

Ordinary Least Squares and Maximum Likelihood Regression with Cauchy Errors

By

SAMAH JRADI

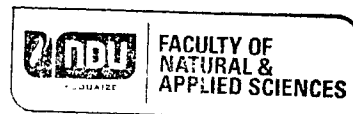
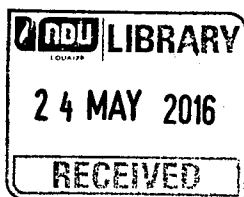
Thesis Advisor: Dr. John Haddad

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Mathematics
in the Department of Mathematics and Statistics
in the Faculty of Natural and Applied Sciences
of Notre Dame University-Louaize

Lebanon

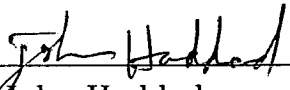
August 25th, 2015



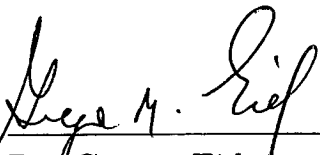
Ordinary Least Squares and Maximum Likelihood Regression with Cauchy Errors

SAMAH JRADI

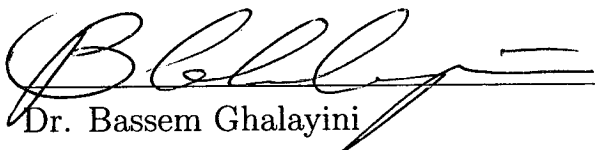
Approved by:



Dr. John Haddad



Dr. George Eid



Dr. Bassem Ghalayini

Acknowledgments

The success and final outcome of this project required a lot of guidance and assistance from many people and I was extremely fortunate to work and learn from those people.

I owe first my profound gratitude to my advisor, Dr. John Haddad, for his excellent guidance, caring and motivation. His guidance has helped me in all the time of research and in writing this thesis. Besides, my advisor I would like to express my special appreciation and thanks to Doctor George Eid and Dr. Bassem Ghalayini for their continuous support and kindness.

Finally I would like to appreciate my family, my elder brother, my younger sisters and especially my mom and dad for standing beside me during this period of hard work because I would have never reached my goals without their sincere care and continuous support.

Abstract

In this paper, we'll examine the effect of Cauchy errors in a linear model on the performance of the least squares and maximum likelihood estimators with the aid of two factors; the sample size and the Cauchy scale parameter. A sampling distribution of one hundred experiments was done to judge the estimation process in the case of least squares method. On the other hand, multivariate Newton Raphson method was used to calculate the unique solution of the partial derivatives of the log likelihood function. The uniqueness of the solution is proved for a known location parameter and unknown scale parameter. At the end, a comparison is held based on the experimental methods done.

The methodologies used were implemented on R language.

Contents

0.1	Continuous Random Variables	7
0.2	Cumulative Distribution Function	7
0.3	Probability Density Function	7
0.4	Expectation and Variance	7
0.4.1	Expectation	7
0.4.2	Variance	8
0.4.3	Moment Generating Function	8
0.5	Linear Regression	8
1	Cauchy Distribution	9
1.1	The Standard Cauchy Distribution	10
1.1.1	Cumulative distribution function	10
1.1.2	Expected Value	10
1.1.3	Characteristic Function	11
1.1.4	Ratio of independent normal variables	12
1.2	General Cauchy Distribution	12
1.2.1	Probability density function	12
1.2.2	Cumulative distribution function	13
1.2.3	Simulating Cauchy Random Variables	13
2	Least Squares Method	14
2.1	Statistics Review	14
2.2	The Method of Least Squares	14
2.3	The Gauss-Markov Assumptions	16
2.4	Gauss-Markov Theorem	16
2.4.1	Unbiased	16
2.4.2	Linear	17
2.4.3	Minimum Variance	17
2.5	Normality of the Stochastic Error	18
3	Cauchy Stochastic Errors	20
3.1	Law of Large numbers and Central Limit Theorem	21
3.2	Sampling distribution of the mean	23

3.3	Outliers	23
3.4	How to deal with outliers?	24
3.5	Experimental Method	24
4	Maximum Likelihood Estimation	27
4.1	MLE of Cauchy distribution with zero mean and unknown scale parameter	28
4.2	Location parameter Known	28
4.3	Newton Raphson Method	29
4.4	Multi-Dimensional Case For Newton Raphson Method	30
4.5	Experimental Method	31
4.6	Optimum Properties of Maximum Likelihood Estimation	33
4.7	Disadvantages of Newton Raphson Method	34

List of Figures

1.1	Cauchy distribution description	9
3.1	Variation of sample mean as a function of sample size (Cauchy case)	22
3.2	Variation of sample mean as a function of sample size (normal case)	22
3.3	Visual detection of outliers	23

List of Tables

3.1	Summary Table	23
3.2	A small sample size result table, n=20	25
3.3	A large sample size result table, n=100	26
4.1	MLE estimators, n=20	32
4.2	MLE estimators, n=100	33

Introduction

The assumption of normality in stochastic modeling has been adapted to almost statistical inference. This assumption had not been questioned until the time of Pareto. As a result, it is very important to test the effect of infinite variance distributions on the performance of the conventional statistical methods' estimators specifically the least squares method and the maximum likelihood estimation. This study will mainly focus one of the most extreme members of the infinite variance family which is the Cauchy distribution. The fact that Cauchy distribution can result as a ratio of normal variables or even the ratio of non-normal variables makes our study very reasonable and efficient. Moreover, it is not unreasonable to suggest that the Cauchy can't arise in some applied economic research because the modeling of investment expenditures Resek's calls for the ratio of investment in constant dollars to capital stock which may lead to a Cauchy distribution.

Blattberg and Sargent have shown that the OLS (ordinary least squares) estimator doesn't perform well in the estimation of linear models with non gaussian errors. Consequently, the relevant question becomes whether the performance of OLS with models following Cauchy errors can be applicable for small samples or not. Moreover, this arouses the question about the factors that affect the performance of OLS estimators. Since the theoretical results don't give any evidence for the threshold sample size for the use of OLS, a series of sampling experiments have been performed to test the effect of the sample size and the Cauchy spread parameter on the OLS estimators' performance.

The maximum likelihood estimation method is also tested regarding the sample size and the Cauchy spread parameter. Multivariate Newton Raphson method is used to find the solution of the log of the likelihood function due to the lack of closed form solutions for a sample size greater than 4. The Newton Raphson method is implemented using R language.

Preliminaries

0.1 Continuous Random Variables

Definition 0.1. *Continuous Random Variables*

A random variable is a mapping $X: S \rightarrow \mathbb{R}$ from the sample space S to the real numbers inducing a probability measure $P_X(B) = P(X^{-1}(B))$, $B \in \mathbb{R}$.

A random variable is said to be continuous if there exists $f_X: \mathbb{R} \rightarrow \mathbb{R}$ such that $P_X(B) = \int_{x \in B} f_X(x) dx$ where f_X is the probability density function (pdf) of X .

0.2 Cumulative Distribution Function

Definition 0.2. *Cumulative Distribution Function*

The cumulative distribution function (cdf) is given by $F_X(x) = \int_{-\infty}^x f_X(t) dt \forall x \in \mathbb{R}$.

0.3 Probability Density Function

Definition 0.3. *Probability density function*

The probability density function is the derivative of the cumulative distribution function given by $f_X(x) = \frac{d}{dx} F_X(x)$.

0.4 Expectation and Variance

0.4.1 Expectation

Let X be a continuous random variable. The expectation or the mean is μ_X or $E_X(x)$ defined by $E_X(X) = \int_{-\infty}^{\infty} x f_X(x) dx$.

Linearity of Expectation

Clearly, we can see that $E_X(aX + b) = aE_X(X) + b$

0.4.2 Variance

Definition 0.4. The variance of a continuous random variable is given by σ^2 or $\text{Var}_X(X) = E_X(X - \mu_X)^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$

It is easy to show that $\text{Var}_X(X) = E_X(X^2) - E_X(X)^2$. Moreover, for a linear transformation $\text{Var}_X(aX + b) = a^2 \text{Var}_X(X)$.

0.4.3 Moment Generating Function

Definition 0.5. Moment Generating Function

The moment generating function is defined as $M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$.

Theorem 1 (Central Limit Theorem). Central limit theorem is the second fundamental theorem of probability. It states that if S_n is the sum of n identically independent random variables X_i having finite mean (μ_X) and variance (σ^2) then $\lim_{n \rightarrow \infty} P\left(\frac{S_n - n\mu_X}{\sigma\sqrt{n}} \leq x\right) = \Phi(x)$

where Φ is the normal cumulative distribution function.

0.5 Linear Regression

Definition 0.6. Regression Analysis

Regression analysis is the art and science of fitting straight lines to patterns of data. The independent variable Y in a linear regression model is predicted from k other independent variables X_1, X_2, \dots, X_k . The value of Y at time t is determined by the following linear equation

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \varepsilon_t$$

where the betas are constants and the epsilons are independent identically distributed normal random variables with mean 0 and variance equal to 1. Epsilons are called the noise. β_0 is called the intercept of the model and $\beta_{i's}$ are called the multiplier or coefficient of the variables $X_{i's}$.

The corresponding equation for predicting Y_t from the corresponding values of the $X_{i's}$ is

$$\hat{Y}_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + \dots + b_k X_{kt}$$

where the $b_{i's}$ are the estimates of the $\beta_{i's}$ obtained by least squares method.

Chapter 1

Cauchy Distribution

The Cauchy distribution is, also called the Lorentzian distribution, is a continuous function which describes the distribution of horizontal distances at which a line segment titled at a random angle cuts the x-axis. let θ be the angle that a line with a fixed point of rotation makes with the vertical axis as shown in the following figure.

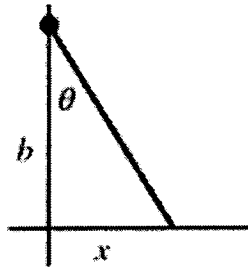


Figure 1.1: Cauchy distribution description

$$\begin{aligned}\tan \theta &= \frac{x}{b} \\ \theta &= \arctan \frac{x}{b} \\ d\theta &= \frac{1}{1 + \frac{x^2}{b^2}} \frac{dx}{b}\end{aligned}$$

so the distribution of θ is given by:

$$\frac{1}{\pi} d\theta = \frac{1}{\pi} \frac{1}{1 + \frac{x^2}{b^2}} \frac{dx}{b}$$

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{\pi} d\theta = 1$$

$$\int_{-\infty}^{\infty} \frac{b}{\pi} \frac{1}{1 + \frac{x^2}{b^2}} dx = \left[\frac{1}{\pi} \arctan \frac{x}{b} \right] = \frac{1}{\pi} \left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right) \right] = 1$$

1.1 The Standard Cauchy Distribution

Let X be a random variable, then X has the standard Cauchy distribution if it has the following probability density function:

$$f(x) = \frac{1}{\pi(1+x^2)}; \quad x \in \mathbb{R}$$

1.1.1 Cumulative distribution function

X has a cumulative distribution function given by:

$$F(x) = \int_{-\infty}^x f(t) dt = \frac{1}{\pi} \arctan t \Big|_{-\infty}^x = \frac{1}{\pi} \arctan x + \frac{1}{2}$$

1.1.2 Expected Value

The expected value of a Cauchy distribution doesn't exist.

Proof. By definition

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

For the latter improper integral not to exist at least one of the integrals $\int_a^{\infty} x f(x) dx$ or $\int_{-\infty}^a x f(x) dx$ doesn't exist.

$$E(X) = \int_a^{\infty} x f(x) dx = \int_a^{\infty} \frac{x}{\pi(1+x^2)} dx = \frac{1}{2\pi} \ln(1+x^2) \Big|_a^{\infty} = \infty$$

Consequently, the expected value doesn't exist. □

1.1.3 Characteristic Function

The expected value of the function $\exp(itx)$ is called the characteristic function for the probability distribution $f(x)$, where t is parameter that can have any real value and i is the square root of -1 . That is to say, the characteristic function of $f(x)$ is

$$\Theta(t) = E(\exp(itx)) = \int_{-\infty}^{\infty} \exp(itx)f(x)dx$$

X has a characteristic function Θ given by

$$\Theta(t) = E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} \frac{1}{\pi(1+x^2)} dx$$

This integral will be computed using Cauchy's integral formula. Suppose first that $t \geq 0$. For $r > 1$, let Γ_r denote the curve in the complex plane consisting of the line segment L_r on the x -axis from $-r$ to r and the upper half circle C_r of radius r centered at the origin. We give Γ_r the usual counter-clockwise orientation. On the one hand we have:

$$\int_{\Gamma_r} \frac{e^{itz}}{\pi(1+z^2)} dz = \int_{L_r} \frac{e^{itz}}{\pi(1+z^2)} dz + \int_{C_r} \frac{e^{itz}}{\pi(1+z^2)} dz$$

We have on L_r $z=x$ so $dz=dx$ then

$$\int_{L_r} \frac{e^{itz}}{\pi(1+z^2)} dz = \int_{-r}^r \frac{e^{itx}}{\pi(1+x^2)} dx$$

Now on C_r , we have $|e^{itz}| \leq 1$ and $|\frac{1}{1+z^2}| \leq \frac{1}{r^2-1}$ thus we obtain

$$\left| \int_{C_r} \frac{e^{itz}}{\pi(1+z^2)} dz \right| \leq \frac{1}{\pi(r^2-1)} \pi r = \frac{r}{r^2-1} \rightarrow 0 \text{ as } r \rightarrow \infty$$

Let $g(x) = \frac{e^{itz}}{\pi(1+z^2)}$

We notice that $g(x)$ has one singularity inside Γ_r at i . Thus, the residue

$$\lim_{z \rightarrow i} (z-i) \frac{e^{itz}}{\pi(1+z^2)} = \lim_{z \rightarrow i} \frac{e^{itz}}{\pi(z+i)} = \frac{e^{-t}}{2\pi i}$$

Hence by Cauchy's integral formula,

$$\int_{\Gamma_r} \frac{e^{itz}}{\pi(1+z^2)} dz = 2\pi i \frac{e^{-t}}{2\pi i} = e^{-t}$$

Letting $r \rightarrow \infty$ we get

$$\int_{-\infty}^{\infty} \frac{e^{itx}}{\pi(1+x^2)} dx = e^{-t}$$

For $t < 0$, choose the change of variable $u = -x$ to get

$$\int_{-\infty}^{\infty} \frac{e^{itx}}{\pi(1+x^2)} dx = \int_{-\infty}^{\infty} \frac{e^{i(-t)u}}{\pi(1+u^2)} du = e^t$$

Hence the characteristic function of the Cauchy distribution is given by $\Theta(t) = \exp(-|t|)$.

1.1.4 Ratio of independent normal variables

Let $X = \frac{Y}{Z}$ where both Y and Z follow a standard normal distribution. Then X follows a standard Cauchy distribution. By definition, Z^2 follows the Chi-Square distribution with a degree of freedom equal to 1 and independent with Y . Hence also by definition $X = Y/\sqrt{Z^2} = Y/Z$ has the Student t distribution with 1 degree of freedom. Using the general formula for the student t PDF, we see that the PDF of $\frac{Y}{Z}$ is

$$t \mapsto \frac{\Gamma(1)}{\sqrt{\pi}\Gamma(1/2)}(1+t^2)^{-1} = \frac{1}{\pi} \frac{1}{1+t^2}$$

1.2 General Cauchy Distribution

The Cauchy distribution is generalized by adding scale and location parameters. Suppose that Z has the standard Cauchy distribution. Now, consider $X = a + bZ$ where $a \in \mathbb{R}$ and $b \in (0, \infty)$. Then, X has the Cauchy distribution with location parameter a and scale parameter b .

1.2.1 Probability density function

Suppose that X has the Cauchy distribution with location parameter $a \in \mathbb{R}$ and scale parameter $b \in (0, \infty)$.

The probability density function is given by

$$g(x) = \frac{b}{\pi[b^2 + (x-a)^2]}, \quad x \in \mathbb{R}$$

Proof. Let $g(x) = \frac{1}{b} f\left(\frac{x-a}{b}\right)$ where f is the standard Cauchy PDF. □

1.2.2 Cumulative distribution function

X has a distribution function given by $G(x)$.

$$G(x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-a}{b}\right), \quad x \in \mathbb{R}.$$

Proof. $G(x) = F\left(\frac{x-a}{b}\right)$

□

1.2.3 Simulating Cauchy Random Variables

If U has the standard uniform distribution, then $X = a + b \tan\left[\pi\left(U - \frac{1}{2}\right)\right]$ has the Cauchy distribution with location parameter a and scale parameter b .

Proof. Let $X = F^{-1}(U)$ where $U \sim U(0,1)$ then X has the same distribution as F . Now, $P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$. As a result, we set $F(x) = U$ and solve for X in terms of U to get the result. □

Chapter 2

Least Squares Method

In the real world linear relations exist in many aspects. For example, the force of the spring (y) linearly depends on the displacement of the spring (x) where $y=kx$ and k is the spring constant. Moreover, the gravitational potential energy linearly depends on the height of the object where it is related to the height by this linear equation $E_p=mgh$. The mass of the object is denoted by m and the gravity is denoted by g .

Unfortunately, it is unlikely that we observe a perfect linear relationship but rather approximately linear. The method of least square method finds the best fitting straight line for a set of points as seen later on.

2.1 Statistics Review

Given a sequence of data x_1, x_2, \dots, x_N , we define the mean by $\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$. The mean is the average value of the data.

The variance of the data is denoted by σ^2 where $\sigma^2 = \frac{1}{N} \sum_{n=1}^N (x_i - \bar{x})^2$. The standard deviation is the square root of the variance:

$$\sigma = \sqrt{\frac{1}{N} \sum_{n=1}^N (x_i - \bar{x})^2}$$

We mean by best fitting straight line $y = ax + b$ that $y - (ax + b)$ must be zero. Given the N -observations $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, we look at $y_1 - (ax_1 + b), y_2 - (ax_2 + b), \dots, y_N - (ax_N + b)$.

The variance of the given data set is $\sigma^2 = \frac{1}{N} \sum_{n=1}^N (y_n - (ax_n + b))^2$. We use the squaring instead of the absolute value to give higher weight to large errors than small ones. Also, the absolute value function is not differentiable which makes the tools of calculus inaccessible.

2.2 The Method of Least Squares

The general linear model assumes that the dependent variable Y is determined by one or more factors X_t in a given linear relationship

$$Y = X\beta + \varepsilon$$

$Y = m \times 1$ is a vector of m observations

$X = m \times (n + 1)$ is a matrix of $(n + 1)$ observations for m regressors

$\beta = (n + 1) \times 1$ is a parameter vector

$\varepsilon = m \times 1$ is a vector of m values for the noise

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{1n} \\ 1 & X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{m1} & X_{m2} & \dots & X_{mn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix}$$

Our goal is to obtain estimates of the population parameters in the β -vector. The estimates of β is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Proof. Define the objective function

$$S = \|Y - \beta X\|^2 = \sum_{i=1}^m |y_i - \sum_{j=0}^n X_{ij}\beta_j|^2$$

Let $\varepsilon_i = y_i - \sum_{j=0}^n X_{ij}\beta_j$ then

$$S = \sum_{i=1}^m \varepsilon_i^2 = (Y - X\beta)^T (Y - X\beta) = \varepsilon^T \varepsilon$$

S is minimized when its gradient vector is zero. The elements of the gradient vector are the partial derivatives of S with respect to the parameters:

$$\frac{dS}{d\beta_j} = 2 \sum_{i=1}^m \varepsilon_i \frac{d\varepsilon_i}{d\beta_j}$$

we have

$$\frac{d\varepsilon_i}{d\beta_j} = -X_{ij}$$

thus substituting in (1) we get:

$$\frac{dS}{d\beta_j} = 2 \sum_{i=1}^m (y_i - \sum_{k=0}^n X_{ik}\beta_k)(-X_{ij})$$

$\hat{\beta}$ minimizes S when

$$2 \sum_{i=1}^m (y_i - \sum_{k=0}^n X_{ik} \beta_k) (-X_{ij}) = 0$$

Thus we get

$$\sum_{i=1}^m \sum_{k=0}^n X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^m X_{ij} y_i$$

hence we get $\hat{\beta} = (X^T X)^{-1} X^T Y$. □

2.3 The Gauss-Markov Assumptions

1. $Y = X\beta + \varepsilon$.

This assumption states that there is a linear relationship between Y and X .

2. X is $n \times k$ matrix of full rank.

We mean the columns of X are linearly independent. This assumption is known as the identification condition.

3. $E(\varepsilon | X) = 0$.

The zero conditional mean assumption states that the stochastic errors average to zero for any value of X . This implies that $E(Y) = E(X\beta)$.

4. $E(\varepsilon \varepsilon^T) = \sigma^2 I$ where $\Omega = \sigma^2 I$ is the variance-covariance matrix of the stochastic error.

5. X may be fixed or random but it should be generated by a mechanism where it is unrelated to ε .

2.4 Gauss-Markov Theorem

The Gauss-Markov theorem states, based on the assumptions mentioned above, that the least squares method estimator is the best linear, unbiased and efficient estimator (BLUE).

2.4.1 Unbiased

$\hat{\beta}$ is an unbiased estimator of β .

Proof. We have shown that $\hat{\beta} = (X^T X)^{-1} X^T Y$, and we have $Y = X\beta + \varepsilon$. This means that

$$\begin{aligned}
\widehat{\beta} &= (X^T X)^{-1} X^T X \beta + \varepsilon \\
\widehat{\beta} &= \beta + (X^T X)^{-1} X^T \varepsilon \\
E(\widehat{\beta}) &= E(\beta) + E((X^T X)^{-1} X^T \varepsilon) \\
E(\widehat{\beta}) &= E(\beta) + (X^T X)^{-1} X^T E(\varepsilon) \\
E(\widehat{\beta}) &= E(\beta) + 0
\end{aligned}$$

□

$E(\widehat{\beta}) = E(\beta)$ since $E(\varepsilon) = 0$ Thus, $\widehat{\beta}$ is an unbiased estimator of β .

2.4.2 Linear

We want to show that $\widehat{\beta}$ is a linear estimator.

Proof. $\widehat{\beta} = \beta + (X^T X)^{-1} X^T \varepsilon$.

Take $A = (X^T X)^{-1} X^T$ then we can write $\widehat{\beta} = \beta + A\varepsilon$. Hence, $\widehat{\beta}$ is a linear estimator. □

2.4.3 Minimum Variance

Remember that our goal is to find the estimator $\widehat{\beta}$ that minimizes the sum of the squared residuals ($\sum_{i=1}^n \varepsilon_i^2$).

The vector of residuals is given by $\varepsilon = Y - X\widehat{\beta}$

The sum of the square residuals is given by $\varepsilon^t \varepsilon$.

We have

$$\begin{aligned}
\varepsilon^t \varepsilon &= (Y - X\widehat{\beta})^t (Y - X\widehat{\beta}) \\
&= Y^t Y - \widehat{\beta}^t X^t Y - Y^t X \widehat{\beta} + \widehat{\beta}^t X^t X \widehat{\beta} \\
&= Y^t Y - 2\widehat{\beta}^t X^t Y + \widehat{\beta}^t X^t X \widehat{\beta}
\end{aligned}$$

since the transpose of a scalar is a scalar.

To find $\widehat{\beta}$ that minimizes the sum of the squared residuals we find the partial derivative of $\varepsilon^t \varepsilon$ with respect to $\widehat{\beta}$.

$$\frac{d\varepsilon^t \varepsilon}{d\widehat{\beta}} = -2X^t Y + 2X^t X \widehat{\beta} \tag{2.1}$$

To check that $\widehat{\beta}$ minimizes 2.1, we take the partial derivative of 2.1 with respect to $\widehat{\beta}$ again to obtain:

$$d^2 \frac{\varepsilon^t \varepsilon}{d^2 \widehat{\beta}} = 2X^t X$$

$2X^t X$ is a positive definite matrix hence a minimum.

2.5 Normality of the Stochastic Error

The general linear model

$$Y = X\beta + \varepsilon$$

with all of its properties mentioned above kept the same. ε follows a normal distribution where

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

as seen in section 3.2.

In this case, where the stochastic errors follow a normal distribution the least squares method turns out to be equivalent to the maximum likelihood method.

Proof. First, consider the density function for a single error term. ε_i follows a normal distribution $N(0, \sigma^2)$.

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\varepsilon_i^2}{2\sigma^2}\right\}$$

We have $\varepsilon_i = y_i - x_i\beta$ with x_i as the i -th row of the matrix X . Now, y_i is a linear function of ε_i hence it will be normally distributed.

Now, consider

$$L = \prod_{i=1}^n f(y_i, x_i, \beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^t(y - X\beta)\right\}$$

Consider the logarithm of this function

$$L^*(\beta, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^t(y - X\beta)$$

We take the partial derivative of L^* with respect to β and σ^2 . As a result, we get

$$\frac{dL^*}{d\beta} = \frac{1}{\sigma^2}(Y - X\beta)^t X \tag{2.2}$$

$$\frac{dL^*}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)'(Y - X\beta) \quad (2.3)$$

To find the maximum likelihood estimator of β we set 2.2 equal to zero to get $\hat{\beta} = (X^T X)^{-1} X^T Y$ as found in the least squares method.

To find the maximum likelihood estimator of σ^2 we set 2.3 equal to zero to get $\hat{\sigma}^2 = \frac{1}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta})$.

This is the same as least squares method in large samples. Now, it is clear that the maximum likelihood and least squares method estimates are equivalent when the error terms are assumed to be normally distributed. \square

Chapter 3

Cauchy Stochastic Errors

The assumption of normality for the stochastic error has been basic to nearly all statistical inference. However, the departure from normality haven't been taken seriously. In this paper, we will study the distribution of the least squares method estimates considering stochastic errors following a Cauchy distribution. Moreover, we will concentrate on the properties of the least squares method estimates as the stochastic errors follow a Cauchy distribution.

Now, consider the general linear model

$$Y = X\beta + \varepsilon$$

$Y = m \times 1$ is a vector of m observations

$X = m \times (n + 1)$ is a matrix of $(n + 1)$ observations for m regressors

$\beta = (n + 1) \times 1$ is a parameter vector

$\varepsilon = m \times 1$ is a vector of m values for the noise

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_m \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdot & \cdot & \cdot & X_{1n} \\ 1 & X_{21} & X_{22} & \cdot & \cdot & \cdot & X_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & X_{m1} & X_{m2} & \cdot & \cdot & \cdot & X_{mn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_m \end{bmatrix}$$

where $\varepsilon \sim C(0, \theta)$.

We have seen that $\hat{\beta} = (X^T X)^{-1} X^T Y$.

It will be shown that $\hat{\beta}$ follow a Cauchy distribution as well.

Proof. we have proved in chapter 3 that

$$\hat{\beta} = \beta + (X^T X)^{-1} X^T \varepsilon$$

The proof proceeds from the characteristic function. We consider the first element of the β - vector. we get

$$\hat{\beta}_1 = \beta_1 + \sum_{t=1}^m C_{1t}\varepsilon_t$$

where C_{1t} is the first row of $(X^t X)^{-1} X^t$; $i = 1, \dots, m$.

The characteristic function of a given ε_t , centered at zero with scale parameter θ_t is given as:

$$\Theta(s) = E\{\exp(is\varepsilon_i)\} = E\{(\exp(is\theta_t))\} = \exp(-|s\theta_t|)$$

Consequently, the characteristic function for the sum of $C_{1t}\varepsilon_t$ is given by

$$\prod_{t=1}^m \exp\{-|sC_{1t}|\theta_t\} = \exp\{-|s| \sum_{t=1}^m |C_{1t}|\theta_t\} \quad (3.1)$$

□

Equation 3.1 shows that $\sum_{i=1}^m C_{1t}\varepsilon_t$ follows a Cauchy distribution centered at zero with scale parameter $\sum_{t=1}^m |C_{1t}|\theta_t$. Then $\hat{\beta}$ follows a Cauchy distribution with β location parameter and $\sum_{t=1}^m |C_{1t}|\theta_t$ as a scale parameter.

Accordingly, the deviations of $\hat{\beta}_1$ from β_1 may also be shown to obey the following statement.

$$P \left[|\hat{\beta}_1 - \beta_1| < \sum_{t=1}^m |C_{1t}|\theta_t \right] = 0.5 \quad (3.2)$$

Proof. Let $X = \hat{\beta}_1 - \beta_1$ and $b = \sum_{t=1}^m |C_{1t}|\theta_t$
Then we have

$$P[|X| < b] = \int_{-b}^b \frac{b dx}{\pi[b^2 + x^2]} = \int_{-b}^b \frac{b dx}{\pi b^2[1 + \frac{x^2}{b^2}]} = \frac{1}{\pi} \arctan \frac{x}{b} \Big|_{-b}^b = 0.5$$

□

3.1 Law of Large numbers and Central Limit Theorem

The law of large numbers and central limit theorem don't apply for the Cauchy errors, consequently the justification for the use of least squares method with a model of Cauchy errors should be tested. As we have seen in equation 3.2 that the deviation of $\hat{\beta}$ from β is

independent of the sample size and the Cauchy scale parameter. This will be tested by an experiment which will be discussed later on.

For now, we will simulate standard Cauchy random variables vs standard normal random variables to check whether the sample mean is representative of the population mean.

With the help of a statistical software, we will generate first standard Cauchy random variables and we will calculate the sample mean as the sample size increases. This will be illustrated in the following figure

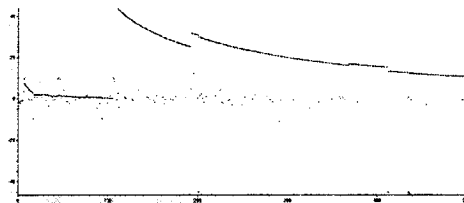


Figure 3.1: Variation of sample mean as a function of sample size (Cauchy case)

The grey dots represent the simulated standard Cauchy random variables whereas the blue dots represent the mean of these grey dots as sample size increases.

Because the parameters of the Cauchy distribution don't correspond to a mean and variance, attempting to estimate the parameters of the Cauchy distribution by using a sample mean and a sample variance will not succeed.

One can obviously see how the sample mean varies as a function of the sample size which ensures that the estimation process won't work.

Now, let us have a look at the simulated standard normal random variables. We will apply the same procedure i.e. we will calculate the sample mean as the sample size increases. The result is illustrated in the following figure

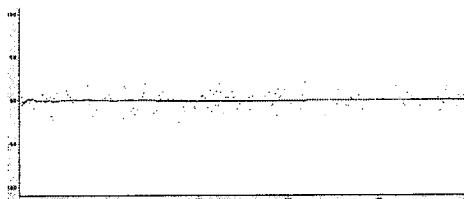


Figure 3.2: Variation of sample mean as a function of sample size (normal case)

One can obviously notice that the mean is almost the same as the sample size increases.

Thus, we can conclude that the law of large numbers which states that as the number of identically distributed, randomly generated variables increases, their sample mean (average) approaches their theoretical mean fails in the case of Cauchy errors.

3.2 Sampling distribution of the mean

In this section, 1000 samples of size 100 were drawn from a standard cauchy distribution. We will calculate the mean of each sample to give us an idea about the sampling distribution of the mean. This simulation is done using excel; here are the descriptive statistics for these 1000 sample means:

mean	minimum	Q1	median	Q3	Maximum	standard deviation
-0.026	-1010	-1.174	0.047	1.1369	218	58.56

Table 3.1: Summary Table

One can notice that the minimum and the maximum values are very extreme outliers. Moreover, we can notice that the standard deviation is much larger than the interquartile range. A scatter plot has been performed to visually detect the outliers.

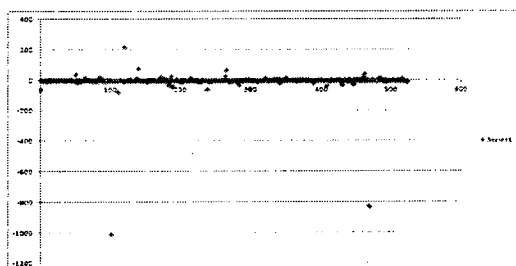


Figure 3.3: Visual detection of outliers

3.3 Outliers

Definition 3.1. *An outlier is an observation that is distant from other observations. The common sources of outliers are measurement error and experimental error. Outliers may also exist by chance, but they are often indicative of measurement error or heavy-tailed distribution of the population. In the former case, we use robust statistics to outliers while in the latter case they indicate that the distribution has high kurtosis.*

One can note that the Cauchy distribution is a heavy tailed one due to the non- existence of the moment generating function for all $t > 0$.

In our case, we are dealing with errors that follow a Cauchy distribution instead of a normal one. Outliers can have deleterious effects on statistical analyses. They increase error variance and reduce the power of statistical tests. Moreover, they seriously bias or influence estimates which is obvious in our case.

3.4 How to deal with outliers?

Robust methods are robust against the presence of outliers. Common robust estimation are the use of trimmed mean or the Winsorized mean. We mean by the trimmed mean is that we eliminate extreme observations at both ends of the sample while Winsorized mean replaces the extreme residuals with the next closest value in the dataset .

Our main point in this chapter is to study the effect of Cauchy errors on the bias and efficiency of the β – estimate. Thus our next step will cover an experimental method to asses which factors influence the performance of β – estimate.

3.5 Experimental Method

A one regressor model given in 3.3 will be used as a frame of reference for the sampling experiments. One hundred experiments were performed to estimate β_0 and β_1 .

$$Y_t = \beta_0 + \beta_1 X_t + C_t \tag{3.3}$$

These experiments are distinguished by the sample size (20 and 100 observations) and the Cauchy spread parameter θ .

The Cauchy errors will be generated according to this equation:

$$C = m + \theta \tan(\pi(U - 0.5))$$

where m is the location parameter, U a uniformly random variable and θ is the Cauchy scale parameter.

Ten values of θ ($\theta = 1, 2, \dots, 10$) each for the two sample sizes will be taken.

Due to the lack of real data, we'll consider a hypothesized model with $\beta_0 = 1$ and $\beta_1 = 5$. We construct this data by adding the Cauchy errors to a certain data for X and to get the Y values.

The OLS estimates of β_0 and β_1 are computed starting with a sample size of 20 observations.

The following table shows the mean, median, bias and the RMSE (root mean square error) of β - estimates.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
Theta	mean	mean	median	median	bias	bias	RMSE	RMSE
1.00	7.98	5.15	1.31	5.00	6.98	0.15	19.93	0.07
2.00	14.92	5.13	1.63	5.00	13.92	0.13	39.85	0.15
3.00	21.86	5.10	1.94	5.00	20.86	0.10	59.78	0.22
4.00	28.80	5.08	2.25	5.00	27.80	0.08	79.71	0.29
5.00	35.74	5.06	2.57	5.00	34.74	0.06	99.64	0.37
6.00	42.68	5.03	2.88	5.00	41.68	0.03	119.56	0.44
7.00	49.62	5.01	3.20	4.99	48.62	0.01	139.49	0.51
8.00	56.56	4.99	3.51	4.99	55.56	-0.01	159.42	0.59
9.00	63.51	4.96	3.82	4.99	62.51	-0.04	179.34	0.66
10.00	70.45	4.94	4.14	4.99	69.45	-0.06	199.27	0.73

Table 3.2: A small sample size result table, n=20

One can see that the increase in the Cauchy spread parameter, θ , tends to worsen the OLS performance of the β -estimates. However, we can notice that the estimated parameter $\hat{\beta}_1$ seems to evidence less sensitivity to that increase.

Examining the values of the mean of $\hat{\beta}_0$ as the scale parameter increases by one unit, we can see that the mean corresponding to scale parameter 1 tends to be multiplied by the values of θ . This huge increase worsen the estimation of $\hat{\beta}_0$ which is obviously tested using Bias and the root mean squared error (RMSE). We conclude that the mean is a biased estimator which doesn't represent the population.

As a next step, I've estimated the median for β -estimates to test its performance as the scale parameter increases. It turned out as shown in the table that the median is a better estimator for both β_0 and β_1 . Now, the same experiment is done but dealing with a sample size of 100. The following table shows the mean, median, bias and the RMSE (root mean square error) of β - estimates as above.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
Theta	mean	mean	median	median	bias	bias	RMSE	RMSE
1.00	-1.09	5.18	0.76	5.00	-2.09	0.18	6.13	0.03
2.00	-3.22	5.19	0.52	5.00	-4.22	0.19	12.26	0.06
3.00	-5.34	5.19	0.28	5.00	-6.34	0.19	18.39	0.09
4.00	-7.47	5.20	0.04	5.00	-8.47	0.20	24.52	0.13
5.00	-9.59	5.21	-0.20	5.00	-10.59	0.21	30.65	0.16
6.00	-11.72	5.22	-0.44	5.00	-12.72	0.22	36.78	0.19
7.00	-13.84	5.22	-0.68	5.01	-14.84	0.22	42.91	0.22
8.00	-15.97	5.23	-0.92	5.01	-16.97	0.23	49.05	0.25
9.00	-18.09	5.24	-1.16	5.01	-19.09	0.24	55.18	0.28
10.00	-20.22	5.25	-1.40	5.01	-21.22	0.25	61.31	0.32

Table 3.3: A large sample size result table, n=100

The increase in the sample size (100) made the performance of the mean for β -estimates better than dealing with a sample size of 20. The OLS performance pattern is mainly focused on the estimation of β_0 since it is clearly noticed that $\hat{\beta}_1$ is much less sensitive to the sample size and scale parameter variation.

Although this increase has lessened the bias and root mean squared error of the mean, it still doesn't represent the real β 's which all goes back to the large sensitivity that $\hat{\beta}_0$ shows. Also, we can notice that the median is a much better estimator than the mean.

This experiment has shown that the sample size and the Cauchy spread parameters are two main factors that affect the performance of OLS. Moreover, this experiment has shed the light on the costs in terms of estimator performance associated with using conventional methods such as OLS and MLE as it will be shown later on. Using such methods to estimate the parameters of a linear regression model where the errors follow a Cauchy distribution maybe misleading taking into consideration the infinite variance of the Cauchy errors. Furthermore, we conclude that as the scale parameter and the sample size increases we totally doubt the performance of OLS method due to the high probability of outliers occurrence that messes that whole estimation process leading to a huge variance.

As a next step, we will test the performance of Maximum Likelihood Estimation in estimating beta parameters. This experiment will allow us to compare the results of both .

Chapter 4

Maximum Likelihood Estimation

Definition 4.1. Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. The Maximum likelihood estimate of parameter θ is the value of θ which maximizes the likelihood $L(\theta)$. For data values of n -sample x_1, x_2, \dots, x_n the likelihood is given by

$$L_n(\theta) = \prod_i f_X(x_i, \theta)$$

This is equivalent to maximize

$$\ell(\theta) = \text{Log}(L_n(\theta)) = \sum_i \log f_X(x_i, \theta)$$

since \log is an increasing function.

Definition 4.2. Likelihood

The likelihood is defined as the joint density or probability of the outcomes, with the roles of the values of the outcomes \mathbf{y} and the values of the parameters $\boldsymbol{\theta}$ interchanged. Let $f(\mathbf{y}, \boldsymbol{\theta})$ be a class of joint densities with the parameter vector $\boldsymbol{\theta}$ in a set (parameter space) Θ . The likelihood is defined as the function

$$L(\boldsymbol{\theta}, \mathbf{y}) = f(\mathbf{y}, \boldsymbol{\theta})$$

The maximum likelihood estimator of θ for the model given by the joint densities or probabilities $f(\mathbf{y}, \boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \Theta$, is defined as the value of $\boldsymbol{\theta}$ at which the corresponding likelihood $L(\boldsymbol{\theta}, \mathbf{y})$ attains its maximum:

$$ML_{\hat{\boldsymbol{\theta}}} = \text{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{y})$$

This definition is not complete because there is no guarantee that such a maximum exists or, when it does exist, it is unique. However, in many settings this definition turns out to be very useful and constructive, yielding an estimator with good properties.

4.1 MLE of Cauchy distribution with zero mean and unknown scale parameter

We will consider the estimation of the scale parameter b of a Cauchy distribution with zero mean and an unknown scale parameter say $C(m=0, b)$.

Haas et al. (1970) have shown that \hat{b} , the maximum likelihood of b , does exist and it is unique. We will solve the MLE using numerical methods due to the lack of existence of a closed form solution.

4.2 Location parameter Known

The problem of estimating \hat{b} using MLE of b from the Cauchy distribution with known m , is considered to be a simpler problem from estimating the location parameter given the scale one. This will be shown according to the following theorem.

Theorem 2. *Let $L(x, b)$ be the likelihood function for the Cauchy distribution with known location parameter m then there exists a unique \hat{b} such that*

$$\left. \frac{\partial \log L(x, b)}{\partial b} \right|_{b=\hat{b}} = 0.$$

Proof. The Likelihood function as defined previously is given by

$$L(x, b) = \prod_{i=1}^n \frac{1}{\pi b [1 + \frac{x_i^2}{b^2}]}$$

and

$$\log L(x, b) = \sum_{i=1}^n -\log(\pi b [1 + \frac{x_i^2}{b^2}])$$

or

$$\log L(x, b) = -n \log \pi - n \log b - \sum_{i=1}^n \log [1 + \frac{x_i^2}{b^2}]$$

Hence,

$$\frac{\partial \log L(x, b)}{\partial b} = -\frac{n}{b} + \frac{2}{b} \sum_{i=1}^n \frac{x_i^2}{b^2 + x_i^2}$$

Now, we seek the solution of

$$h(b) = \frac{\partial \log L(x, b)}{\partial b} = -\frac{n}{b} + \frac{2}{b} \sum_{i=1}^n \frac{x_i^2}{b^2 + x_i^2} = 0$$

Notice that h is a continuous function and $h(0) = n$ and $h(\infty) = -n$ and

$$h'(b) = -2b \sum_{i=1}^n \frac{x_i^2}{(b^2 + x_i^2)^2} < 0$$

for all $b \geq 0$. Then it follows that there is a unique \hat{b} such that $h(\hat{b}) = 0$.

The function $h(b)$ is nonlinear in b . To find \hat{b} we'll use Newton-Raphson method starting with a value that causes convergence. \square

4.3 Newton Raphson Method

The Newton-Raphson method is used to solve equations of the form $f(x) = 0$. First, we begin by making an initial guess for the root we are trying to find, we call this initial guess x_0 . Now, the sequence $x_0, x_1, x_2, \dots, x_n, \dots$ generated in the manner described below should converge to the exact root. To implement it analytically we need a formula relating each approximation with the previous one i.e x_{n+1} in terms of x_n .

This method is constructed by finding the equation of the tangent line of $f(x)$ at the point $(x_0, f(x_0))$. The tangent line is given by this equation :

$$y - f(x_0) = f'(x_0)(x - x_0)$$

The tangent line intersects the x-axis when $y = 0$ $x = x_1$, so solving for x_1 we get:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

and more generally we get:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

4.4 Multi-Dimensional Case For Newton Raphson Method

Now, consider the system of n non-linear equations and n unknowns.

Consider

$$\begin{aligned}f_1(x_1, x_2, x_3, \dots, x_n) &= 0 \\f_2(x_1, x_2, x_3, \dots, x_n) &= 0 \\f_3(x_1, x_2, x_3, \dots, x_n) &= 0 \\f_4(x_1, x_2, x_3, \dots, x_n) &= 0 \\&\vdots \\f_n(x_1, x_2, x_3, \dots, x_n) &= 0\end{aligned}$$

One technique used to solve this problem is called the Multivariate Newton Raphson Method (MNRM). The basic idea comes from the fact the the derivative of a function of two variables , f_j , is

$$df_j = \frac{df_j}{dx_1} dx_1 + \frac{df_j}{dx_2} dx_2$$

For the n variable case we have:

$$df_j = \sum_{i=1}^n \frac{df_j}{dx_i} dx_i$$

where j represents the index over the functions, i the index over the variables and the superscript in parenthesis stands for iterations. We suppose now that we have n equations so j goes from 1 to n. This system can be written of the form:

$$f(x^{(2)}) - f(x^{(1)}) = \sum_{i=1}^n \frac{df}{dx_i} (x_i^{(2)} - x_i^{(1)})$$

We want our next iteration to lead us to the root so we take $f(x^{(2)}) = 0$ Thus the iterative method for solving a system of n non-linear equations and n unknowns is given by:

$$x^{k+1} = x^k - J^{-1} f(x^k)$$

where

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

4.5 Experimental Method

In this experiment, we will test the same data used in the OLS method. Sample sizes of 20 and 100 will be taken to examine the performance of the MLE estimators in the presence of Cauchy errors. As we mentioned before, we need a starting value which is close to the real solution to run the multivariate Newton-Raphson method to insure convergence.

This experiment is implemented using R program with the aid of RootSolve and MASS packages and the solution of the likelihood function is given to the nearest 10^{-12} .

To find the maximum likelihood estimates for regression parameters with Cauchy errors, we just look at that likelihood:

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^n \frac{1}{\pi\sigma \left(1 + \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2\right)}$$

We take the log of the likelihood function $\ell(\beta_0, \beta_1, \sigma) = \log(L(\beta_0, \beta_1, \sigma))$, set the partial derivatives equal to zero as given below and solve the obtained system using multivariate Newton Raphson method.

After differentiating we get the following non-linear system:

$$\begin{aligned} \frac{\partial \ell(\beta_0, \beta_1, \sigma)}{\partial \beta_0} &= \sum_{i=1}^n \frac{2(y_i - \beta_0 - \beta_1 x_i)}{(\sigma^2 + (y_i - \beta_0 - \beta_1 x_i)^2)} &= 0 \\ \frac{\partial \ell(\beta_0, \beta_1, \sigma)}{\partial \beta_1} &= \sum_{i=1}^n \frac{2x_i(y_i - \beta_0 - \beta_1 x_i)}{(\sigma^2 + (y_i - \beta_0 - \beta_1 x_i)^2)} &= 0 \\ \frac{\partial \ell(\beta_0, \beta_1, \sigma)}{\partial \sigma} &= \frac{n}{\sigma} - \sum_{i=1}^n \frac{2\sigma}{(\sigma^2 + (y_i - \beta_0 - \beta_1 x_i)^2)} &= 0 \end{aligned}$$

To solve this system using multivariate Newton Raphson method we need an initial guess. This initial guess will be the vector $c = (\text{median}\hat{\beta}_0, \text{median}\hat{\beta}_1, IQR(\text{error})/2)$ where IQR is the interquartile range of the estimated error vector. This choice is based on the

results we have discussed in chapter 2. Moreover, we choose the IQR since it is a robust estimator of the scale parameter.

In the case of the sample size 20, we run the R program 10,000 times to get the vector of $\hat{\beta}_0$, $\hat{\beta}_1$ and the estimated scale parameter of the Cauchy distribution $\hat{\sigma}$. The result is $c=(1.3,5,1.01)$. As we can see that $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}$ are very close to the real emphasizing that the multivariate Newton Raphson method has converged very close to the real parameters in the case of a small sample size.

In the case of the sample size 100, we get the estimated vector $c=(1.3,5,1.00)$ where we can notice that the estimation of β_0 and β_1 and σ are also very close to the real parameters (1,5,1) illustrating that the method has also converged to the same point for big sample size.

We can notice that solving the log of the maximum likelihood function using multivariate Newton Raphson method has certainly gave very impressive results which are very close to the real ones. Moreover, one should be careful of the choice of the starting point since this method is expected to converge only near the solution. As a result, we can say that the MLE method is much more preferable than the OLS method regarding sample size factor and limited scale parameter.

As a next step, we will test the MLE while increasing our scale parameter to check whether the losses may be torelable or not. In our former case, the scale parameter was 1 and it turned out that the estimation of the parameters was very close to the real ones. Now, we will see the effect of the increase of the scale parameter on the MLE solution as shown in the following table.

The results below are computed for a sample size of 20.

σ	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$
1.00	1.31	5.00	1.01
2.00	1.31	4.99	2.11
5.00	1.31	4.99	5.29
10.00	1.31	5.00	10.57

Table 4.1: MLE estimators, n=20

Now, the results are computed for a sample size of 100.

σ	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$
1.00	1.30	5.00	1.00
2.00	1.3	5.00	2.03
5.00	1.3	5.00	5.07
10.00	1.3	5.00	10.10

Table 4.2: MLE estimators, n=100

N.B: The results of both tables 4.1 and 4.2 are taken to be the mean of 10000 runs.

As we can see, that the increase in the Cauchy spread parameter has no effect on the estimation of the linear model parameters while using multivariate Newton Raphson method. We can notice that that estimated parameters are insensitive to both changes: sample size and Cauchy scale parameter. As a result, MLE could be used in the parameter estimation for linear models with Cauchy errors.

In conclusion, we can say that the MLE results using Newton Raphson method were valid to equation 3.2 which stated that the convergence of β 's are independent of the sample size and the Cauchy scale parameter unlike the OLS method. These results have unified the theoretical and experimental aspects in the case of MLE contrary to the OLS results. Moreover, we can't ignore the importance of OLS in providing a starting point for the Newton Raphson iterations with the help of robust estimators such as the median and the interquartile range.

4.6 Optimum Properties of Maximum Likelihood Estimation

Let us consider the maximum likelihood estimation of the parameter θ , which is to be estimated on the basis of a random sample from a density $f(\cdot; \theta)$, where θ is assumed to be a real number. That is, let us consider the unidimensional-parameter case and estimate θ itself. Recall that for the observed sample x_1, x_2, \dots, x_n the maximum likelihood estimate of θ is the value, say $\hat{\theta}$, of θ which maximizes the likelihood function $L(\theta, x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta)$. Let $\hat{\Theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ denote the maximum likelihood estimator of θ based on a sample of size n.

Theorem 3. *If the density $f(x; \theta)$ satisfies certain regularity conditions and if $\hat{\Theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ is the maximum likelihood estimator of θ for a random sample of size n from $f(x; \theta)$, then:*

$\hat{\Theta}_n$ is asymptotically normally distributed with mean θ and variance $1/n E_{\theta} \left[\left[\frac{\partial}{\partial \theta} \log f(X; \theta) \right]^2 \right]$.

4.7 Disadvantages of Newton Raphson Method

- ◆ The method is very time consuming and computationally challenging. We mean the calculation of the inverse of the Jacobian matrix and the evaluation of the function and its derivative.
- ◆ The method doesn't converge if the tangent is parallel or nearby parallel to the x-axis.
- ◆ Usually the Newton method is expected to converge only near the solution.

Conclusion

The objective of this paper was to test the effect of Cauchy errors in a linear model on the performance of the OLS and MLE estimator. In the OLS method, one hundred sampling experiments were performed to test the effect of the sample size and the Cauchy spread parameter on the performance of the OLS estimators. It turned out that the sample size and the Cauchy spread parameter affect the performance of OLS. In this study, a one regressor linear model was tested and the results showed that $\hat{\beta}_0$ was very sensitive to the increase in the sample size and the Cauchy scale parameter contrary to $\hat{\beta}_1$.

In the MLE case, we encountered the initial guess problem in which we took the values of the sample median and the interquartile range of the error vector to start with. Moreover, we have shown that the derivative of log of the likelihood function has a unique root in the case of known location parameter and unknown scale parameter. This unique root exhibited close values to real beta values in which it was insensitive to the increase of both factors; the sample size and the Cauchy scale parameter.

At the end, we can say that the application of OLS in a linear model with Cauchy errors is not representative in which it is important to be aware of the costs in terms of estimator performance associated with using such conventional method. On the other hand, the MLE is strongly recommended for the estimation of a linear model parameters although it is very time consuming and computationally demanding.

Appendix

This appendix will display the R-programming of multivariate Newton-Raphson method.

```
sum=c(0,0,0)
k=0
norm=1
for (i in 1:10000)
{
xvect=c(500*runif(100))
evect=c(10*tan(180*(runif(100)-0.5)))
yvect=1+5*xvect+evect
evect=yvect-1.31-5.0007272*xvect
Q1=quantile(evect,0.25)
Q3=quantile(evect,0.75)
IQR=Q3-Q1
c=c(1.31,5,IQR/2)
f=function(t=0,beta,parms = NULL,epsilon=10-12)c(f1=sum((2*1/beta[3]2*(yvect-beta[1]-
beta[2]*xvect))/(1+1/beta[3]2*(yvect-beta[1]-beta[2]*xvect)2)), f2=sum((2*xvect*1/beta[3]2*(yvect-
beta[1]-beta[2]*xvect))/(1+1/beta[3]2*(yvect-beta[1]-beta[2]*xvect)2)), f3=100/beta[3]-
sum(2*beta[3]/(beta[3]2+(yvect-beta[1]-beta[2]*xvect)2))

  while (norm > 10-12 && is.finite(norm)==TRUE && is.nan(norm)==FALSE )
  {
J=jacobian.full(y=c,func=f)
c=c-ginv(J)% * %f(t=0,beta=c)
delta=-ginv(J)% * %f(t=0,beta=c)
absdelta=abs(delta)
norm=sum(absdelta)
}
if(is.finite(norm)==TRUE && is.nan(norm)==FALSE)
sum=sum+c
k=k+1
}
solution=sum/k
print(solution)
```

Bibliography

- [1] SMITH, V. K. (1973), LEAST SQUARES REGRESSION WITH CAUCHY ERRORS. *Oxford Bulletin of Economics and Statistics*, 35: 223-231. doi: 10.1111/j.1468-0084.1973.mp35003004.x
- [2] Thomas S. Ferguson, *Journal of the American Statistical Association*, Vol. 73, No. 361 (Mar., 1978), pp. 211-213. doi: 10.2307/2286549
- [3] JOHN N. HADDAD, A Robust Maximum Likelihood Estimator for the Correlation coefficient, *International Journal of Mathematics and Computer Science*.
- [4] Karline Soetaert, Package RootSolve: rootemknus, gradients and steady-states in R, , *Royal Netherlands Institute of Sea Research (NIOZ) Yerseke, The Netherlands*.
- [5] Haas, Gerald Nicholas, "Statistical inferences for the Cauchy distribution based on maximum likelihood estimators" (1969). *Doctoral Dissertations*. Paper 2274.
- [6] D. Keffer, (1998), *ChE 301 Lecture Notes*
- [7] Charles J. Geyer, (2003), *Maximum Likelihood in R*.