$\frac{d^2\log L}{d\theta^2} = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n}(y_i - a) \tag{24}$$

$\Rightarrow$

$$\begin{aligned}\frac{n}{\widehat{\theta}^2} - \frac{2}{\widehat{\theta}^3}\sum_{i=1}^{n}(y_i - a) &= -\frac{1}{\widehat{\theta}}\left[-\frac{n}{\widehat{\theta}} + \frac{1}{\widehat{\theta}^2}\sum_{i=1}^{n}(y_i - a)\right] - \frac{1}{\widehat{\theta}^3}\sum_{i=1}^{n}(y_i - a)\\ &= -\frac{1}{\widehat{\theta}^3}\sum_{i=1}^{n}(y_i - a).\end{aligned}$$

Since $\theta > 0$ and $\sum_{i=1}^{n} y_i > na$ because all $y_i > a \Rightarrow \frac{d^2\log L}{d\theta^2} < 0$.

**Proof of Proposition 2**

We write
$P(\sqrt{n}\left|y_{(1)} - a\right| > \delta) = P\left(\left|\inf_{1\le i\le n} y_i - a\right| > \frac{\delta}{\sqrt{n}}\right)$ and $\inf_{1\le i\le n} y_i \ge a$.
Then we get

$$\begin{aligned}P(\sqrt{n}\left|y_{(1)} - a\right| &> \delta) = P\left(\inf_{1\le i\le n} y_i > \frac{\delta}{\sqrt{n}} + a\right)\\ &= 1 - P\left(\inf_{1\le i\le n} y_i \le \frac{\delta}{\sqrt{n}} + a\right)\\ &= 1 - F_{\widehat{a}_{mv}}\left(\frac{\delta}{\sqrt{n}} + a\right)\\ &= e^{-\frac{\sqrt{n}}{\theta}\delta}\end{aligned}$$

then

$$\lim_{n\to\infty} P(\sqrt{n}\left|y_{(1)} - a\right| > \varepsilon) = 0. \tag{25}$$

**Proof of Lemma 1**

We prove (10) by splitting the event into two events

- $P\left(\frac{1}{\log n}y_{(1)} \le \theta - \delta\right) \to 0$ when $n \to \infty$
- $P\left(\frac{1}{\log n}y_{(1)} \ge \theta + \delta\right) \to 0$ when $n \to \infty$

The first is rewritten as:

$$\begin{aligned} P\left(\frac{1}{\log n}y_{(1)} \leq \theta-\delta\right) &= 1-P\left(\frac{1}{\log n}y_{(1)} \geq \theta-\delta\right) \\ &= 1-\left(P\left(y_1 \geq (\theta-\delta)\log n\right)\right)^n \\ &= 1-\left(1-P\left(y_1 \leq (\theta-\delta)\log n\right)\right)^n \\ &= 1-\left(e^{-\frac{((\theta-\delta)\log n-a)}{\theta}}\right)^n. \end{aligned}$$

We assume that $\delta < \theta$.
The second is rewritten as:

$$\begin{aligned} P\left(\frac{1}{\log n}y_{(1)} \geq \delta+\theta\right) &= P\left(y_{(1)} > (\delta+\theta)\log n\right) \\ &\leq nP\left(y_1 > (\delta+\theta)\log n\right). \end{aligned}$$

This increase by sub-additivity of $P$ is sufficient, and

$$\begin{aligned} nP\left(y_1 > (\delta+\theta)\log n\right) &= n-\left(1-e^{-\frac{(\delta+\theta)\log n-a}{\theta}}\right)n \\ &= ne^{-\frac{(\delta+\theta)\log n-a}{\theta}} = ne^{\frac{a}{\theta}}e^{-\frac{(\delta+\theta)\log n}{\theta}} \\ &= e^{\frac{a}{\theta}}n^{-\frac{(\delta+\theta)}{\theta}+1} = e^{\frac{a}{\theta}}n^{-\frac{\delta}{\theta}}. \end{aligned}$$

**Proof of Theorem 1**

We can already predict that for $n$ sufficiently large the expectation of $\widehat{a}_{ml}$ will be close to $a$. This needs to be clarified. We know that

$$\lim_{n\to\infty}(1-\tfrac{e^{-a}}{n})^n = e^{-e^{-a}}.$$

The sequence of general term $y_{(1)} - \theta\log n$ converges in distribution towards a limit that one seeks to determine.

$$\begin{aligned} P(y_{(1)}-\theta\log n \leq t) &= P(y_{(1)} \leq t+\theta\log n) \\ &= 1-\left(P(y_{(1)} \geq t+\theta\log n)\right)^n \\ &= 1-\left(1-e^{-\frac{(t+\theta\log n-a)}{\theta}}\right)^n \\ &= 1-\left(1-\frac{e^{-\frac{t-a}{\theta}}}{n}\right)^n. \end{aligned}$$

We know that

$$\lim_{n\to\infty}\left(1-\frac{e^{-\frac{t-a}{\theta}}}{n}\right)^n = \exp\left(-e^{-\frac{t-a}{\theta}}\right) = G(\tfrac{t-a}{\theta}),$$

where $G$ is the distribution function of Gumbel's law.
The cumulative distribution function of the Gumbel distribution is :

$$F_{Gumbel}(x;\mu,\beta) = \exp\left(-e\left(\frac{\mu-x}{\beta}\right)\right). \tag{26}$$

The standard Gumbel distribution is the case where $\mu = 0$ et $\beta = 1$. The Gumbel distribution is used to model the distribution of the maximum (or the minimum) of a number of samples of various distributions. We conclude that, when $n \to \infty$

$$y_{(1)} - \theta\log n \to Z \quad \text{in distribution}, \tag{27}$$

where the distribution of $Z$ is defined by

$$P(Z \geq t) = 1 - F_Z(t) = \exp\left(-e^{-\frac{t-a}{\theta}}\right) = F_{Gumbel}(a; t, \theta).$$

Gumbel has shown that the maximum value in a sample of a random variable following an exponential distribution minus the logarithm of the sample size approaches the Gumbel distribution closer with increasing sample size.

**Proof of Proposition 3**

The error term accounts for the variation in the dependent variable that the independent variables do not explain. For the model to be unbiased, the average value of the error term must equal zero.

Let $y_i = a + u_i = a + \theta + \varepsilon_i$ which implies $E\varepsilon_i = 0$ and $E\left(Y_i\right) = a + \theta$. Note $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The OLS method consists of minimizing:

$$S(a) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a - \theta)^2. \tag{28}$$

The solution to the problem of minimization (28), denoted $\widehat{\alpha}_{ols}$, is given by

$$\widehat{a}_{ols} = \frac{1}{n}\sum_{i=1}^{n} y_i - \theta. \tag{29}$$

Let's calculate the expectation and variance of $\widehat{a}_{ols}$

- $E\left(\widehat{a}_{ols}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} y_i - \theta\right) = \frac{1}{n}\sum_{i=1}^{n} Ey_i - \theta = a$
- $Var(\widehat{a}_{ols}) = Var(\overline{y}) = \frac{\theta^2}{n}$.