Brunel University London

College of Engineering, Design and Physical Sciences

Numerical Comparison of Algorithms for Fitting Regression to a Burr Type XII Distribution

---

Author:
Murad Ahmed

Supervisor:
Professor Keming Yu

A dissertation submitted in part-fulfillment of the requirements for the master's degree in Statistics with Data Analytics

September 6, 2018

Abstract

Extreme distributions have wide applications in risk analysis. One of the extreme distribution is the Burr type XII distribution. Let Y follow a Burr type XII distribution and once covariate information $\boldsymbol{x}$ is available, Beirlant et al. considered a regression model via one shape parameter

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}).$$

Three algorithms were discussed in this paper and how it estimates the regression model parameters. The three algorithms are maximum likelihood estimation; ordinary least squares with one shape parameter being held constant and the ordinary least squares with shape parameter being unknown. This paper aims to compare the estimation accuracy of the three methods. In addition, the second part of the paper was to known how one of the algorithms can overcome big data problems. Therefore, the paper discussed and showed how the ordinary least squares with one shape parameter being held constant deals with big data.
All three algorithms were implemented on several sample sizes namely 25, 250, 750 and 1000. Results have shown that maximum estimation, had the best accuracy for the parameters for large samples and the least for small samples. The results from the ordinary least squares with one shape parameters being held constant was very accurate for a small sample but poor large samples. Results from the ordinary least squares without the shape parameter being constant was very accurate in general although it does overestimate for small samples.
It was recommended to compare the three algorithms on other various sample sizes with different initial values. Also, it was discussed that an artificial neural network can estimate the shape parameters but further research is needed to be done in order to perform regression analysis.

Acknowledgment

I would like to express my deepest appreciation to my supervisor, Dr. Keming Yu, for his guidance throughout this academic year and without his exceptional statistical knowledge, I would not have been able to complete this project. I would like to thank my family, close friends and the schools that I have attended to for their kind support and guidance during my hard times.

Contents

Chapter 1

Introduction

1.0.1 Heavy Tail Distribution

The distribution of a random variable Y with distribution function F is said to have a heavy right tail [18] if for $t > 0$

$$\lim_{Y\to\infty} e^{\lambda y}\mathbb{P}[Y > y] = \infty \quad \forall \lambda > 0. \tag{1.1}$$

When a distribution satisfies (1.1) the distribution tends to have very large values with many extreme values. The heavier the tail, the larger the probability that you'll get one or more very large values in a sample. Heavy-tailed distributions appear in many industries including insurance, Internet services, and natural disaster management.
Thus, being able to fit heavy-tailed distributions to empirical data and accurately calculate functionals of heavy-tailed distributions, such as the ruin distribution with heavy-tailed claim sizes, is of utmost importance. However, heavy-tailed distributions are in many cases intractable.[8] One example of a heavy-right tailed distribution is the Burr type XII distribution and we will discuss this distribution in the next section.

1.0.2 Introduction To a Burr type XII Distribution

The Burr type XII distribution was introduced in 1942 by I.W. Burr and has been widely applied in many areas such as business, engineering quality control, reliability and mineralogy as a failure model.[2] The distribution is one of twelve distributions based on the differential equation

$$\frac{dF(y)}{dy} = F(y)(1-F(y))g(y,F(y)) \tag{1.2}$$

where g(y,z) is positive for $0 \leq z \leq 1$ and y is in the domain of $F(y)$ [13] . The cumulative distribution function (CDF) of a Burr type XII $F_Y(y)$ distribution is defined as

$$F_Y(y) = 1 - \frac{1}{(1+y^\tau)^\lambda}; \quad y > 0, \tag{1.3}$$

where $\tau > 0$ and $\lambda > 0$.

Given that we know the cumulative distribution of a Burr type XII distribution we can derive its probability density function as

$$f(y) = \frac{\lambda\tau y^{\tau-1}}{(1+y^\tau)^{\lambda+1}}. \tag{1.4}$$

Let $Q(p)$ denote the quantile function

$$Q(p) = \inf\{y \in \mathbb{R} : p \leq F_Y(y)\}.$$

Then the quantile function of a Burr type XII distribution function is given by

$$Q(p) = \Big((1-p)^{-1/\alpha} - 1\Big)^{1/\lambda}; \quad 0 < p < 1. \tag{1.5}$$

It can be shown that $Q(p)$ decreases with respect to α for a given value of with $0 < p < 1$ and a value of λ or $\lambda > 0$; and for a given $0 < p < 1$, $Q(p)$ increases with respect to λ of

$$\alpha > -\frac{(1-p)}{ln(2)},$$

otherwise $Q(p)$ decreases with respect to λ[1] .

1.0.3 Method of Moments & Moment-generating Functions of a Burr Type XII distribution

Method of Moments

The nth central moment of random variable Y is given by

$$\mu_n = \mathbb{E}\big[Y - \mu\big]^n,$$

where μ is the first moment of random variable Y. It is worth mentioning that the second, third and the fourth central moments is the variance, skewness and kurtosis of random variable Y respectively.
The skewness and the kurtosis of a Burr type XII distribution is extremely important especially it can be used to estimate the shape parameters using a neural network model[12].

If the random variable Y follows a burr type XII distribution, BXII(τ, λ) then we can obtain the nth moment of Y by

$$\mathbb{E}\big(Y^n\big) = \int_0^\infty y^n f(y)dy = \tau\lambda \sum_{j=0}^{\infty} \frac{(-1)^j \Gamma\Big(\frac{n}{\tau} + 1\Big)}{n!\Gamma\Big(\frac{n}{\tau} + 1 - j\Big)\Big[\tau(\lambda + j) - n\Big]},$$

where Γ is the gamma function.

Moment-Generating Function

The moment generating function of a random variable Y is a function $M_Y : \mathbb{R} \to [0, \infty)$ given by

$$M_Y(t) = \mathbb{E}\big[e^{tY}\big],$$

provided that the expectation exists for t in the vicinity of zero [12]. It is worth mentioning that if the moment generating function of Y exists then

$$\mathbb{E}\left[Y^n\right] = M_Y^n(0).$$

If the random variable Y follows a burr type XII distribution, BXII(τ, λ), then the moment generating function of Y is given by

$$M_Y(t) = \mathbb{E}\left[e^{tY}\right] = \int_0^\infty e^{tY} f(y)dy = \tau\lambda \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{(-1)^j t^i \Gamma\left(\frac{i}{\tau}+1\right)}{i!\,j!\,\Gamma\left(\frac{i}{\tau}+1\right)\left[\tau(\lambda+j)-i\right]}.$$

1.0.4 Tail Index of a Burr Type XII Distribution

The parameters τ and λ are shape and scale parameters of the Burr type XII distribution respectively and each parameter has its own statistical meaning. The density function $f(y)$ is unimodal if $\tau > 1$ with mode at

$$y = \left(\frac{\tau-1}{\lambda\tau+1}\right)^{1/\tau}$$

and $f(y)$ is L-shaped if $\tau \leq 1$.

1.0.5 Important Distributions Derived From a Burr type XII Distribution

There are many specials distribution such as the Lomax, Weibull, Log-logistic and many more distributions that are useful in reliability and they are all a special case of a Burr type XII distribution and in this section, we will show how the Lomax, Weibull, and the Log-logistic distribution arise from a Burr type XII distribution.
Recalling back to the CDF of a Burr type XII distribution, the scale parameter can be represented in many different ways and if we rewrite the CDF given in (1.3) by introducing a scale parameter γ, then

$$F_Y(y) = 1 - \frac{1}{(1+\frac{y^\tau}{\gamma})^\lambda}. \tag{1.6}$$

$$F_Y(y) = 1 - \frac{1}{(1+(\frac{y}{\gamma})^\tau)^\lambda}. \tag{1.7}$$

Lomax Distribution

Lomax (1954) used the following cumulative distribution to fit some business failure data:

$$F(y) = 1 - (\kappa/(y+\kappa))^{-\lambda} = 1 - (1+y/\kappa)^{-\lambda}. \tag{1.8}$$

He derived this distribution based on the failure rate function

$$\pi(y) = \lambda/(y+\gamma),$$

where $\pi(y) = f(y)/(1-F(y))$[14]. It can be shown that (1.8) is a special case of a Burr type XII cumulative distribution function using (1.6) or (1.7), with scale parameter $\gamma = \kappa$

and the shape parameter $\tau = 1$[14]. Let Y be a random variable that follows an exponential distribution with probability density function

$$f(y|\beta) = \beta e^{-\beta y}, \quad y \geq 0, \beta > 0. \tag{1.9}$$

and if the parameter β has a gamma distribution with Probability distribution function

$$g(\beta) = \kappa^{\lambda} \beta^{\lambda-1} e^{-\kappa\beta}, \quad \beta \geq 0, \kappa > 0, \lambda > 0. \tag{1.10}$$

Then the probability density function of Y can be obtained as

$$f(y) = \int_0^{\infty} f(y|\beta) g(\beta) d\beta = \lambda/\gamma\big[1 + y/\gamma\big]^{-(\lambda+1)}. \tag{1.11}$$

This is the Lomax distribution and the cumulative distribution function corresponding to (1.11) is the same as (1.8) and we have shown that Lomax distribution is a special case of a Burr type XII distribution [14].

The Weibull Distribution

Let X follow an Weibull distribution, $X \sim \text{Weibull}(k, \delta)$, then its cumulative distribution function is given by

$$F(x) = 1 - e^{-(x/\delta)^k}, \quad x \geq 0. \tag{1.12}$$

Similarly, If Y follows an extended (three parameters) Burr type XII distribution, Y$\sim$ EBXII(τ, λ, γ), then its cumulative distribution function is given by

$$F(y) = 1 - \frac{1}{(1 + (\frac{y}{\gamma})^{\tau})^{\lambda}}. \tag{1.13}$$

Recall that

$$\lim_{z \to \infty} (1 + z/n)^{-n} = e^{-z}.$$

To get the cumulative distribution function of Y to approach the cumulative distribution function of X in a pointwise sense, then take the limit of $F_Y(y)$ as $\lambda \longrightarrow \infty$.
Let $\gamma = \lambda^{1/k}\delta$ If γ is allowed to approach infinity independently of λ, then of course the limit of $F_Y(y)$ is indeterminate [14]. If we set $Y \sim \text{EBXII}(\tau, \lambda, \lambda^{1/k}\delta)$ then we obtain

$$\begin{aligned} \lim_{\lambda \to \infty} F_Y(y) &= 1 - \lim_{\lambda \to \infty} \left[1 + (\frac{y}{\lambda^{1/k}\delta})^k\right]^{-\lambda} \\ &= 1 - \lim_{\lambda \to \infty} \left[1 + \frac{(y/\delta)^{k\,k}}{\lambda}\right]^{-\lambda} \\ &= 1 - e^{-(y/\delta)^k} \\ &= F_X(x). \end{aligned} \tag{1.14}$$

The Log-Logistic Distribution

Consider the logistic random variable X with cumulative distribution function given by

$$F(x) = 1 + \exp(-(x - \ln\kappa)/\beta)^{-1} \tag{1.15}$$

where β is a scale parameter and $\ln\kappa$ is chosen as the location for convenience[14]. Consider the random variable $Y = \exp(X)$, then its corresponding probability distribution function is given by

$$f(y) = \beta\kappa y^{\beta-1}/(\kappa^{\beta} + y^{\beta})^2, y \geq 0. \tag{1.16}$$

The CDF corresponding to (1.16) can be written as

$$\begin{aligned} F(y) &= y^{\beta}/(\kappa^{\beta} + y^{\beta}) \\ &= (y/\kappa)^{\beta}/(1 + (y/\kappa)^{\beta}). \end{aligned} \tag{1.17}$$

It should be noted that (1.17) is a special case of (1.7) when $\tau = 1$ and $\lambda = \beta$.

1.0.6 Literature Review

In Beirlant's research paper (1998) he focuses on a more generalised Burr type XII distribution with density

$$f(y, \lambda, \tau, \beta) = \frac{\lambda\beta^{\lambda\tau}y^{\tau-1}}{(\beta + y^{\tau})^{\lambda+1}}, \quad \lambda > 0, \tau > 0, \beta > 0, y > 0. \tag{1.18}$$

Beirlant mentions that (1.18) can be extended to a regression model. He considered maximum likelihood estimation of the regression and nuisance parameters for two different parametrisations. In parametrisation I the shape parameter τ is allowed to vary with covariate $\boldsymbol{x}$ (k-dimension vector), so

$$Y = y|X = x \sim \text{EBXII}(\beta, \lambda, \tau(\boldsymbol{x})).$$

There are many functions that are possible for $\tau(\boldsymbol{x})$. Since $\tau > 0$, it is appropriate to use the exponential link function, $\tau(\boldsymbol{x}) = \exp(\boldsymbol{\theta}^T\boldsymbol{x})$ with $\boldsymbol{\theta}$ denoting the k-dimensional vector of regression coefficients[1].

A second parametrisation can be derived from the relation between the Burr Type XII and the generalized logistic distributions. Indeed specifying $\beta(\boldsymbol{x}) = \exp(\tau\boldsymbol{\theta}^T\boldsymbol{x})$ leads to a linear representation for the location parameter $\delta = \boldsymbol{\theta}^T\boldsymbol{x}$ and in case of generalised logistic random variables Z defined by $Z = \ln(Y) + \delta$, with Y distributed as in (1.4). Note that by a simple re-parametrisation τ can be included in θ, so that τ would disappear from (1.6)[1].

A simulation study was done under parametrization I, where 2000 data sets of sizes $n = 200, 500, 1000$ and 2000 were generated from a Burr regression model with $\lambda = 0.5$, $\beta = 5$, $\boldsymbol{\theta}^T = (-1, 2)$, $\boldsymbol{x}^T = (1, x)$. The explanatory variable $\boldsymbol{x}$ is continuous with values generated from a uniform distribution, $U[0, 1]$ [1].

For the location-scale model given by parametrization II a similar simulation experiment was performed. Again 2000 datasets of sizes of $n = 200, 500, 1000, 2000$ were generated with parameters $\lambda = 2, \tau = 1.5, \boldsymbol{\theta}^T = (-1, 2)$ [5].

From Beirlant's simulation study, the results for parametrisation I showed strong convergence for all regression coefficients and the shape parameters for all sample sizes. The

sample mean for each parameter is approximately equal to its true value and it is observed that as the sample size increases the accuracy gets better for each parameters. Also, each parameters sample variance is relatively small. However, from the residual plots (Appendix), the distributions of the regression parameters are approximately normally distributed and this indicated that the covariates are a good fit for the Burr regression model. This is especially true for small sample sizes but in the case of the scale parameters τ and λ larger samples are needed for the normal approximation to be appropriate.

The results from parametrisation II showed strong convergence rate for all parameters in all sample sizes. The sample variances for all parameters were approaching zero as the sample sizes were increasing. The residual plots showed similar results that were found under parametrisation I. From this simulation study it is observed that for small sample sizes the maximum likelihood estimation is biased and has a minimal unbiased variance for large samples. Therefore, the method is very useful in estimating the parameters and it is worth knowing how well each parametrisation work in different sample sizes.

In another paper it was discussed the importance of a Burr type XII distribution in reliability analysis and the paper aimed to study the reliability in a multicomponent stress-strength based on X, Y be two independent random variables, where $X \sim \text{BXII}(\gamma_1, \beta)$ and $Y \sim \text{BXII}(\gamma_2, \beta)$. It was assumed for a system with k identical components, functions if at least s $(1 \leq s \leq k)$ components simultaneously operate. In its operating environment, the system is subjected to a stress Y which is a random variable with distribution function G(.)[5]. The strengths of the components, that is the minimum stresses to cause failure, are independently and identically distributed random variables with distribution function F(.)[5]. The reliability of the system was obtained by the equation

$$R_{s,k} = \sum_{i=s}^{k} \binom{k}{i} \int_{-\infty}^{\infty} [1 - F(y)]^i [F(y)]^{k-i} dG(y). \tag{1.19}$$

Two real dataset that were studied by Zimmer et al. and Lie et al. were

X: 0.19, 0.78, 0.96, 0.31, 2.78, 3.16, 4.15, 4.67, 4.85, 6.50, 7.35, 8.01, 8.27, 12.06, 31.75, 32.52, 33.91, 36.71 ,72.89

and

Y: 0.9, 1.5, 2.3, 3.2, 3.9, 5.0, 6.2, 7.5, 8.3, 10.4, 11.1, 12.6, 15.0, 16.3, 19.3, 22.6, 24.8, 31.5, 38.1 ,53.0 [5].

Rao and his collaborators implemented the iterative procedure to obtain the maximum likelihood estimation of γ_1, γ_2 and β. The final estimates for the real data sets were $\gamma_1 = 0.287835$, $\gamma_2 = 0.232784$ and $\hat{\beta} = 1.799809$. Based on the estimates of γ_1, γ_2 and the maximum likelihood estimate of $R_{1,3}$ =0.689914 and $R_{2,4}$ = 0.533462. The 95% confidence intervals for $R_{1,3}$ and $R_{2,4}$ is (0.505449, 0.874379) and (0.337642, 0.729282) respectively. From the simulation study, the maximum likelihood estimation of the parameters was very good and the simulation study results have shown that the average bias and the average MSE decreases as the sample size increases for the iterative procedure. Among the parameters the absolute bias and MSE

increases (decreases) as γ_1 increases (γ_2 increases) in both the cases of (s, k)[5].

1.0.7 Aims and Objective

This paper aims to study a Burr type XII distribution but primarily we are interested when the covariate information $\boldsymbol{x}$ is given and Beirlant and his collaborators (1998, 1999, 2004) considered a regression model via one shape parameter λ:

$$\lambda = \lambda(\boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x})$$

This paper introduces three algorithms that estimates the regression coefficients $\boldsymbol{\beta}$'s. The aim of this paper is to compare the accuracy of these three methods. The three methods are: maximum likelihood estimations; ordinary least square estimation with one of the shape parameters (τ) being held constant ; ordinary least squares estimation with τ being unknown. To assess the accuracy of the three methods we will consider the simple case where $\lambda = \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x)$ where the actual value (initial) of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are both set to 1 and where it is necessary we will estimate τ, where its actual value (initial) is set to 0.5. In addition, each methods will be implemented on four samples sizes 25, 250, 750 and 1000. For each methods there are various numerical methods that can be used to estimate the parameters and therefore where it is necessary the iterative procedure and Newton's method (algorithm will be implemented on several random dataset) for optimisation will be implemented.
From Rao et al. research paper it was stated that the iterative procedure had strong convergence in estimating the parameters in their simulation studies and therefore in our project we will determine whether this is true for our simulation study. Furthermore, this paper will show how the ordinary least squares estimation with τ held constant can overcome big data problem. Therefore, we will introduce and implement the split and combination algorithm[5] and we use bootstrapping method to compute the confidence intervals for the parameters.

Chapter 2

Numerical Methods

2.1 Estimation method 1: Maximum Likelihood Estimate

2.1.1 Maximum Likelihood Estimation

Suppose that the random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a sample from a distribution $f(x|\theta)$; if x is continuous then $f(x|\theta)$ is known as the probability density function (PDF). For every observed $(x_1, \ldots, x_n)$, we define

$$f(X_1 = x_1, \ldots, X_n = x_n|\theta) = f(x_1|\theta) \ldots f(x_n|\theta). \tag{2.1}$$

Now we call $f(X_1 = x_1, \ldots, X_n = x_n|\theta)$ as the likelihood function. As you can see, the likelihood function depends only on the unknown parameter θ, and it is always denoted as $L(\theta)$. The meaning of the likelihood function is as follows; We choose a parameter that makes the likelihood function of having the obtained data at hand maximum.
Theoretically, if we had no actual data, maximising the likelihood function will give us a function of n random variables $X_1, \ldots, X_n$, and we shall call this as the maximum likelihood estimation of $\hat{\theta}$.
The maximum likelihood estimation requires us to maximise the likelihood function ,$L(\theta)$ with respect to the unknown parameter θ. However, in equation (2.1), $L(\theta)$ is defined as a product of n terms, which is not easy to maximise. Maximising $L(\theta)$ is equivalent to maximising $\ln L(\theta)$ because logarithms are monotonic increasing functions and if we take the logarithm of the likelihood function then

$$\ell(\theta) = \ln L(\theta) = \ln \prod_{i=1}^{n} f(X_i = x_i|\theta) = \sum_{i=1}^{n} \ln\Big(f(X_i = x_i|\theta)\Big) \tag{2.2}$$

2.1.2 Properties of Maximum Likelihood Function: Consistency

Let us recall one of the most important theorem in probability and statistics namely the law of large numbers.

The Law of Large Numbers

Let $X_1, \ldots, X_n$ be independent and identical random variables, each with mean $\mathbb{E}[X_i] = \mu$ and standard deviation σ, we define $\bar{X_n} = \frac{x_1 + \ldots + x_n}{n}$ [9]. The weak Law of Large Numbers states that for all $\varepsilon > 0$ then

$$\lim_{n \to \infty} P(|\bar{X_n} - \mu| > \varepsilon) = 0. \tag{2.3}$$

We will show that the maximum likelihood estimation satisfies the consistency property. We say that $\hat{\theta}$ is consistent if $\hat{\theta} \longrightarrow \theta_0$ in probability as n $\longrightarrow \infty$, where θ_0 is the true value of $\hat{\theta}$.

Consistency of Maximum Likelihood Estimation

The maximum likelihood estimation of $\hat{\theta}$ is the maximiser of

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln\Big(f(\boldsymbol{X}_i|\theta)\Big) \tag{2.4}$$

This is simply the log-likelihood function normalised by $\frac{1}{n}$ and if we consider the function $\ell(X|\theta) = \ln f(X|\theta)$ and define

$$\ell(\theta) = \mathbb{E}_{\theta_0} l(X|\theta), \tag{2.5}$$

where $\mathbb{E}_{\theta_0}$ denotes the expectation with respect to the true unknown parameter θ_0 of the sample $X_1, \ldots, X_n$[4].

$$\mathbb{P}(\sqrt{n}(\bar{X}_n \in [a,b]) \longrightarrow \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx. \tag{2.6}$$

Assuming a continuous distribution then

$$\ell(\theta) = \int \Big[\ln f\big(x|\theta)\big) f(x_0|\theta)\Big] dx. \tag{2.7}$$

By the law of large numbers, for any θ,

$$\ell_n(\theta) \longrightarrow \mathbb{E}_{\theta_0} l(\boldsymbol{X}|\theta) = \ell(\theta). \tag{2.8}$$

Using the lemma that states $\ell(\theta) \leq \ell(\theta_0)$ we can say that under some regularity conditions on the family of the distributions, the maximum likelihood estimation of $\hat{\theta}$ is consistent, that is $\hat{\theta} \longrightarrow \theta_0$ as n $\longrightarrow \infty$[4].

2.1.3 MLE:Estimating the tail index of Burr type XII Distribution

Recall that the density function of Burr XII distribution is given by

$$f(y) = \frac{\lambda \tau y^{\tau-1}}{(1+y^{\tau})^{\lambda+1}}$$

then the likelihood function is defined as,

$$L(\tau, \lambda|y) = \prod_{i=1}^{n} f(y) = \prod_{i=1}^{n} \frac{\lambda \tau y_i^{\tau-1}}{(1+y_i^{\tau})^{\lambda+1}} = \lambda^n \tau^n \frac{\prod_{i=1}^{n} y_i^{\tau-1}}{\prod_{i=1}^{n}(1+y_i^{\tau})^{\lambda+1}}. \tag{2.9}$$

Taking the natural logarithm of (2.9) becomes

$$\ell = n\ln(\lambda) + n\ln(\tau) + (\tau-1)\sum_{i=1}^{n} \ln(y_i) - (\lambda+1)\sum_{i=1}^{n} \ln\Big(1+y_i^{\tau}\Big) \tag{2.10}$$

Now if we take the partial derivative of ℓ with respect to λ and τ we get

$$\begin{aligned}\frac{\partial \ell}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{i=1}^{n} \ln(1+y_i^{\tau}) = 0 \\ \frac{\partial \ell}{\partial \tau} &= \frac{n}{\tau} + \sum_{i=1}^{n} \ln(y_i) - (\lambda+1)\sum_{i=1}^{n} \frac{\ln(y_i) y_i^{\tau}}{1+y_i^{\tau}} = 0\end{aligned} \tag{2.11}$$

There are two known numerical methods that can solve (2.11). The first method is the iterative procedure and the second is Newton's method for optimisation. The iterative procedure is computational expensive but on the positive side it was stated by Rao that the estimation for the parameters gets better as the number of iteration increases[**6**].

2.1.4 Estimating the tail Index of a Burr type XII Regression

Let Y denote the dependent variable that is linearly related to k independent variables $X_1, X_2, \ldots, X_k$ through the parameters $\beta_1, \beta_2, \ldots, \beta_k$ and we write

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon. \tag{2.12}$$

This is called the multiple linear regression model. The parameters $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients associated with covariate $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k$ respectively and β_0 is known as the intercept and ε is the random error component reflecting the difference between the observed and fitted linear relationship. We can rewrite (2.12) in matrix notation

$$\boldsymbol{Y} = \boldsymbol{\beta}^T \boldsymbol{X} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2). \tag{2.13}$$

Once covariate information $\boldsymbol{x}$ is available, Beirlant and his collaborators considered a regression model via one shape parameter

$$\lambda \equiv \lambda(\boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}) \tag{2.14}$$

This transformation in (2.14) is not unique but we have chosen the exponential function since $\lambda > 0$. There are other transformations that we can choose from and in general we can transform both shape parameters as long as they satisfy their inequality[**jan**]. Let $Y_1, Y_2, \ldots, Y_k$ be independent random variables with

$$Y_i = y_i | X_i = x_i \sim \text{BXII}(\tau, \lambda_i), \quad \lambda_i = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i).$$

The likelihood function for a sample of n independent observations, $y_1, y_2, \ldots, y_n$, is then given by

$$L = \tau^n \prod_{i=1}^{n} \frac{y_i^{\tau-1} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\left[1+y_i^{\tau}\right]^{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)+1}}$$

and

$$\ell = n\ln(\tau) + \sum_{i=1}^{n} \boldsymbol{\beta}^T \boldsymbol{x}_i + (\tau-1)\sum_{i=1}^{n} \ln(y_i) - \sum_{i=1}^{n} \left[\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) + 1\right] \ln(1+y_i^{\tau}).$$

The partial derivatives of the log-likelihood function with respect to all parameters is given by:

$$\begin{aligned}\frac{\partial \ell}{\partial \tau} &= \frac{n}{\tau} + \sum_{i=1}^{n} \ln(y_i) - \sum_{i=1}^{n} \Big[\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) + 1\Big] \frac{y_i^{\tau} \ln(y_i)}{1 + y_i^{\tau}}. \\ \frac{\partial \ell}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n} \boldsymbol{x}_{ij} - \sum_{i=1}^{n} \Big[\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) \boldsymbol{x}_{ij} \ln(1 + y_i^{\tau})\Big], \quad j = 0, 1, \ldots, j = k - 1.\end{aligned} \tag{2.15}$$

Set

$$\boldsymbol{\beta}^{*T} = (\tau, \boldsymbol{\beta}^T) = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \ldots, \boldsymbol{\beta}_{p+2}^*).$$

Solving

$$\frac{\partial \ln L(\boldsymbol{\beta}^* | y)}{\partial \boldsymbol{\beta}_j^*} = 0, j = 1, 2, \ldots, j = k + 2,$$

subject to the restriction $\boldsymbol{\beta}_1^* > 0, \boldsymbol{\beta}_2^* > 0$ leads to the maximum likelihood estimate $\hat{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}^*$[2].

2.2 Method 2: Ordinary Least Squares

In section 1 we have shown that λ, τ and the regression coefficients $\boldsymbol{\beta}$ can be estimated using the maximum likelihood estimation method and in this section we will show how the ordinary least square estimation with τ held constant and ordinary least square estimation with τ is unknown estimates the regression coefficients.

2.2.1 Linear Regression

Suppose, we have n observations on $k + 1$ independent variables. Then a multiple linear regression models takes the form

$$y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{i1} + \ldots + \boldsymbol{\beta}_k x_{ik} + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n. \tag{2.16}$$

We can write a linear regression model in matrix.

Let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{d2} & x_{k3} & \ldots & x_{kn} \end{bmatrix} \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_k \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix}, \tag{2.17}$$

Then the regression model can be written in the form

$$\mathbb{E}[Y = y | X = x] = \boldsymbol{\beta}^T \boldsymbol{X}. \tag{2.18}$$

Given data we can compute β using the ordinary least squares estimation

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \tag{2.19}$$

provided that $\boldsymbol{X}$ is invertible.

2.2.2 Online analytical processing (OLAP) with regression for massive data

When the sample size is too large or consists of different stages of data from different time period and location, one will face challenge in data storage. We can tackle this problem by implementing the divide and conquer algorithm. The idea behind divide and conquer is to split data into small blocks and compute the coefficients for all blocks and combine all the results so that the solution of the combination computes the coefficient of the original problem.

Recall the linear regression:

$$\mathbb{E}[Y = y|X = x] = \boldsymbol{\beta}^T\boldsymbol{X} \tag{2.20}$$

Assume that we are able to partition the raw design matrix X in K-sub-matrix. Suppose that the entire data set comes from an aggregative data set or K-cells. All K cells have more or less similar patterns.
Let Y_k and X_k be the collections of values of the response and independent variables respectively in the k^{th} cell $(k = 1,2,\ldots,K)$[6]. That is, $\boldsymbol{Y}_k$ and $\boldsymbol{X}_k$ are a sub-vector and sub-matrix of raw observations $\boldsymbol{Y}$ and $\boldsymbol{X}$ respectively and satisfy

$$\boldsymbol{X}^T\boldsymbol{X} = \sum_{i=1}^{K}\boldsymbol{X}_k^T\boldsymbol{X}_k \text{ and } \boldsymbol{X}^T\boldsymbol{Y} = \sum_{i=1}^{n}\boldsymbol{X}_k^T\boldsymbol{Y}_k. \tag{2.21}$$

Let $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ be the estimator of $\boldsymbol{\beta}$ from the aggregate data[6]. Using (2.21) we have

$$\hat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{K}\boldsymbol{X}_k^T\boldsymbol{X}_k\Big)^{-1}\Big(\sum_{i=1}^{K}\boldsymbol{X}_k^T\boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k\Big). \tag{2.22}$$

2.2.3 Ordinary Least Square Estimation with τ Held Constant

Recall that the cumulative distribution function of a Burr type XII distribution is given by

$$F_Y(y) = 1 - \frac{1}{(1+y^\tau)^\lambda}$$

Let $0 < u < 1$ then,

$$\begin{aligned}\mathbb{P}[F_Y(y) < u] &= \mathbb{P}[Y < F_Y^{-1}(u)] \\ &= F[F_Y^{-1}(u)] \\ &= u\end{aligned} \tag{2.23}$$

From this we have $F_Y(y)$ which follows an uniform distribution $F_Y(y) \sim U[0,1]$. As a result of this the complement of a Burr type XII cumulative distribution function also follows an

uniform distribution, $1 - F_Y(y) \sim U[0,1]$. The cumulative distribution of a Burr type XII distribution, $F_Y(y)$ can be rearranged to

$$F_Y(y) = -\ln\Big(1 - F_Y(y)\Big) \sim U[0,1].$$

Let Y be a continuous random variable with CDF $F_Y(y)$ and domain R_Y, and let $U = g(Y)$ where $g : R_Y \longrightarrow \mathbb{R}$ is continuous , on-to-one function defined over R_Y. Recall that if g is one-to-one function, it has an unique inverse g^{-1} and recall that if g is increasing (decreasing), then so is g^{-1}.
Suppose that $g(y)$ is a strictly increasing function of y defined over R_y. Then, it follows that

$$u = g(y) \Longleftrightarrow g^{-1}(u) = y \tag{2.24}$$

and

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ \mathbb{P}[g(Y) \leq u] &= \mathbb{P}[g(Y) \leq u] \\ \mathbb{P}[Y \leq g^{-1}(u)] &= F_Y[g^{-1}(u)]. \end{aligned} \tag{2.25}$$

Differentiating $F_U(u)$ with respect to u, we get

$$\begin{aligned} f_U(u) &= \frac{d}{du} F_U(u) \\ &= f_U(g^{-1}(u)) \frac{d}{du} g^{-1}(u). \end{aligned} \tag{2.26}$$

As g is increasing, so is g^{-1}. Thus, $\frac{d}{du} g^{-1}(u) > 0$.[7]
If $g(y)$ is strictly decreasing,

then

$$f_U(u) = f_Y\Big(g^{-1}(u)\Big)\Big|\frac{d}{du} g^{-1}(u)\Big|.$$

We can use this method and show that

$$-\ln(Y) \sim \exp(1),$$

provided that Y is uniformly distributed.

In addition, we can show that

$$\ln[1 - F_Y(y)] \equiv \lambda \ln[(1 + y^{\tau}], \tag{2.27}$$

follows an exponential distribution , exp(1)[7].

Recall that if $X \sim \exp(\lambda)$ then $cX \sim \exp(\lambda/c)$

and

$$\exp(\lambda) = \frac{1}{2\lambda}\exp(\frac{1}{2}) \sim \frac{1}{2\lambda}\chi_2^2.$$

Let $S = \ln(1+y^{\tau})$ and if we use the property above then $2\lambda S$ follows a chi-squared distribution with 2 degrees of freedom.

If $X \sim \chi^2(v)$ and $c > 0$, then cX follows a gamma distribution, $cX \sim \Gamma(k = \frac{v}{2}, \theta = 2c)$. Then, we have $2\lambda S \sim \Gamma(1,2)$[16].

If $X \sim \Gamma(k,\theta)$.

Then

$$\mathbb{E}\Big[\ln(X)\Big] = \Phi(k) + \ln(\theta)$$

where,

$$\Phi(k) = \frac{d}{dx}\Big[\ln(\Gamma(x))\Big],$$

is the Digammma function[17] .

Using the Digamma function we can show that

$$\begin{aligned} \mathbb{E}\Big[(2\lambda S)\Big] &= \mathbb{E}\Big[\ln(2\lambda S)\Big] \\ \ln(2) + \ln(\lambda) + \mathbb{E}\Big[\ln(S)\Big] &= \Phi(1) + \ln(2) \\ \mathbb{E}\Big[-\ln(S) + \Phi(1)\Big] &= \ln(\lambda). \end{aligned} \tag{2.28}$$

Let $\boldsymbol{U} = -\ln(S) + \Phi(1)$ then the equation in (2.27) becomes

$$\mathbb{E}\Big[U\Big] = \ln\Big[\ln(1+y^{\tau})\Big] - 0.5772 = \boldsymbol{\beta}^T \boldsymbol{X}. \tag{2.29}$$

This is simply an ordinary linear regression where $\mathbb{E}\Big[\boldsymbol{U}\Big] = \boldsymbol{\beta}^T \boldsymbol{X}$ and $\text{Var}[\boldsymbol{U}] = \frac{\pi^2}{6}$. Given data $\boldsymbol{X}$, we can compute the regression coefficients $\boldsymbol{\beta}$ using the ordinary least square methods provided that τ is held constant using the formula

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}\boldsymbol{U}. \tag{2.30}$$

2.2.4 Ordinary Linear Regression Estimate with τ being Unknown

We can estimate the parameters in (2.29) even if τ is not known.

Recall that for each $2\lambda_i S_i \sim \chi^2(2)$. Let X_i be a chi-squared distribution and for $i = 1,2,\ldots,n$ then we have

$$Z = \sum_{i=1}^{n} X_i \sim \chi^2(2n).$$

Using this information we can show that for $i = 1, 2, \ldots, n$

$$2\sum_{i=1}^{n} \lambda_i S_i \sim \chi^2_{2n}. \tag{2.31}$$

This is equivalent to

$$2\sum_{i=1}^{n} \exp(\boldsymbol{\beta}^T X_i)\ln(1+x_i^{\tau}) \sim \chi^2_{2n}. \tag{2.32}$$

Given the observations $\{X_i, Y_i\}_{i=1}^{n}$ of (X, Y), we have

$$2\sum_{i=1}^{n} \exp(\boldsymbol{\beta}^T X_i)\ln(1+y_i^{\tau}) = 2(n-1) \tag{2.33}$$

$$\begin{aligned} 2\sum_{i=1}^{n} \exp(\boldsymbol{\beta}^T X_i)\ln(1+y_i^{\tau}) &= 2(n-1) \\ &\iff \frac{1}{n-1}\sum_{i=1}^{n} \exp(\boldsymbol{\beta}^T X_i)\ln(1+y_i^{\tau}) \\ &= 1. \end{aligned} \tag{2.34}$$

The equation in (2.34) has an unique solution of τ, saying $\hat{\tau}$ and the $\hat{\tau}$ is almost sure convergence or strong consistency.

Chapter 3

Implementations

3.0.1 Estimation of τ and λ with Maximum Likelihood Estimation

In this section, we will implement the maximum likelihood estimation method to find the shape parameters τ and λ. The estimation will be done in two different ways namely by the iteration procedure and Newton's method.

Iterative Procedure Method

To estimate the shape parameters iteratively we need come up with an algorithm. If we manipulate both equations in (2.11) we get the following algorithm

$$\begin{aligned} \text{Initial: } \hat{\lambda}^{(1)} &= \alpha > 0, \hat{\tau}^{(1)} = \beta > 0 \\ \text{For j} &= 1, \ldots, k \\ \hat{\lambda}^{(j+1)} &= \frac{n}{\sum_{i=1}^{n} \ln(1+y_i^{\hat{\tau}^{(j)}})} \\ \hat{\tau}^{(j+1)} &= \frac{n}{\frac{(\hat{\lambda}^{(j)}+1)\sum_{i=1}^{n} \ln(y_i) y_i^{\hat{\tau}^{(j)}}}{1+y_i^{\hat{\tau}^{(j)}}} - \sum_{i=1}^{n} \ln(y_i)} \end{aligned} \tag{3.1}$$

$$\begin{aligned} \lambda^* &= (\hat{\lambda}^{(1)}, \hat{\lambda}^{(2)}, \ldots, \hat{\lambda}^{(k)}) \\ \tau^* &= (\hat{\tau}^{(1)}, \hat{\tau}^{(2)}, \ldots, \hat{\tau}^{(k)}). \end{aligned}$$

Having implemented the algorithm, 25 estimation of the parameters were made in sample sizes of length 25, 250 ,750 and 1000 with initial value $\hat{\tau}^{(1)} = 0.5$ and $\hat{\lambda}^{(1)} = 1$. The results are summarised in the table (3.1).
The estimation for $\tau = 0.5$ in sample size 25, the result showed a strong convergence rate with little variation. Although, the sample mean is slightly above its actual value. When the number of data points increases we observed that the maximum likelihood estimation is an excellent estimation of the actual value since in sample size group 250 and 750 had zero variance. When the sample size increases too much the estimation becomes poor and we observed this in sample size 1000. The group mean was $(\bar{\tau} \approx 0.79)$, which clearly diverges away from the actual value.

The estimation of $\lambda = 1$ in sample size 25 is a good estimation to the actual value but strong convergence is observed in sample size 250 and 750. Their sample means are $\bar{\lambda} \approx 0.96$ and $\bar{\lambda} \approx 0.92$. Both sample group almost have zero variance. However, in sample size 1000 the

estimation diverged away from the actual value significantly and it even had a large variance. Thus, the maximum likelihood estimation of $\tau = 0.5$ and $\lambda = 1$ is approximately accurate in general but the best estimation was only observed in sample size 250 and 750. However, in a sample size of 1000, the maximum likelihood estimation of $\tau = 0.5$ and $\lambda = 1$ was the poorest since the estimation diverged from the actual value.

TABLE 3.1: Maximum Likelihood Estimate of τ and λ Iterative Method

Shape Parameter	**Sample Size**	**Sample Mean**	**Sample Variance**
$\tau = 0.5$	25	0.5867324	0.0003335
	250	0.5093234	3.8782e-06
	750	0.5010097	4.49273e-8
	1000	0.7905523	0.0037524
$\lambda = 1$	25	0.8896705	0.00052
	250	0.9667026	7.2396e-06
	750	0.928224	8.561215e-08
	1000	2.51322	0.05314993

The convergence and the volatility can be observed for the maximum likelihood estimation for $\tau = 0.5$ in figure 3.1. In sample size 25, 250 and 750 the volatility was the highest for the first 5 iterations. Although the maximum likelihood estimation of $\tau = 0.5$ is approximately equal to the actual value for these sample sizes and after the 6th iteration the estimation converges to a value in the vicinity of $\tau = 0.5$. Furthermore, for sample size 1000 we observed that the volatility is non-constant and exponentially decreasing to a value. Based on the nature of an exponential function it can be said that as the number of iteration increases the estimation has the tendency to approach the actual value.

The convergence and the volatility for the maximum likelihood estimation of $\lambda = 1$ are identical for sample sizes 25, 250, 750 but the difference is that for the first 5 iterations the variance is decreasing significantly and then the estimation converges to a value that is close to the actual value. In sample size 1000 the variation is extremely high for the first 5 iterations and then it converges to a value slowly in the vicinity of 1 (exponentially nondecreasing). Overall, it has been observed that the variance is relatively small for $\tau = 0.5$ and $\lambda = 1$ in all sample sizes but in sample size 1000 the variation can be explained by an exponential distribution and hypothetically it can be said that as the number of iteration increases the estimation converges to the actual value slowly. Thus, the maximum likelihood estimation of $\tau = 0.5$ and $\lambda = 1$ through the iterative method has shown strong convergence.

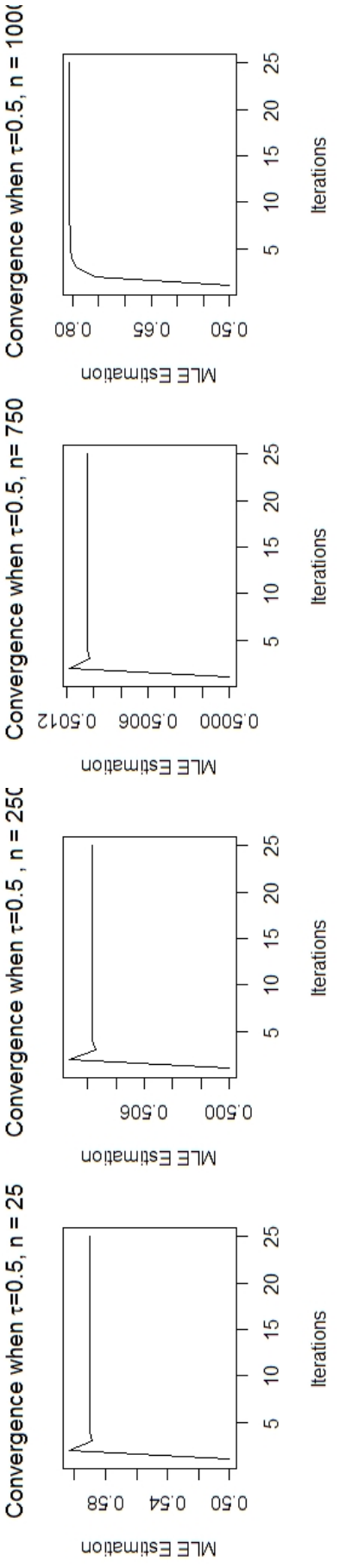

FIGURE 3.1: Maximum Likelihood Estimation of $\tau = 0.5$

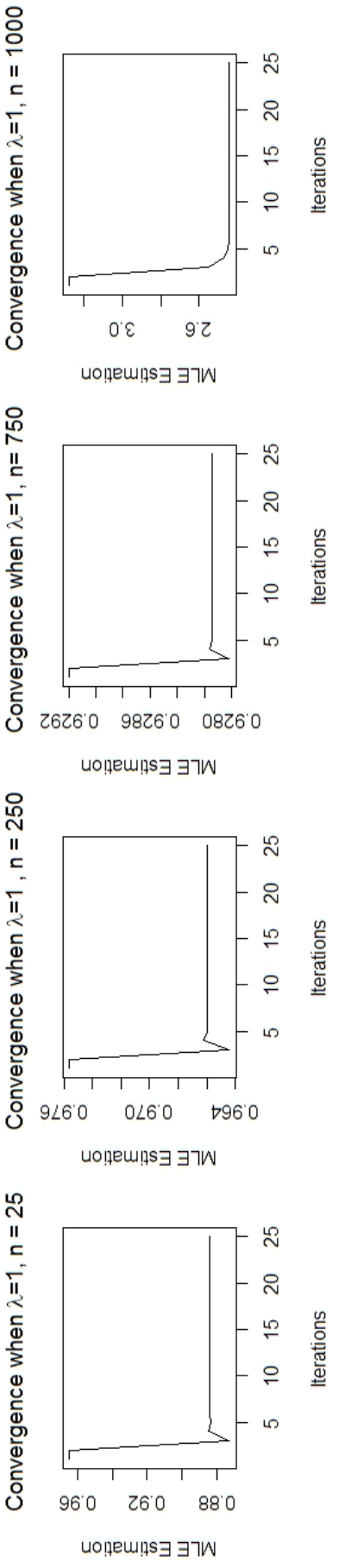

FIGURE 3.2: Maximum Likelihood Estimation of $\lambda = 1$

3.0.2 Maximum Likelihood Estimation of τ and λ through Newton's Method

Newton's method was used to solve (2.11) directly to the find the maximum likelihood estimation of τ and λ given that we know its actual value [same as 3.01]. In the previous analysis, 25 estimations of the parameters were found, whereas in this analysis Newton's method was implemented 10 times (trials) in each sample sizes. Each trial is independent and was implemented on different datasets and thus the variation between the trials came out to be high. Therefore, the average of the maximum likelihood estimation of the parameters in each sample size was considered.

From table 3.2, it is observed that the average maximum likelihood estimation of τ is an excellent measure overall across all sample groups with approximately zero variance. The estimation is only closest to the actual value in sample size 250 and 750 and the estimation was least closest to the actual value in sample size 1000 and this result was observed in the results of the previous method.
The average maximum likelihood estimation of λ was a good measure since the sample means was approximately around the actual value and the group's variance is roughly zero. The estimation was only closest to the actual value in sample size 250 and 1000 and the least in sample size 25. This result is different from the result found in the previous section. Since the estimation for sample size 25, 250 and 750 was good in the previous analysis whereas the estimation for this was only poor in a sample size of 1000.

TABLE 3.2: Maximum Likelihood Estimation of τ and λ : Newtons Method

Shape Parameter	**Sample Size**	**Sample Mean**	**Sample Variance**
$\tau = 0.5$	25	0.52064575	0.01102014
	250	0.51262603	0.001021143
	750	0.50464759	0.000359249
	1000	0.49447207	0.000223454
$\lambda = 1$	25	1.127257844	0.078927827
	250	0.98300694	0.001449356
	750	1.02635352	0.002039668
	1000	1.01398016	0.00190146

From figure 3.3 and 3.4 the maximum likelihood estimation of τ and λ is good overall but interestingly in sample size 1000, the estimation is actually closer to the actual value. Whereas the result from the iterative method, we observed that the estimation for both parameters in sample size 1000 diverged away from its actual value. Thus, based on the results we found, it can be said that the maximum likelihood estimation of τ and λ via Newton's method is more accurate and efficient on average than the iterative method for all sample sizes although the iterative method did converge to the parameters actual faster.

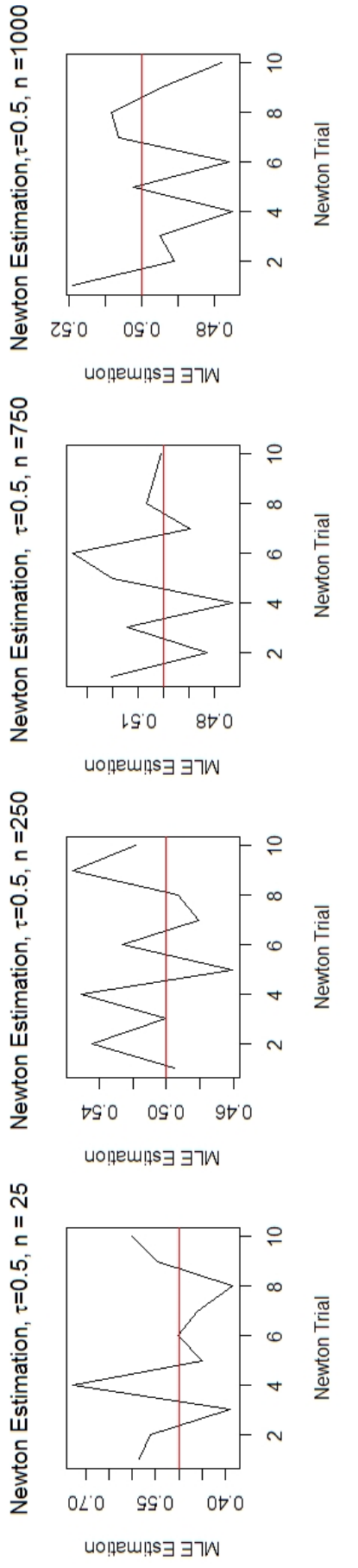

FIGURE 3.3: Maximum Likelihood Estimation of $\tau = 0.5$

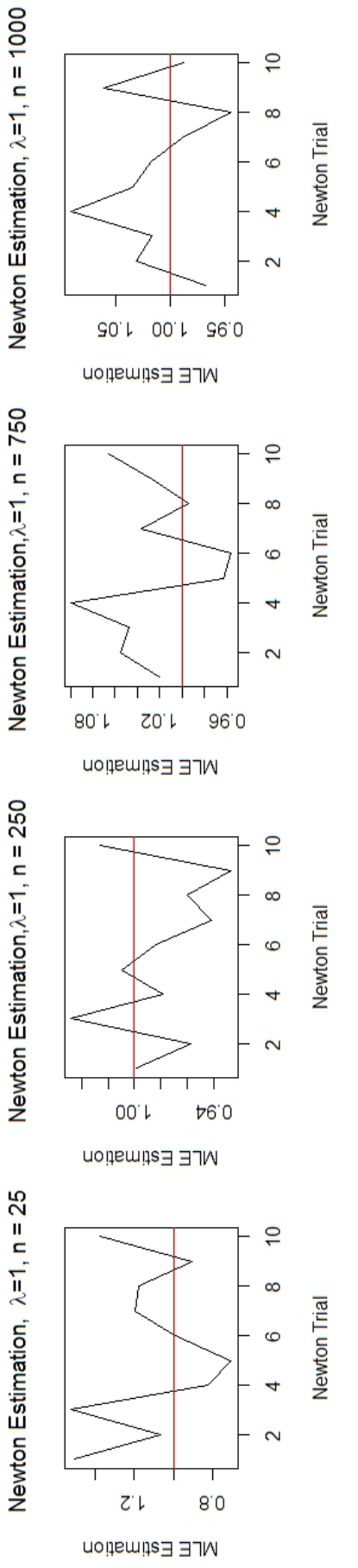

FIGURE 3.4: Maximum Likelihood Estimation of $\lambda = 1$

3.0.3 Maximum Likelihood Estimation with Data

If a dataset is given (positive points), a linear regression can be fit through a Burr type XII distribution via the transformation

$$\lambda \equiv \lambda(\boldsymbol{x}) = \exp(\beta_0 + \beta_1 x). \tag{3.2}$$

Let ℓ denote the log likelihood function

$$\ell = n\ln\tau + \sum_{i=1}^{n} + (\tau - 1)\sum_{i=1}^{n} \ln(y_i) - \sum_{i=1}^{n} \Big[(e^{\beta_0+\beta_1 x_i} + 1)\Big] \ln(1 + y_i^{\tau}) \tag{3.3}$$

Then if we take the partial derivative of the log-likelihood function with respect to τ and β_0 and solve τ and β_0 we get

$$\frac{\partial \ell}{\partial \tau} = \frac{n}{\tau} + \sum_{i=1}^{n} \ln(y_i) - \sum_{i=1}^{n} \Big[(1 + \exp(\beta_0 + \beta_1 x_i))\frac{y_i^{\tau}\ln(y_i)}{1 + y_i^{\tau}}\Big] = 0 \tag{3.4}$$

$$\tau = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{(e^{\beta_0+\beta_1 x_i}+1)y_i^{\tau}\ln(y_i)}{1+y_i^{\tau}}\Big]}$$

$$\frac{\partial \ell}{\partial \beta_0} = 0 \Longleftrightarrow \beta_0 = -\ln\Big(\frac{1}{n}\sum_{i=1}^{n} e^{\beta_1 x_i}\ln(1 + y_i^{\tau})\Big) \tag{3.5}$$

The partial derivative of ℓ with respect to β_1 and expressed in terms of β_0 is given by

$$\frac{\partial \ell}{\partial \beta_1} = 0 \Longleftrightarrow \beta_0 = \ln\Big(\frac{\frac{1}{n}\sum_{i=1}^{n} x_i}{\frac{1}{n}\sum_{i=1}^{n} e^{\beta_1 x_i}\ln(1 + y_i^{\tau})}\Big) \tag{3.6}$$

Since the expression for β_0 in (3.5) and the expression for β_0 in (3.6) are equal then the equation $(3.5) - (3.6)$ should evaluate to zero and we use this equation to find the position and the value of $\hat{\beta}_1$. To find the position β_1 a sequence of $\hat{\beta}_1$'s of length m must be created first. To find the index of $\hat{\beta}_1{}^{(1)}$ an initial $\hat{\tau}^{(1)}$ must be provided and then a search is performed and a position of $\hat{\beta}_1{}^{(1)}$ is returned and this is the position where $\hat{\beta}_1{}^{(1)}$ makes the equation $(3.5) - (3.6)$ closest to zero and we choose this position value to be $\hat{\beta}_1{}^{(1)}$. To find the $\hat{\beta}_0{}^{(1)}$ and $\hat{\beta}_1{}^{(1)}$ is simple since we can use the formula provided and this process is recursive.

Iterative Procedure Method: Estimation of β_0, β_1 and τ.

The recursive algorithm was implemented and 10 estimations of β_0, β_1 and τ were found in sample sizes of length 25, 250, 750 and 1000 and the average was considered to be the maximum likelihood estimation. To find the maximum likelihood estimation of the parameters, all data points within the sample groups were from a uniform distribution,$[0, 1]$, and it was decided that the actual value of the parameters of β_0, β_1 and τ were 1,1 and 0.5 respectively.

The maximum likelihood estimation of $\tau = 0.5$ was approximately accurate for all sample groups but it had the best estimation in sample size 1000 whereas the estimation in

sample size 250 slightly deviated from the actual value since it had a large variance. The estimation of $\beta_0 = 1$ was good for all groups but the variances were too high and it was seen in sample size 1000 the average estimation for β_0 was 1.14 which is quite high. Nevertheless, the estimation for sample size 25 was the best since and had the lowest variance. The estimations for $\beta_1 = 1$ in all sample groups were good on average although it did overestimate it slightly. The estimation in sample size 25 was very poor since the sample variance was too high in comparison to other sample groups.

TABLE 3.3: Table Maximum Likelihood Estimation for τ, β_0 and β_1

Parameter	**Sample Size**	**Sample Mean**	**Sample Variance**
$\tau = 0.5$	25	0.5108412	0.005752276
	250	0.5302954	0.009826608
	750	0.5292545	0.02147648
	1000	0.4958412	0.001226574
$\beta_0 = 1$	25	0.9997072	0.008576555
	250	1.001944	0.04606975
	750	0.9924185	0.1324405
	1000	1.148225	0.2854984
$\beta_1 = 1$	25	1.142833	0.1525783
	250	1.049506	0.04948899
	750	1.029952	0.0205219
	1000	1.031995	0.007592003

From figure 3.5 it was observed that the convergence rate for $\tau = 0.5$ was very rapid for all sample groups and we clearly see that after the fourth iteration the estimation of τ approaches the actual value in sample size 25 and 1000. However, for the first three iterations for all sample groups, the estimation was slightly inaccurate and had high variances. It must be noted that for sample sizes 25, 250 and 750 for the first three iterations it followed a similar volatility pattern whereas, for sample size 1000 the behavior was the opposite. From figure 3.6, the convergence rate for parameter τ was rapid in general, but in particular, it was the fastest after the third iteration in sample size 25. Since the estimation for the parameter, τ almost evaluates to its actual value. However, the convergence rate in sample size 1000 was not linear until the eighth iteration. Also, the first four iteration in all sample sizes they had the same volatility but for sample size 1000 the volatility was much higher than other groups.

From figure 3.7, it was observed that the estimation evaluates to the actual value after the fourth iteration for sample size 25, 250 and 750 but the convergence in sample size 1000 happened after its sixth iteration. Based on the results, the estimations were very accurate and the convergence rate was rapid for all groups although only 10 iterations (estimation) were found. This demonstrates that the iterative procedure is very powerful in estimating the parameters.

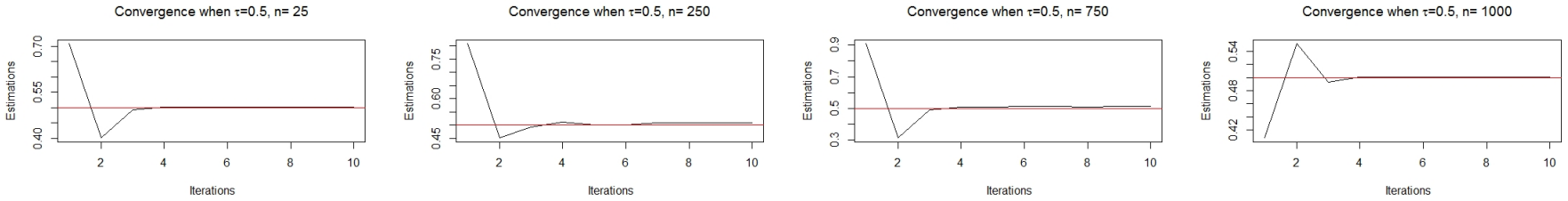

FIGURE 3.5: The Maximum Likelihood Estimation of $\tau = 0.5$

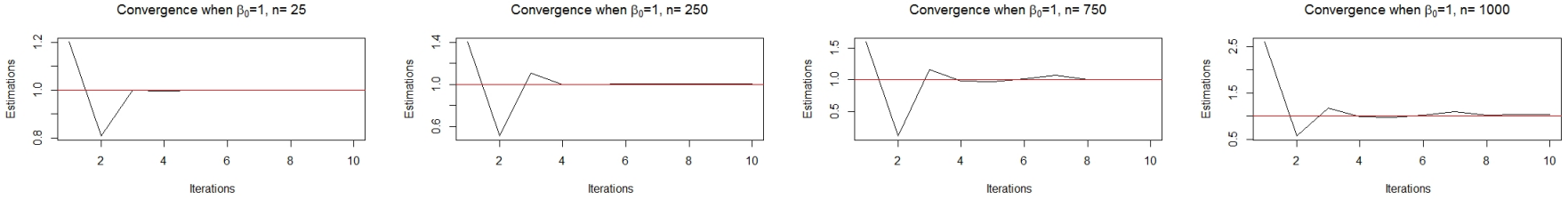

FIGURE 3.6: The Maximum Likelihood Estimation of $\beta_0 = 1$

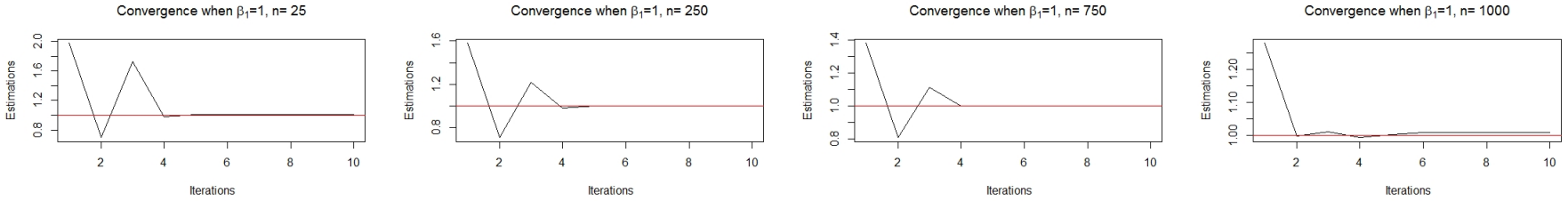

FIGURE 3.7: The Maximum Likelihood Estimation of $\beta_1 = 1$

Newton Method: Estimation of β_0, β_1 and τ.

Newton method was implemented to solve the following system of non-linear equations directly

$$\begin{aligned} \frac{\partial \ell}{\partial \tau} &= 0 \\ \frac{\partial \ell}{\partial \beta_0} &= 0 \\ \frac{\partial \ell}{\partial \beta_1} &= 0, \end{aligned} \tag{3.7}$$

with the same initial conditions and the method was repeated 10 times (trials) in each sample groups with different uniform distribution dataset (I.I.D). The average was considered to be the maximum likelihood estimation of parameters and the table below summarises the result.

TABLE 3.4: Newton's Estimation of τ, β_0 and β_1.

Parameter	Sample Size	Sample Mean	Standard Variance
$\tau = 0.5$	25	0.5056424	1.478712e-05
	250	0.4970208	0.1686975
	750	0.5006147	0.0192996
	1000	0.5090208	0.003854503
$\beta_0 = 1$	25	1.009755	0.000530539
	250	1.057408	0.002583182
	750	1.019616	0.001576827
	1000	0.9974079	0.001082324
$\beta_1 = 1$	25	1.009996	0.001529842
	250	1.006223	1.59577
	750	1.000149	1.091382
	1000	0.9932228	1.182038

It was observed that the estimation for $\tau = 0.5$ was very accurate on average for all groups but it was the best in sample size 25 since it had the lowest variance. The estimations for $\beta_0 = 1$ was good on average for all groups and was the best in sample size 25 and the least in sample size 250. The estimations for β_0 and β_0 was excellent across all groups on average. From figure 3.8, all trials estimation for $\tau = 0.5$ was very close to its actual value but it never hit the actual value. It should be noted that for all sample groups the estimations were in the neighborhood of the true value whereas in sample size 750 and 1000 the estimation did equal to the value 0.5 in trial 3 in both cases. Also, from figure 3.9 the estimations were excellent for all groups but it did overestimate it most of the times.

Based on figure 3.9 and 3.10 we observed that in sample size 1000 the estimations for each trial was good and the variation between the trials was happened to be low, it can be concluded that in general Newton's method performs the best on large datasets.

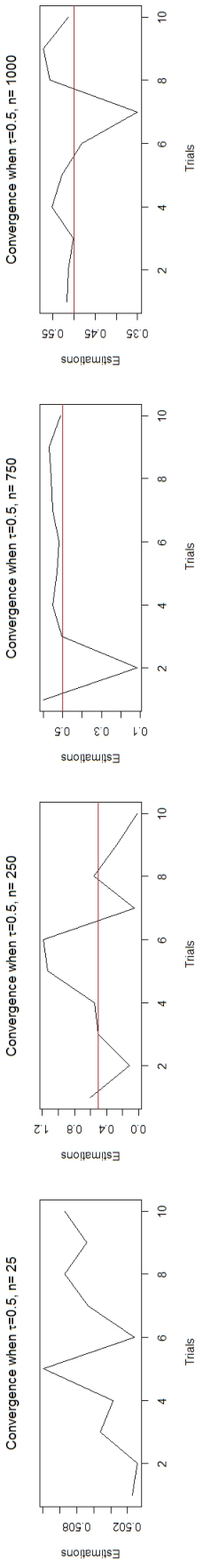

FIGURE 3.8: Newton Estimate of $\tau = 0.5$

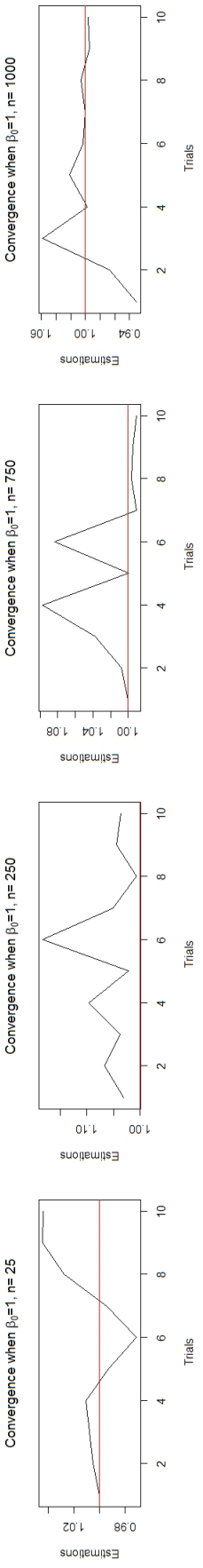

FIGURE 3.9: Newton Estimate of $\beta_0 = 1$

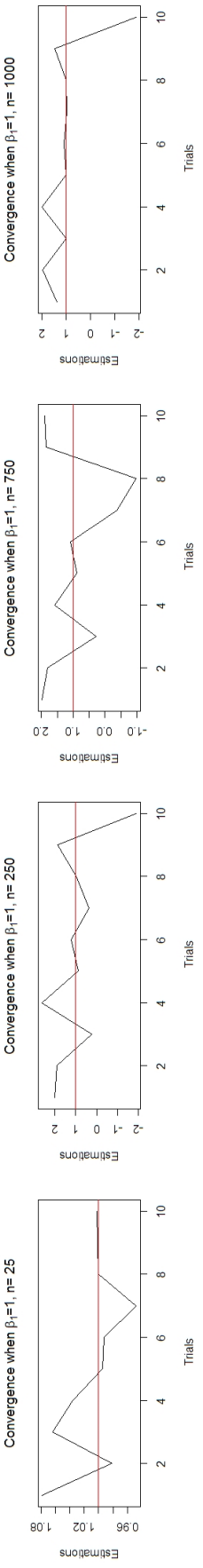

FIGURE 3.10: Newton Estimate of $\beta_1 = 1$

3.0.4 Ordinary Least Squares Estimation with Fixed τ.

For $i = 1, \ldots, n$, define U_i as

$$U_i = ln[ln(1 + y_i^{\tau})] - 0.5772,$$

and $\mathbf{U}$ be the collection vector of all U_i's and $\mathbf{X}$ be the design matrix, then the linear estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is given by

$$\boldsymbol{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{U} = \beta_0 + \beta_1 x + \boldsymbol{\varepsilon} \tag{3.8}$$

Data in sample size 25, 250, 750 and 1000 was chosen to be from a normal distribution with mean zero and variance one. The initial conditions of β_0 and β_1 were both and τ was fixed at 0.5. Ordinary least square was used 10 times to find the regression coefficients estimation and the average was considered as reference for the analysis. The table below shows the result found.

TABLE 3.5: Table using booktabs.

Parameter	Sample Size	Sample Mean	Standard Error	R Squared
$\beta_0 = 1$	25	1.07156	0.51128	0.056662
	250	1.22856	0.16775	0.0152419
	750	1.395315	0.095416	0.002239774
	1000	1.438978	0.081995	0.000807328
$\beta_1 = 1$	25	1.00289	0.88677	
	250	0.54026	0.29261	
	750	0.181163	0.165333	
	1000	0.099437	0.141676	

From the table above it can be seen that the estimation for $\beta_0 = 1$ was very poor and inaccurate on average for all groups. Also, each group respected average standard error was very high. Nevertheless, it can be seen that for the estimation of $\beta_0 = 1$ was only closest to its actual value in sample size 25 and the estimation was the poorest in the sample size of 1000. The estimation for $\beta_1 = 1$ was also poor in general for all sample sizes but the estimation is very accurate in sample size 25 although it has a large average standard error. The variation explained by $\boldsymbol{\beta}$ in each sample size was poor (R-squared $< 5\%$) on average. From figure 3.11 it can be seen that the ordinary least squares estimation for β_0 in sample size 25 was very accurate for the first 7 trials but for the remaining trials the estimation deviated significantly from the actual value. For the remaining sample sizes, their respected plot did show that the ordinary least square overestimates the parameter.

From figure 3.12, it can be seen that the estimation for β_1 is only good for sample size 25 but for the remaining sample sizes the ordinary least squares estimation of β_1 was poor and underestimated for all trials. Thus, it can be said that the ordinary least squares estimation for β_0 and β_1 when τ is held constant is accurate for small sample sizes and for large sample size the algorithm is poor and inaccurate.

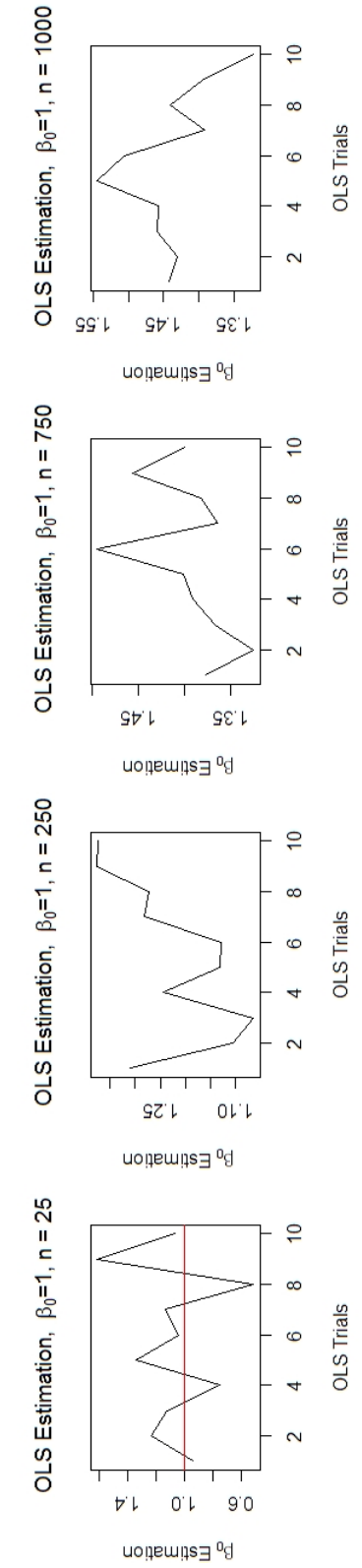

FIGURE 3.11: Maximum Likelihood Estimation of $\beta_0 = 1$

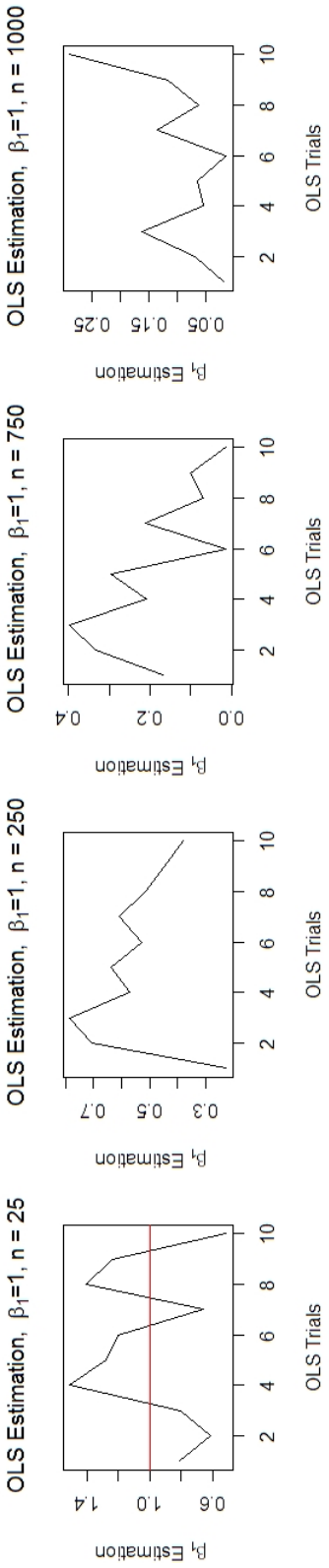

FIGURE 3.12: Maximum Likelihood Estimation of $\beta_1 = 1$

3.0.5 Ordinary Least Square Estimate of Regression : Big Data

The split and combination algorithm was implemented. The sample data had 1000 observations and generated from a uniform U[0,1]. The dataset was partitioned into three sub-matrix ($\boldsymbol{X}_1$,$\boldsymbol{X}_2$ and $\boldsymbol{X}_3$) and regression was fit to each partitioned sub-matrix using the ordinary least square estimation where τ was fixed at 0.5 and the initial values of β_0 and β_1 were both one.

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{U}_1 = c(1.45026572, -0.01203779)^T \\ \hat{\boldsymbol{\beta}}_2 &= (\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T\boldsymbol{U}_2 = c(1.49084683, 0.03627158)^T \\ \hat{\boldsymbol{\beta}}_3 &= (\boldsymbol{X}_3^T X_3)^{-1}\boldsymbol{X}_3^T\boldsymbol{U}_3 = c(1.57011747, -0.28473598)^T.\end{aligned} \tag{3.9}$$

Combining the results in (3.9) using the following formula solves the original problem

$$\hat{\boldsymbol{\beta}} = (\Sigma_{k=1}^3\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}(\Sigma_{k=1}^3\boldsymbol{X}_k^T\boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k) = c(1.502079, -0.08453294)^T \tag{3.10}$$

Since the dataset is small enough we can still apply the ordinary least squares estimation method and find $c(1.50652, -0.09033)^T$. The regression coefficients are identical to the coefficients found by the split and combination method. Thus, it can be said that the split and combination algorithm is a fast and efficient algorithm for overcoming big data problems especially when data are streaming.

Bootstrapping Ordinary Least Square Estimation for Big Data

Recalling back to the split and combination results it was assumed that the residuals were normally distributed and whenever it is not fulfilled the regression coefficients can be misleading. Thus, bootstrapping is a resampling method without replacement and a tool for making statistical inferences when standard parametric assumptions are questionable.
Bootstrapping was implemented to determine the accuracy of the regression from a non-parametric perspective. The data was re-sampled 1000 times with replacement and the β_0 and β_1 estimate were 1.507016 and -0.09002111 respectively. If compared to the ordinary least square (OLS) estimate the values are identical and it can be said the OLS estimation of the Burr regression coefficient is accurate without the need of checking the normality of the residuals. From figure 3.13 it can be observed that when re-sampling the data 1000 the regression coefficients are close to the OLS estimations (dashed line) and their respected quantiles standard normal plot shows that the bootstrap estimation of the regression coefficients are normally distributed and their respected values are close to the ordinary least squares estimation. According to the law of large numbers, it can be said that as the number of the observation increases the estimation improves and has the tendency to reach the actual ordinary least squares estimation.

Having bootstrapped the Burr regression coefficients 1000 it happens that the ordinary least squares estimation was accurate from a non-parametric approach but it is a question whether the coefficients are significant or not. Thus, the following hypothesis can be tested

$$\boldsymbol{H}_0 : \beta_0 = \beta_1 = 0, \quad \boldsymbol{H}_1 : \text{ One of } \beta's \text{ are non-zero.}$$

The bootstrapped 95% confidence intervals for β_0 and β_1 are (1.352, 1.675) and (-0.3752, 0.2056) respectively. It can be said that the coefficient β_0 was significant since its 95% confidence interval does not contain zero, whereas the coefficient β_1 is insignificant. Thus, at 95% confidence level, the regression intercept does explain most of the variation of the response variable. However, the data is random and therefore it is fine for the β_1 to be insignificant but the objective was to show how the split and combination and bootstrapping works in Burr regression.

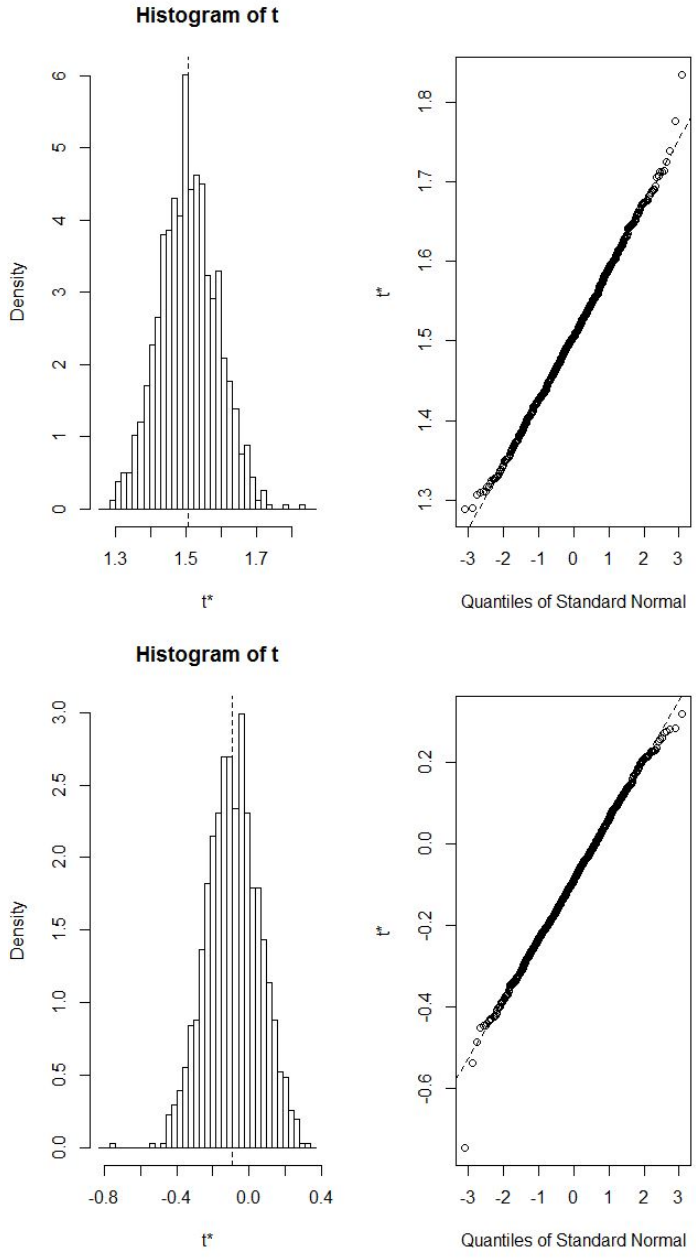

FIGURE 3.13: Histogram: Bootstrap of Burr regression coefficients

3.0.6 Ordinary Least Square Estimation: Simultaneously

The third algorithm requires Newton's method to solve the following equation

$$\frac{1}{n-1}\sum_{i=1}^{n}\exp(\boldsymbol{\beta}^T\boldsymbol{X})\ln(1+\boldsymbol{y}^{\tau}) = 1 \tag{3.11}$$

For our simulation study, the sum of squares of the differences between the LHS and the RHS of (3.11) was solved

$$\left(\frac{1}{n-1}\sum_{i=1}^{n}\exp(\boldsymbol{\beta}^T\boldsymbol{X})\ln(1+\boldsymbol{y}^{\tau}) - 1\right)^2 = 0. \tag{3.12}$$

For the study, the same sample size was used and the same initial conditions that were used in the previous section. Newton's method was implemented 10 times and the average was considered the best estimation. The table below are the results found.

TABLE 3.6: Ordinary Least Estimation : Newton's Methods Simultaneous

Parameter	Sample Size	Sample Mean	Standard Variance
$\tau = 0.5$	25	0.53163619	0.0035674
	250	0.54063387	0.000738936
	750	0.5509587	0.000431207
	1000	0.55869593	0.000152406
$\beta_0 = 1$	25	0.98914562	0.006788014
	250	0.98386316	0.001802332
	750	0.97369175	0.000531695
	1000	0.97172947	0.000217455
$\beta_1 = 1$	25	1.02899552	0.00053084
	250	1.02850355	0.000145426
	750	0.97369175	0.000152425
	1000	1.02138458	0.000105794

The estimation for $\tau = 0.5$ was good on average for all sample sizes and the estimation was only closest to the actual value in sample size 25 although the estimation is overestimated on average. The estimation for $\beta_0 = 1$ was very accurate for all sample sizes although it is underestimated on average. The estimation was only closest to the actual value in sample size 25 and the least in sample size 250. Also, the estimation for $\beta_1 = 1$ was good although it is overestimated. It is seen again that the estimation was good in sample size 25 and the least in 250 on average.

It can be seen that the estimation in Newton trial 5 in sample size 25 was the lowest and highest for trial 7 but for other trials, it was precisely close to the actual value. The estimation in sample size 750 and 1000 was overestimated across all trials and for all trials, it did fluctuate around $0.51 < \tau < 0.58$.
The estimation for $\beta_0 = 1$ was good across all trials and the estimations in sample size 25 and 250 were only closest to the actual value for all trials. It is visible that the estimations

were underestimated in sample size 750 and 1000 significantly.

The estimation for $\beta_1 = 1$ was very good across all trials. The estimation in sample size 25 was above the actual value and only below only for trial 7. For sample size 250 and 750, the estimations were slightly higher than the actual value but the estimation was extremely accurate for trial 4 in sample size 750 since the estimation was exactly the actual value.

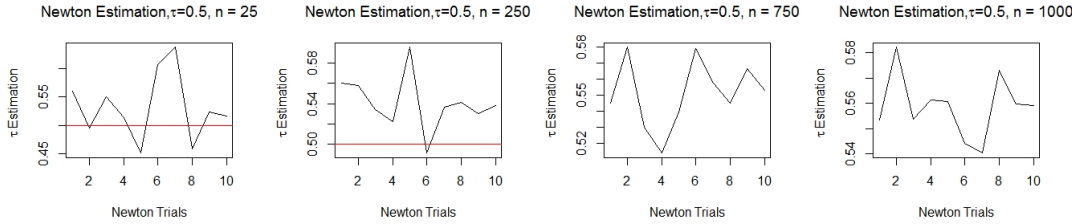

FIGURE 3.14: Ordinary Least Square Estimate of $\tau = 0.5$

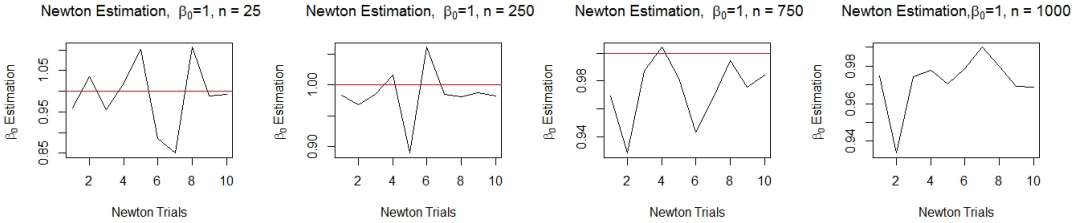

FIGURE 3.15: Ordinary Least Square Estimateof $\beta_0 = 1$

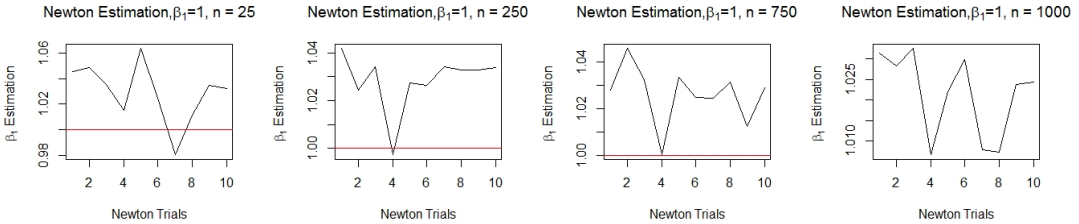

FIGURE 3.16: Ordinary Least Square Estimate of $\beta_1 = 1$

Chapter 4

Discussion

The aim of this project was to fit a regression model to a Burr type XII distribution through a transformation on one of its shape parameter namely the location parameter λ and the accuracy of three numerical algorithms that estimate the coefficients of a Burr regression model and the shape parameter. The algorithms that were implemented were maximum likelihood estimation, ordinary least square estimate where one shape parameter is held constant and ordinary least square where both the shape parameter and the regression coefficients are unknown. It must be noted that for each of the three algorithms the unknown parameters were found using the iterative procedure and optimisation namely Newton's method. To compare the accuracy of the three algorithms we used several sample sizes namely 25, 250, 750 and 1000 and in regards to the regression part for simplicity, we estimated the coefficients β_0 and β_1. Thus, the transformation $\lambda = \exp(\beta_0 + \beta_1 x)$ was fitted to a Burr type XII distribution.

From the first method, Maximum likelihood estimation was used to estimate the parameters τ and β iteratively. The estimation for $\lambda = 1$ and $\tau = 0.5$ through the iteration procedure was very good in general and it can be seen that the algorithm performed well on sample size 750 and worse on sample size 1000. Nevertheless, this method has shown fast convergence for the actual value of β and τ after a certain amount of iterations but however for sample size 1000 its convergence was rather complicated and non-linear.
When the transformation was applied [$\lambda = \exp(\beta_0 + \beta_1 x)$] with the actual value being $\tau = 0.5$, $\beta_0 = 1$ and $\beta_1 = 1$, both iterative procedure and Newton's method was implemented The results from the iteration method did show strong convergence even after few iterations for all sample sizes. The estimation of the parameters using Newton's method was very accurate on average since it was implemented in 10 different datasets. Thus, two methods were applied to compute the MLE estimation of the regression coefficients and τ and the results were very accurate and our results do support Rao's argument that the maximum likelihood estimation of parameters through the iterative procedure converges to its actual value as the number of iteration increases.

When ordinary least square with $\tau = 0.5$ was repeated 10 times and the average estimation of the regression coefficients were very poor and inaccurate especially for β_1 whereas the estimation of the coefficients for sample size 25 was very accurate. The estimation of the regression coefficients gets poorer on average as the sample size increases and their respected average R-squared value is getting closer to zero.

Also, we have used the second method to show how to overcome big data problem. A dataset with 1000 observations and split and combination was implemented with $K = 3$. The result was shown to be very accurate and the validity of the result was checked where the data

set was re-sampled 1000 and the results did show that the split and combination algorithm estimations to be very accurate. Newton's method was implemented 10 times with different data and estimation of the regression coefficients and τ simultaneously was very accurate for all sample sizes on average with a low standard error. However, after repeating the algorithm 10 times it was observed that the estimation for τ and β_1 was slightly overestimated and the estimation for β_0 was slightly underestimated.

Thus, based on the results it is highly recommended that when the sample size is large maximum likelihood estimation should be used. This is supported by an article on maximum likelihood estimation which stated that MLE is biased for a small dataset and for large samples one expects to get the true value on average[3]. For small samples, it is recommended that ordinary least squares with the shape parameter being held constant are applied since from our study it was shown to be very accurate and precise.

Chapter 5

Recommendation

The objectives of this paper were to determine the accuracy of the three methods that estimate the parameters after applying a transformation to one of its shape parameters. The results were very accurate in general for all three methods, however, if we made some improvements to the simulation study the results could have been much better. The first improvement suggested is that instead of estimating the parameters with a single actual value we could have applied each of the three methods on different initial and true values. This would give us a good overview of the accuracy of each method as one method may work for one initial value but not for another.
In the simulation study the chosen sample sizes were small in general and this is a problem since some methods tend to favor small dataset and some prefer large dataset, for example, maximum likelihood estimation of the parameters get closer to the initial value provided that the sample size is large and vice versa. Therefore, if each method were implemented small sample sizes and very large sample then we can generalise which sample sizes the algorithms work best on and secondly we avoid any biases.
In addition, when the algorithm in section (3.03) was implemented we expected a bias estimation for β_1 and it does have a direct effect on other parameters estimation. This is because we chose the list of β_1's and we only chose 200 values ranging from 0.05 to 2 for the search to happen. This is a problem for the algorithm because it would only search the list and the estimation for β_1 is always positive. Although, the list is random if we allowed the list to have negative values the estimation of the parameters could have been much better and avoid biases.
Also, Newton's method for optimisation was used in the simulation study and this algorithm has two problems. The first problem is that the initial guess must be close to the actual value else the estimation diverges away from the actual value. Secondly, it is computationally expensive if we have too many parameters to estimate[10] and if we have too many observations the algorithm becomes complicated and less efficient. Thus, care should be taken when implementing Newton's method.
Although, it was shown how the ordinary least square estimates the parameters with one shape parameter being held constant deals with big data we recommend statisticians to study how ordinary least square estimation with τ being unknown deals with big data.

This research paper introduced three algorithms that estimates shape parameters and the regression coefficients when the $\lambda \equiv \lambda(\boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x})$ transformation was applied to the distribution. Some of these methods are complicated for example the maximum likelihood estimation is computationally expensive. Having studied Abbasi and his collaborator's research paper he mentions that a multilayer perceptron neural networks can estimate the Burr type XII distribution parameters.

Let Y follow a Burr distribution then there is a standardise transformation between Y and a random variable X. The transformation is expressed as

$$\frac{Y-\mu_Y}{\sigma_Y}=\frac{X-\mu_X}{\sigma_X}.$$

Considering $Y=\Big((X-\mu_X)/\sigma_X\Big)\sigma_Y+\mu_Y$[11]. Then the PDF of Y is given by

$$f(y)=\frac{\sigma_Y}{\sigma_X}\lambda\tau\Big[\Big(\frac{X-\mu_X}{\sigma_X}\Big)\sigma_Y+\mu_Y\Big]^{\tau-1}\Big\{1+\Big[\Big(\frac{X-\mu_X}{\sigma_X}\Big)\sigma_Y+\mu_Y\Big]^{\tau}\Big\}^{-(\lambda+1)},$$

where

$$x\geq\mu_X-\mu_Y\Big(\frac{\sigma_X}{\sigma_y}\Big),\quad \tau,\lambda>0.$$

Since the moment-generating function of a Burr type XII distribution exists we can use it to find the mean (μ_X) and the variance (σ_X) of the distribution.

$$\begin{aligned}\mu_X&=\frac{\lambda}{\Gamma(\lambda+1)}\Gamma\Big(\lambda-\frac{1}{\tau}\Big)\Gamma\Big(1+\frac{1}{\tau}\Big)\quad\text{exists only for } \tau\lambda>1.\\ \sigma_X^2&=\frac{\lambda}{\Gamma(\lambda+1)}\Big(\Gamma(\lambda-\frac{2}{\tau}\Big)\Gamma\Big(1+\frac{2}{\tau}\Big)-\lambda\Big[\Gamma\Big(\lambda-\frac{1}{\tau}\Big)\Gamma\Big(1+\frac{1}{\tau}\Big)\Big]^2\quad\text{exists only for } \tau\lambda>2.\end{aligned}\tag{5.1}$$

If we use the multilayer neural perceptron with four layers to estimate the Burr type XII distribution parameters [11]. Where, skewness and kurtosis are used as input variables (input later) and μ_Y, σ_Y and $1/\tau$ are the outputs of the multilayer neural perceptron neural network model (output layer) [11]. Then we can estimate the parameter λ using the formula

$$\hat{\lambda}=\frac{n}{\sum_{i=1}^{n}\ln\Big[1+(((x_i-\mu_X/\sigma_X)\sigma_Y+\mu_Y)^{\tau})\Big]}.$$

This equation is derived using the maximum likelihood estimation.

Once the covariate information $\boldsymbol{x}$ is available then if we use the transformation $\lambda=\lambda(\boldsymbol{x})=\exp(\boldsymbol{\beta}^T\boldsymbol{x})$ then the estimation of $\boldsymbol{\beta}$'s becomes a hot topic. Therefore, it would be highly appreciated if a PhD research was done in this area.

Chapter 6

Appendix

6.1 Newton's Method for Optimisation

We can find solutions to a homogeneous system of equations that have the form

$$\begin{aligned} f_1(x_1,x_2,x_3,\ldots,x_n) &= 0 \\ f_2(x_1,x_2,x_3,\ldots,x_n) &= 0 \\ f_3(x_1,x_2,x_3,\ldots,x_n) &= 0 \\ &\vdots \\ f_n(x_1,x_2,x_3,\ldots,x_n) &= 0 \end{aligned} \tag{6.1}$$

For convenience we can think of $(x_1,x_2,\ldots,x_n)$ as a vector $\mathbf{x}$ and $(f_1,f_2,\ldots,f_n)$ as a vector-valued function $\mathbf{f}$. With this notation, we can write the homogeneous system of equation as

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}, \tag{6.2}$$

As in Newton's method for one variable, we need to start with an initial guess x_0. In general, it is hard to find a good initial guess for a multivariate function.[15] Given $\mathbf{x}_0$,
let

$$\Delta\mathbf{x} = \mathbf{x}_1 - \mathbf{x}_0. \tag{6.3}$$

In the single variable case, Newton's method was derived by considering the linear approximation of the function f at the initial guess x_0.[15] From multivariable calculus, the following is the linear approximation of $\mathbf{f}$ at $\mathbf{x}_0$, for vectors and vector-valued functions:

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{x_0}) + D\mathbf{f}(\mathbf{x_0})(\mathbf{x_1} - \mathbf{x_0}). \tag{6.4}$$

Here $\mathrm{D}\mathbf{f}(\mathbf{x}_0)$ is an n x n matrix whose entries are the various partial derivative of the components of $\mathbf{f}$, evaluated at $\mathbf{x}_0$. Specifically,

$$D\mathbf{f}(\mathbf{x_0}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x_0}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x_0}) & \frac{\partial f_1}{\partial x_3}(\mathbf{x_0}) & \ldots & \frac{\partial f_1}{\partial x_n}(\mathbf{x_0}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x_0}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x_0}) & \frac{\partial f_2}{\partial x_3}(\mathbf{x_0}) & \ldots & \frac{\partial f_2}{\partial x_n}(\mathbf{x_0}) \\ \frac{\partial f_3}{\partial x_1}(\mathbf{x_0}) & \frac{\partial f_3}{\partial x_2}(\mathbf{x_0}) & \frac{\partial f_3}{\partial x_3}(\mathbf{x_0}) & \ldots & \frac{\partial f_3}{\partial x_n}(\mathbf{x_0}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x_0}) & \frac{\partial f_n}{\partial x_2}(\mathbf{x_0}) & \frac{\partial f_n}{\partial x_3}(\mathbf{x_0}) & \ldots & \frac{\partial f_n}{\partial x_n}(\mathbf{x_0}) \end{pmatrix} \tag{6.5}$$

We need to find $\mathbf{x}$ such that $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, so let's choose $\mathbf{x}_1$ so that

$$\mathbf{f}(\mathbf{x_0}) + D\mathbf{f}(\mathbf{x_0})(\mathbf{x_1} - \mathbf{x_0}) = \mathbf{0}. \tag{6.6}$$

Since $\mathrm{D}\mathbf{f}(\mathbf{x}_0)$ is a square matrix, we can solve this equation by

$$\mathbf{x_1} = \mathbf{x_0} - (D\mathbf{f}(\mathbf{x_0}))^{-1}\mathbf{f}(\mathbf{x_0}), \tag{6.7}$$

provided that the matrix $\mathrm{D}\mathbf{f}(\mathbf{x}_0)$ is invertible.[15] In practice inverse of matrix $\mathrm{D}\mathbf{f}(\mathbf{x}_0)$ is never used but Rather, we can do the following. First solve the equation

$$D\mathbf{f}(\mathbf{x_0})\Delta\mathbf{x} = -\mathbf{f}(\mathbf{x_0}). \tag{6.8}$$

Since $D\mathbf{f}(\mathbf{x}_0)$ is a known matrix and $-\mathbf{f}(\mathbf{x}_0)$ is a known vector, this equation is just a system of linear equations, which can be solved efficiently.[15] Once the solution of vector $\Delta\mathbf{x}$ is found, we can obtain our improved estimate $\mathbf{x}_1$ by

$$\mathbf{x}_1 = \mathbf{x}_0 + \Delta\mathbf{x}. \tag{6.9}$$

For subsequent steps, we have the following process:

- Solve $D\mathbf{f}(\mathbf{x}_i)\Delta\mathbf{x} = -\mathbf{f}(\mathbf{x}_i)$ for $\Delta\mathbf{x}$.
- $\mathbf{x}_{i+1} = \mathbf{x}_i + \Delta\mathbf{x}$.

6.2 Rao's Simulation Study: Quantile Plots

Parameter	Value	Sample mean	Sample variance	Asymptotic variance	Ratio
$n = 200$					
λ	0.50	0.5167	0.0154	0.0117	1.3162
β	5.00	5.3681	1.9300	1.5069	1.2808
θ_0	−1.00	−0.9884	0.0240	0.0222	1.0810
θ_1	2.00	2.0025	0.0274	0.0281	0.9751
$n = 500$					
λ	0.50	0.5077	0.0051	0.0047	1.0851
β	5.00	5.1336	0.6641	0.6027	1.1019
θ_0	−1.00	−0.9990	0.0090	0.0092	0.9783
θ_1	2.00	2.0037	0.0112	0.0114	0.9825
$n = 1000$					
λ	0.50	0.5033	0.0026	0.0023	1.1304
β	5.00	5.0583	0.3201	0.3014	1.0620
θ_0	−1.00	−0.9990	0.0045	0.0045	1.0000
θ_1	2.00	2.0015	0.0057	0.0055	1.0364
$n = 2000$					
λ	0.50	0.5015	0.0011	0.0012	0.9167
β	5.00	5.0372	0.1446	0.1507	0.9595
θ_0	−1.00	−0.9981	0.0023	0.0023	1.0000
θ_1	2.00	1.9995	0.0028	0.0028	1.0000

FIGURE 6.1: Simulation results for parametrisation I

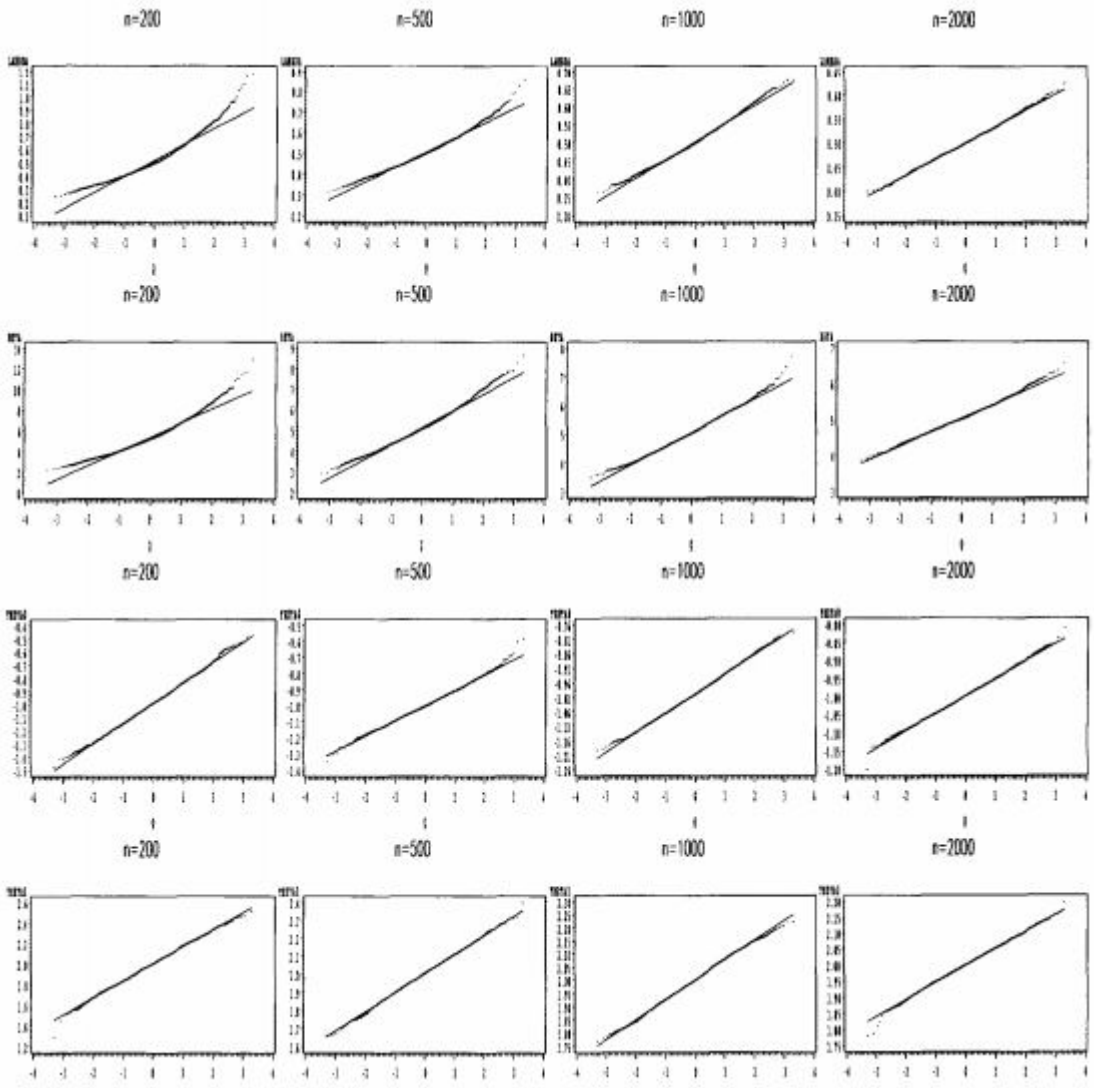

FIGURE 6.2: Normal Quantile plots for $\hat{\lambda}$, $\hat{\beta}$,$\hat{\theta_0}$, $\hat{\theta_1}$, Parametrisation I

6.3 R Codes

6.3.1 Maximum Likelihood Estimation of the Shape Parameter τ and λ via Iterative Procedure

Part 1 of the R code below generates samples from a Burr type XII distribution using the inverse transform sampling method and the second part of the R code computes the maximum likelihood estimation of the shape parameters using an iterative procedure when the sample size is 25.

```
set.seed(15)
#---------- Part 1-------------------------------------------------------
u<- runif(25,0,1)
#Inverse Transformation Sampling Method
Simburr.f <- function(n, lambda, tau) {
u=runif(n)
y= (1/(u^(1/lambda))-1)^(1/tau)
y
}
y25=Simburr.f(n= 25,1,0.5) # Function returns samples
#---------- Part 2-------------------------------------------------------
# Computes the MLE of the shape parameters using the interarive Procecure
n=25
est=function(m){
tau=rep(0,m) # ero matrix
lamd=rep(0,m) #
tau[1]=0.5 #  Initial value
lamd[1]<-n/sum(log(1+y25^(tau[1]))) # MLE of lamda
for(i in 2:m){
lamd[i]=n/sum(log(1+y25^(tau[i-1])))
u=-sum(log(y25))+(lamd[i]+1)*sum((log(y25)*y25^tau[i-1])/(1+y25^tau[i-1]))
tau[i]=n/u # MLE of tau
}
cbind(lamd,tau) # Results stored in a matrix
}
output1=est(25) # Estimation of the parameters.
```

FIGURE 6.3: R code for Maximum Likelihood Estimation of τ and λ

6.3.2 Maximum Likelihood Estimation of the Shape Parameter τ and λ via Newtons Method

The R code below computes the maximum likelihood estimation of the shape parameters via Newton's method for one trial when the sample size is 25. The code can be amended to find the maximum likelihood estimation for other sample sizes by simply changing the sample size. Repeat this code and re-run part 1 of the previous code 10 times to get 10 different MLE values for the shape parameter τ and λ. Take the average of those 10 values and this value is the average maximum likelihood estimation of the shape parameter τ and λ.

```
# Newton Method to compute the MLE of lambda and tau
model= function(s) {
F= numeric(1)
F[1] = n/s[1] + sum(log(y25))-(s[2]+1)*sum((y25^s[1])*log(y)/(1+y25^s[1]))
F[2] = n/s[2] -sum(log(1+y25^s[1]))
# Partial derivatives of the log-likelihood function set to zero
F
}
ss$x = nleqslv(c(0.5,1),model,control = list(allowSingular = TRUE))
# Built in function nleqslv implements Newton's method
# Newton's method solves F[1] = 0 & F[2] = 0 with initial values
ss$x # MLE of the shape parameters for one trial
```

FIGURE 6.4: R code for Maximum Likelihood Estimation of τ and λ via Newton's method for one trial when sample size is 25.

6.3.3 Maximum likelihood estimation of the Burr regression

```
# Collect sample from the distribution with transformation
Simburr.f1 <- function(n, tau) {
u=runif(n)
x<- runif(n)
lambda = exp(1+x)
y= (1/(u^(1/lambda))-1)^(1/tau)
 y
}
y25 = Simburr.f1(25,0.5)
#-------------------------------------

est3 = function(m) {
x=runif(25)
Out = rep(0,m)
k =  seq(0.05,2,l=200)
for ( i in 1:m){
Out[i]=  log(mean(x)) -log(mean(exp(k[i]*x)*log(1+y25^0.5)*x) )
+ log(mean(log(1+y25^0.5)*exp(k[i]*x)))
}
Out
}
p=est3(200)
k[which.min(abs(p-0))] # Estimation of beta_{1}
```

FIGURE 6.5: R code for Maximum Likelihood Estimation of β_0, β_1 and λ iteration procedure when sample size is 25 (I).

```
tau[i] =(( mean((exp(b0[i-1]+b1[i-1]*x)+1)*(y25^tau[i-1])*log(y25)/(1+y25^0.5))
b0[i]= -log(mean(log(1+y25^0.5)*exp(b1[i-1]*x)))
```

FIGURE 6.6: R code for Maximum Likelihood Estimation of β_0, β_1 and λ iteration procedure when sample size is 25 (II).

The R code above is coded badly but we did still find the maximum likelihood estimation of the parameters. Painfully, we found 10 iterations manually.

6.3.4 Maximum likelihood estimation of the Burr regression via Newton's Method

```
# Newton's Method
x = runif(25)
require(nleqslv)
n=25
mod= function(s) {
F= numeric(3)
F[1] = n/s[1] + sum(log(y25))-sum((exp(s[2]+s[3]*x)+1)*(y33^s[1])*log(y25)/(1+y25^s[1]))
F[2] = n -sum(exp(s[2]+s[3]*x)*log(1+y25^s[1]))
F[3] = sum(x)- sum(exp(s[2]+s[3]*x)*x*log(1+y25^s[1]))
# Newton's method solves F[1], F[2] and F[3] = 0 simultaneously
F
}
(ss = nleqslv(c(0.5,1,1),mod,control = list(allowSingular = TRUE)))
# MLE of the regression coefficients are returned
```

FIGURE 6.7: R code for Maximum Likelihood Estimation of β_0, β_1 and τ via Newton's method for one trial when sample size is 25.

6.3.5 Ordinary Least Square Estimation when τ Held Constant

```
x=runif(25)
Simburr.f <- function(n, tau) {
u=runif(n)
lambda = exp(1 + x)
y= (1/(u^(1/lambda))-1)^(1/tau)
y
}
y33=Simburr.f(25,0.5)
U_33 <- -log(log(1+y33^(0.5))) -0.5772
U_33
lm.fit1<- lm(U_33~x)
lm.fit1
```

FIGURE 6.8: R code for Ordinary Least Square Estimation of β_0, β_1 and λ with $\tau = 0.5$ when sample size is 25 (I).

6.3.6 Ordinary Least Square Estimation when τ Held Constant and Big Data.

```
# Collect samples from the distribution
set.seed(19)
Simburr.f1 <- function(n, tau) {
u=runif(n)
x<- runif(n)
lambda = exp(1+x)
y= (1/(u^(1/lambda))-1)^(1/tau)
y
}
y1 = Simburr.f1(1000,0.5)
x4 = runif(1000)
# data
U = -log(log(1+y1^(0.5))) -0.5772
lm(U~x4)
# lm solves the original problem is sample size is small
```

FIGURE 6.9: Split and Combination: Ordinary Least Square Estimation of β_0, β_1 and λ with $\tau = 0.5$ when sample size is 1000 for big data (I).

```
X<- matrix( x4,ncol=1, nrow = 1000 )
X_1 <- matrix(X[1:333,], nrow = 333, ncol = 1 )
X_2 <- matrix(X[334:667,], nrow = 334, ncol = 1 )
X_3 <- matrix(X[668:1000,], nrow =  333 , ncol = 1 )
# Split data into sub-matrix
fit1<- lm(U[1:333]~X_1)
fit2<- lm(U[334:667]~X_2)
fit3<- lm(U[668:1000]~X_3)

# Fit regression to each sub-data
beta<-  c( fit1$coefficients ,fit2$coefficients, fit3$coefficients)
matlist <- list(t(X_1)%*%X_1,t(X_2)%*%X_2,t(X_3)%*%X_3)
combmat <- matlist[[1]]+matlist[[2]]+matlist[[3]]
(solve(combmat))%*% ((t(X_1)%*%X_1)%*%beta[1:2]
+(t(X_2)%*%X_2)%*%beta[3:4] +(t(X_3)%*%X_3)%*%beta[5:6])
# Combines the results
```

FIGURE 6.10: R code for Maximum Likelihood Estimation of τ and λ via Newton's method for one trial when sample size is 25.

```
library(boot)
X = data.frame(U,x4)
bs <- function(formula, data, indices) {
d <- data[indices,] # allows boot to select sample
fit <- lm(formula, data=d)
return(coef(fit))
}
# bootstrapping with 1000 replications
results <- boot(data=X, statistic=bs,
R=1000, formula=U~x4)
plot(results, index=1,) # Histogram
plot(results, index=2)
boot.ci(results)
boot.ci(results, type="bca", index=1)
boot.ci(results, type="bca", index=2)
# Bootstrap 95% confidence for lambda and tau
```

FIGURE 6.11: Split and Combination: Ordinary Least Square Estimation of β_0, β_1 and λ with $\tau = 0.5$ when sample size is 1000 for big data (II).

6.3.7 Ordinary Least Square Estimation with τ being Unknown.

```
# Minimize the square of the difference of the means using optim function.
f <- function(par) {
b0 <- par[1]
b1 <- par[2]
ta <- par[3]
(mean(exp(b0+b1*x1)*log(1+y25^(ta))) - 1)^2
}
Y = Simburr.f1(25,0.5)
optim(c(1,1,0.5), f)
```

FIGURE 6.12: R code for Maximum Likelihood Estimation of τ and λ via Newton's method for one trial when sample size is 25.

Bibliography

[1] Beirlant et al. (1998). *Burr regression and portfolio segmentation*. https://www.sciencedirect.com/science/article/pii/S0167668798000456. [Online; accessed: August 2018].

[2] Zimmer et al. (1998). *The Burr XII Distribution in Reliability Analysis*. https://www.researchgate.net/publication/279549117_The_Burr_XII_Distribution_in_Reliability_Analysis. [Online; accessed: August 2018].

[3] Reliability HotWire (2001). *Maximum Likelihood Estimation*. http://www.weibull.com/hotwire/issue9/relbasics9.htm. [Online; accessed: August 2018].

[4] MIT (2006). *Properties of MLE: consistency, asymptotic normality*. https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture3.pdf. [Online; accessed: August 2018].

[5] Rao et al. (2006). *Burr-XII Distribution Parametric Estimation and Estimation of Reliability of Multicomponent Stress-Strength*. http://home.iitk.ac.in/~kundu/2MSSBD.pdf. [Online; accessed: August 2018].

[6] Chen et al. (2010). *A SPLIT-AND-CONQUER APPROACH FOR ANALYSIS OF EXTRAORDINARILY LARGE DATA*. http://www.stat.rutgers.edu/home/mxie/RCPapers/split_and_conquer_rev1_final.pdf. [Online; accessed: August 2018].

[7] Sigman (2010). *Inverse Transform Method*. http://www.columbia.edu/~ks20/4404-Sigman/4404-Notes-ITM.pdf. [Online; accessed: August 2018].

[8] Taylor (2016). *Heavy-tailed Distributions*. https://math.la.asu.edu/~jtaylor/teaching/Spring2016/STP421/lectures/stable.pdf. [Online; accessed: August 2018].

[9] Dartmouth (2017). *Law of Large Numbers*. https://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/Chapter8.pdf. [Online; accessed: August 2018].

[10] Luke Olson (2017). *newton's method and optimization*. https://relate.cs.illinois.edu/course/cs357-f15/file-version/2978ddd5db9824a374db221c47a33media/cs357-slides-newton2.pdf. [Online; accessed: August 2018].

[11] B.Abbas et al. *A neural network applied to estimate Burr XII distribution parameters*. https://doi.org/10.1016/j.ress.2010.02.001. [Online; accessed: September 2018].

[12] Shao et al. *THE BURR TYPE XII DISTRIBUTION WITH SOME STATISTICAL PROPERTIES*. https://doi.org/10.1623/hysj.49.4.685.54425. [Online; accessed: September 2018].

[13] Robert N. Rodriguez. *A Guide to the Burr Type XII Distributions*. www.jstor.org/stable/2335782. [Online; accessed: September 2018].

[14] Pandu R. Tadikamalla. *A Look at the Burr and Related Distributions*. https://www-jstor-org.ezproxy.brunel.ac.uk/stable/1402945. [Online; accessed: September 2018].

[15] Ohio University. *Nonlinear Systems - Newton's Method*. http://www.math.ohiou.edu/courses/math3600/lecture13.pdf. [Online; accessed: August 2018].

[16] Wikipedia. *Chi-squared distribution*. https://en.wikipedia.org/wiki/Chi-squared_distribution. [Online; accessed: August 2018].

[17] Wikipedia. *Digamma function*. https://en.wikipedia.org/wiki/Digamma_function. [Online; accessed: August 2018].

[18] Wikipedia. *Heavy-tailed Distribution*. https://en.wikipedia.org/wiki/Heavy-tailed_distribution. [Online; accessed: August 2018].