ON THE MAXIMIZATION OF THE LIKELIHOOD FUNCTION AGAINST LOGARITHMIC TRANSFORMATION

1. Obisesan K. O. (ko.obisesan@mail.ui.edu.ng) **2. Udomboso C. G.**

3. Osowole O. I. **4. Alaba O. O.**

Department of Statistics

University of Ibadan

Ibadan, Nigeria

Abstract

We consider maximum likelihood estimation logarithmic transformation irrespective of mass of density functions. The estimators are assumed to be consistent, convergent and existing. They are referred to as asymptotically minimum-variance sufficient unbiased estimators (AMVSU). We find that the likelihood function gives accurate result when maximized than the log-likelihood. This is because logarithmic transformation has potential problems. We consider a uniform case where the parameter θ cannot be estimated by calculus but order-statistics. We fit a truncated Poison distribution into data on damaged done after estimating λ by a Newton-Raphson Iterative Algorithm.

Résumé

Nous considérons l'estimation de la probabilité maximale de la transformation logarithmique indépendante de masse fonctions de densité. Les estimateurs sont assumés pour être logiques, convergent et existant. Ils sont connus sous le nom d'estimateurs asymptotiquement impartial suffisant en désaccord minimum (AMVSU). Nous trouvons que la fonction de densité de probabilité donne le résultat exact quand à la probabilité maximisée. C'est parce que la transformation logarithmique a des problèmes potentiels. Nous considérons un cas constant où le paramètre θ ne peut pas être estimé par calcul mais par ordre - statistique. Nous utilisons une distribution de Poison tronquée dans les données expérimentales endommagées après avoir estimé λ par un Algorithme itératif de Newton Raphson.

1.0 Introduction

The method of maximum Likelihood Estimation is popular with R.A Fisher, who according to history published two papers in the early 1920's. This method yields sufficient estimators whenever they exist, and that the estimators are Asymptotically-Minimum-Variance-Unbiased Estimators (AMVSU).

The essential feature of the method of maximum likelihood is that we look at the values of a random sample and choose as our estimate of the parameter (unknown), the value for which the probability of obtaining the observed data is a 'maximum'.

The method consists of maximizing the likelihood function with respect to θ (say) and refer to the value of θ which maximizes the likelihood function as the maximum likelihood estimate of θ. For both discrete and continuous random variable the procedure is similar but $f(x_1, x_2, ..., x_n / \theta)$ is the value of the joint probability density at the sample point $(x_1, x_2, ..., x_n)$.

Harry Van Zanton, in his paper of the Bernoulli (Bernoulli 11(4), 2005, 643-664) present a unified approaches to obtaining rates of convergence for the maximum likelihood estimator (MLE) in Brownian semi martingale models of the form

$$dX_t = \beta_t^{n,\theta} dt + \sigma_t^n dW_t, \qquad t \le T_n \qquad (A)$$

From maximum likelihood estimation, there are results for various models which state that the rate of convergence of the maximum likelihood estimator (MLE) is determined by the entropy of the (possibly infinitedimensional) parameter space relative to the Hellinger metric. Wong and Shen (1995) and Van de Geer (1995) consider independent and identically distributed (i.i.d) observations from a density ρ_0 belonging to a set ρ of densities with respect to a dominating measure μ. Our interest here is on the transformation of the variable by logarithm. We will therefore assume convergence, existence and consistency of the (AMVSU) estimators. For details of these matters, see Kutoyants (2004), Louikianova and Loukianov (2003a and 2003b), Van Zanten (2001).

2.0 Potential problems with transformation

When non-linear transformation is used we encounter problems estimating quantities such as means, variance, confidence limits, and regression coefficients in the transformed back into the original scale.

It may be difficult to understand or apply results of statistical analysis expressed in the transformed scale. More calculations are required. We may illustrate the bias referred to by lognormal distribution. Let x represent an untransformed lognormal datum, let y=lnx. An unbiased estimator of the log-transformed distribution is $\bar{y}$, the Arithmetic Mean of the y's. If it is transformed back to original scale by computing E $\bar{y}$, the geometric mean, we do not obtain an unbiased estimate of the mean of the untransformed (lognormal) distribution. A similar problem arises when estimating confidence limits for the mean of a lognormal distribution.

Koch and Link (1980, Vol.1, p.233) suggest that transformation may be useful "when the conclusions based on the transformed scale can be understood, when biased estimates are acceptable, or when the amount of bias can be estimated and removed because the details of the distribution are known".

Hoaglin, Mosteller, and Turkey (1983) point out that we lose some of our initiative understanding of data in a transformed scale, and that a judgement must be made as to when the benefits justify the "costs". They indicate that a transformation is likely to be useful when the ratio of the largest datum to the smallest datum in a data set is greater than about 20.

2.1 Limitation of logarithmic formations

Consider the relationship between an **index number** and a **logarithm:**

$$a^x = \log_a^x$$

This is also what happens in an exponential index that can also be written as natural logarithm that

$$e^x = \log_e^x \quad or \quad \ln x$$

We notice that what happens to an index number is not exactly what happens to its logarithmic transformation.

Consider the expansion of the exponential series by Taylor's theorem:

$$e^{x} = 1 + x + \frac{x^{2}}{2!} + \frac{x^{3}}{3!} + \ldots + \frac{x^{n}}{n!} + Rn(x,a)$$

Where $Rn(x,a) = \int_{a}^{x} \frac{(x-t)^{2}}{n!} e^{t} dt$

The term $R_n(x,a)$ is known as the remainder. The omission of the remainder makes the expansion a **Taylor's polynomial approximation** to the function, $f(x) = e^{x}$. The remainder is also known as the error function. The series expansion can only converge to $f(x) = e^{x}$ if and only if the limiting value of the error disappears. That is:

$$\lim_{x \to \infty} Rn(x,a) = 0$$

It is possible to estimate the remainder, but we do not intend to do that in this paper. However, when this remainder fizzles out, then

$$f(x) = \sum_{m=0}^{\infty} \frac{f^{(m)}(a)}{m!}(x-a)^{m}$$

Therefore,

$$e^{x} = \sum_{n=0}^{\infty} \frac{x^{n}}{n!} = 1 + x + \frac{x^{2}}{2!} + \frac{x^{3}}{3!} + \ldots \qquad where \quad a = 0$$

We recall that the resulting value is an approximation. This means too, that the logarithmic transformation would also be an approximation. The result above simply means that

$$\ln x = 1 + x + \frac{x^{2}}{2!} + \frac{x^{3}}{3!} + \ldots \qquad Since \quad \ln x = e^{x}$$

In another example, $\ln(1+x)$, at $a = 0$

$$\ln(1+x) = x - \frac{x^{2}}{2} + \frac{x^{3}}{3} + \frac{x^{4}}{4} + \ldots + (-1)^{n-1}\frac{x^{n}}{n} + \ldots$$

We can also obtain this by integrating the geometric series for $(1+t)^{-1}$ from $t = 0$ to $t = x$

$$\int_{a}^{x} (1+t)^{-1} dt = \int_{a}^{x} (1 - t + t^{2} - t^{3} + \ldots) dt$$

$$= \ln(1+t) \Big|_{0}^{x}$$

Now, by Taylor polynomial approximation expansion

$$\ln(1+t)\Big|_0^x = t\Big|_0^x - \frac{t^2}{2}\Big|_0^x + \frac{t^3}{3}\Big|_0^x - \frac{t^4}{4}\Big|_0^x + \ldots$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \ldots$$

The remainder approaches zero, and leaves an approximation. The procedure here is what also applies to the log-likelihood of probability functions. An approximation is not, and can never be full information.

An important device in many computational methods is reparametrization. A one-to-one transformation produces a new distribution with the Jacobian matrix of derivatives. It is often difficult to theorize about an appropriate normalizing transformation p-dimensional parameter θ hence the need for transformation of one dimension at a time.

The one-dimensional transformation improves the normal approximation asymptotically over the range of $-\infty$ to ∞ of the possible values of θ_i.

If $\phi_i = \log\theta_i$ removes the restriction, allowing θ_i to range over $(-\infty, \infty)$. Which in practical problems the distribution of a parameter constrained to be positive will be positively skewed?

The log transformation then gives the benefit of reducing skewness. If θ_i has a log-normal distribution the transformation is obviously exact.

2.2 Asymptotically-minimum-Variance-Sufficient Unbiased estimators

(*A) Given y successes in n trials, the minimum likelihood estimator of the parameter $\theta = p$ which is binomial $Y \square B(n,p)$ is $L(p)$. Then

$$L(P) = \binom{n}{y} P^y (1-P)^{n-y}$$

$$\frac{d(L(P))}{d(P)} = \binom{n}{y}\left[-P^y[(n-y)(1-P)^{n-y-1} + (1-P)^{n-y}(yP^{y-1})]\right]$$

$$\Rightarrow -\binom{n}{y} P^y\left[(n-y)(1-p)^{n-y-1}\right] + \binom{n}{y}(1-P)^{n-y}(y \cdot P^{y-1}) = 0$$

$$\Rightarrow -\binom{n}{y} P^y\left[(n-y)(1-p)^{n-y-1}\right] = -\binom{n}{y}(1-P)^{n-y}(y \cdot P^{y-1})$$

$$\Rightarrow (n-y)(1-p)^{-1} = \frac{y}{p}$$

$$\Rightarrow \frac{(n-y)}{(1-p)} = \frac{y}{p}$$

$$\hat{P}_{AMVSU} \frac{y}{n}$$

Since estimation in logarithmic transformation leads to biased estimates we may have to prefer equation (2) as obtained as against $\ln L(p)$ when it is transformed. Then we have:

$$\ln L(P) = \ln\binom{n}{y} + x \cdot \ln P + (n-x).\ln(1-P)$$

$$\Rightarrow \frac{d(\ln L(P))}{d(P)} = \left(\frac{x}{p} - \frac{n-x}{1-P}\right)$$

This yields same result. Then transforming is not necessary. We recall here that when non-linear transformation is used in estimating parameters such as means, variances, confidence limits, regression coefficients, we face other problems. It may be difficult to understand or apply results of statistical analysis expressed in the transformed scale. We shall also have to calculate so many quantities.

$(*B)$ *AMVSU*

If $x_1, x_2, \ldots, x_n$ are values of a random sample from an exponential population the maximum likelihood estimate of P in this case is $L(P) = f(x_1, x_2, \ldots, x_n / P)$, then $L(P) = P^n e^{-P\sum_{i=1}^{n} X}$

Solving equation (4) by product rule without transformation

$$\frac{dL(P)}{d(P)} = \left(-P^n \left(\sum X\right) e^{-p\sum X} + e^{-p\sum X} nP^{n-1}\right) = 0$$

$$\Rightarrow -P^n\left(\sum X\right)e^{-p\sum X} = -e^{-p\sum X} nP^{n-1}$$

$$\Rightarrow \Sigma X = \frac{n}{p} \Rightarrow \hat{P} = \frac{n}{\Sigma X} \qquad (5)$$

As we have obtained equation (5), we shall compare (6) below with (5). They give same result. Then

$$L(P) = f(x_1, x_2, ..., x_n / P) = P^n e^{-p\sum X}$$

$$\ln L(P) = n\log_e P - P\sum X_i(\log_e e)$$

$$= n\log_e P - P\sum X = 0$$

$$\frac{d\ln(L(P))}{d(P)} = \frac{n}{p} - \frac{\sum X}{1} = 0$$

$$\Rightarrow \frac{n}{p} - \frac{\sum X}{1}, \hat{P}\sum X = n$$

$$\hat{P} = \frac{n}{\sum X}$$

We should note here quickly that if the exponential $p.d.f$ has been defined as $f_x(x) = Pe^{-Px}$ which could have been $f_x(x) = \frac{1}{P}e^{-Px}$. We could set $f_x(x) = \lambda e^{-\lambda x}$ where $\lambda = \frac{1}{P}$. What we have is the form obtained as a special case of gamma distribution in (8) with $\alpha = 1$ and $\beta = P$ as

$$\begin{cases} \frac{1}{P}e^{-x/p} & x > 0 \\ 0 \ otherwise \end{cases}$$

$$f_x(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma\alpha} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 \ elsewhere & \alpha, \beta > 0 \end{cases}$$

We need to be careful with the usage of P as the parameter as it has to do with convenience. P is binomial is not same in exponential. It is rather convenient to consider P as a parameter depending on the distribution and type being considered.

If in (4), we have decided to use directly, the special case of gamma as $\frac{1}{P}e^{-x/P}$

$$L(P) = \prod f(x_i, P) = \left(\frac{1}{p}\right)^n \cdot e^{-1/p\left(\sum_{i=1}^{n} Xi\right)}$$

We shall see that

$$\Rightarrow \frac{-n}{P} + \frac{1}{P^2}\sum_{i=1}^{n}(X_i) = 0$$

$$\Rightarrow \frac{n}{P} = \frac{\sum X}{P^2}$$

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \quad AMVSU \qquad estimator \qquad (9)$$

3.0 Some simple cases: If in the exponential case again, we are interested only in a sample case, then we have $L(P) = Pe^{-Px}$

$$\frac{dL(P)}{d(P)} = P(e^{-p(x)}(-x)) + e^{-p(x)}(1)$$

$$-xPe^{-P(x)} + e^{-P(x)} = 0$$

$$\Rightarrow \qquad xP\frac{e^{-P(x)}}{e^{-P(x)}} = \frac{e^{-P(x)}}{e^{-P(x)}}$$

xP=1

$$\hat{P} = \frac{1}{x}(AMVSU) \qquad (10)$$

Also if the random variable is discrete, say Poisson, then $P_x(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$

$$L(\lambda)\frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!}\lambda^x e^{-\lambda}$$

$$\frac{d(L(\lambda))}{d(\lambda)} = \frac{1}{x!}\left[\lambda^x e^{-\lambda} + e^{-\lambda}(x\lambda^{x-1})\right] = 0$$

$$\Rightarrow -\lambda^x e^{-\lambda} + xe^{-\lambda}\lambda^{x-\lambda} = 0$$

$$\Rightarrow 1 = \frac{x}{\lambda}, \qquad \hat{\lambda} = \hat{x}(AMVSU) \qquad (11)$$

Also in a case where $X \square B(n, p)$ and $x_1, x_2, \ldots, x_n$ is being considered, we have:

$$P_x(x) = \binom{n}{x} P^x (1-P)^{n-x}$$

$$L(P) = \prod_{i=1}^{n}\binom{n}{x} P^{\sum_{i=1}^{n} x_i}(1-P)^{\sum_{i=1}^{n}(n-x)} = \prod_{i=1}^{n}\binom{n}{x} P^{\sum x_i}(1-P)^{mn-\sum X_i}$$

But if we set $m = n$ and $\sum_{i=1}^{n} x_i = t$, then we shall have

$$\frac{dL(P)}{d(P)} = \left(-P^t(n^2-t)(1-P)^{n^2-t-1} + (1-P)^{n^2-t} tP^{t-1}\right) = 0$$

$$\Rightarrow P^t\left(n^2-t\right)\left(1-P\right)^{n^2-t-1}+\left(1-P\right)^{n^2-t}tP^{t-1}$$

$$\Rightarrow P^t(1-P)^{n^2-t}(1-P)^{-1}(n^2-t)=(1-P)^{n^2-t}tP^tP^{-1}$$

$$\Rightarrow \frac{\left(n^2-t\right)}{1-P}=\frac{t}{P}$$

$$\Rightarrow P\left(n^2-t\right)=t\left(1-P\right)$$

$$\Rightarrow n^2P-Pt-t+Pt=0$$

$$\Rightarrow n^2P=t$$

$$\Rightarrow P=\frac{\sum x_i}{n^2} \text{ since } m=n\text{., then } \frac{\sum x}{n}\cdot\frac{1}{n}$$

$$\Rightarrow \frac{\bar{x}}{n} \qquad (12)$$

Clearly without transformation, we have obtained an asymptotically minimum-variance unbiased estimator which is faster and better so as to obtain more reliable result.

4.0 Problem case: If $x_1, x_2, \ldots x_n$ are values of a random sample from a continuous uniform population with $\alpha=0$ and $\beta=P$ then

$$L(P)=\left(\frac{1}{P}\right)^n=P^{-n}$$

$$\frac{d(L(P))}{d(P)}=P^{-n-1}=0$$

Here, we cannot proceed by using calculus to estimate P for $P\geq$ the largest x and otherwise. Clearly the value of this function will increase as P decreases, P has to be made as small as possible. $\hat{P}=x_n$, the nth order statistics can be applied. There can also be an exceptional case where we can confirm that the 'maximum-likelihood-estimation procedure can be carried out without transformation (logarithmic). It may useful in probability densities and mass functions especially those that contain an exponential component.

5.0 Numerical illustration: If it is believed that the numbers of breakages in a damaged gene y follow a truncated Poisson distribution with probability mass function

$$P(Y=y)=\frac{e^{-\lambda}}{1-e^{-\lambda}}\frac{\lambda^{y}}{y!}, \qquad y=1,2,\ldots$$

The frequency distribution of the number of breakages in a random sample 30 damaged genes as follows

Number of breakage	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
Number of genes	9	5	4	5	0	1	0	2	1	0	1	1	1	30

First is the legitimacy of this truncated Poisson distribution

$$P(Y=y)=\left[\frac{e^{-\lambda}}{1-e^{-\lambda}}\right]\frac{\lambda^{y}}{y!}, \qquad y=1,2,\ldots$$

Then $\sum_{y=1}^{\infty}\left[\frac{e^{-\lambda}}{1-e^{-\lambda}}\right]\frac{\lambda^{y}}{y!}$,

$$\Rightarrow\left[\frac{e^{-\lambda}}{1-e^{-\lambda}}\right]\left[\lambda+\frac{\lambda^{2}}{2!}+\frac{\lambda^{3}}{3!}+\ldots\right]$$

$$\Rightarrow\left[\frac{e^{-\lambda}}{1-e^{-\lambda}}\right]\sum_{y=1}^{\infty}\frac{\lambda^{y}}{y!},$$

$$\Rightarrow\left[\frac{e^{-\lambda}}{1-e^{-\lambda}}\right]\left[\frac{\lambda}{1!}+\frac{\lambda^{2}}{2!}+\frac{\lambda^{3}}{3!}+\ldots\right]$$

$$\Rightarrow\left[\frac{e^{-\lambda}}{1-e^{-\lambda}}\right]\left[e^{-\lambda}-1\right]=\frac{e^{-\lambda+\lambda}-e^{-\lambda}}{1-e^{-\lambda}}$$

$$\Rightarrow\frac{1-e^{-\lambda}}{1-e^{-\lambda}}=1$$

$$L(\lambda)=\frac{e^{-30\lambda}}{\left(1-e^{-\lambda}\right)^{30}}\frac{y\sum x_{i}}{\pi x_{i}!}, \qquad \lambda>0$$

Given that $\sum x_i = 9+10+12+20+0+6+16+9+0+11+12+13=118$

$\ln\ (L(\lambda) = -30\lambda - 30\ln(1-e^{-\lambda}) + 118\ln\lambda - \sum \ln(x_i!)$

$$\frac{d(\ln(\hat{\lambda})}{d(\hat{\lambda})} = -30 - \frac{30}{e^{-\hat{\lambda}} - 1} + \frac{118}{\hat{\lambda}}.$$

With the usual regularity condition, the maximum likelihood estimator $\hat{\lambda}$ satisfies $\frac{d(\ln(\hat{\lambda})}{d(\hat{\lambda})} = 0$

The Newton-Raphson method can be used to find an iterative algorithm for computing the value of $\hat{\lambda}$. We shall obtain the $\frac{d^2(\ln(\hat{\lambda})}{d(\hat{\lambda})}$ as $\frac{d^2(\ln(\hat{\lambda})}{d(\hat{\lambda})} = \frac{30e^{\lambda}}{(e^{\lambda}-1)^2} - \frac{118}{\lambda^2}$. Then if we use an algorithm for finding $\hat{\lambda}$

Numerically as $$\hat{\lambda}_{n+1} = \left[\frac{\lambda_n - \left[\frac{d(\ln(\hat{\lambda}))}{d(\hat{\lambda})}\right]_{\lambda=\lambda_n}}{\left[\frac{d^2(\ln(\hat{\lambda}))}{d(\hat{\lambda}^2)}\right]_{\hat{\lambda}=\lambda_n}} \right]$$

λ_0, initial estimate could be found by plotting in $(L(\lambda))$ against λ. We can also use the non-truncated Poisson, which here would $\lambda_0 = \frac{118}{30} = 3.6$

Carefully let us use λ_0=3.6 as suggested above. If we combine categories 5 to 13 breakages into a single category, we can check the goodness-of-fit of the truncated Poisson distribution.

6.0 Goodness - of - fit

y	1	2	3	4	≥5	Total
Observed	9	5	4	5	7	**30**
Expected	3.030	5.456	6.546	5.892	9.078	**30**

$$\chi^2_{cal} = \frac{(9-3.030)^2}{3.030} + \frac{(5-5.454)^2}{5.454} + \ldots + \frac{(7-9.078)^2}{9.078}$$
$$= 13.4011$$

If k is the number of categories and since we have estimate $\hat{\lambda}$ the maximum likelihood estimator then the degree of freedom is 5-1-1=3. Then $\chi^2_{3,0.95}$ (at 5% level of significance) $\chi^2_{3,0.95}=7.815$. Consider the truncated Poisson

H_0 : *The fit is good*
H_1 : *The fit is not good*

We see this is significant. There is evidence against the null hypothesis of a truncated Poisson distribution.

7.0 Conclusion and recommendation

We have considered both simple and sample cases for different distributions both mass and densities. The asymptotically minimum variance sufficient unbiased estimators (AMVSU) can also be seen as best linear unbiased estimators (BLUE).
We assume convergence, existence and consistency. We have considered a case where calculus cannot even estimate our parameter $\theta = P$ (say). We see that if some set of data is fitted by the truncated Poisson distribution, the goodness-of-fit can be determined, after estimating the parameter $\theta = \lambda$ (say).

References

[1] Harry V.Z (2005).: On the Rate of Convergence of the Maximum Likelihood Estimator in Brownian Semimartingale Models Bernoulli 11 (4), 643-664.

[2] Hoaglin D.C., Mosteller F., and Turkey J.W. (1983).: Understanding Robust and Exploratory Data Analysis. Wiley, New York

[3] Koch G.S. (Jr.), and Link R.F., (1980).: Statistical Analysis of Geological Data Vols. I.II. Dover, New York

[4] Kutoyants, Yu.A. (2004).: Statistical Inference for Ergodic Diffusion Processes. New York: Springer Verlag.

[5] Loukianova D. and Loukianov O. (2003a).: Almost Sure Rate of Convergence of Maximum Likelihood Estimators for Multidimensional Diffusions. Preprint

[6] Loukianova D. and Loukianova O. (2003b).: Uniform Law of Large Numbers and Consistency of Estimators for Multidimensional Harris Diffusions. Preprint

[7] O'Hagan , A. ( ).: Kendall's Advanced Theory of Statistics Vol.2B. Bayesian Inference. Pg. 214.

[8] Thomas G. B. and Finney R. L. (1982).: Calculus and Analytical Geometry. 5th Edition. Addison – Wesley Publishing Company, Reading, UK.

[9] Van de Geer S.A. (1995a).: The Method of Sieves and Minimum Contrast Estimator. Math-Methods. Statist., 4(1), 20-38

[10] Van Zanten J.H (2001).: A Note on Consistent Estimation of Multivariate Parameters in Ergodic Diffusion Models. Scanells J.Statist., 28,617-623

[11] Wong W.H. and Shen X. (1995).: Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES Ann Statist., 23,339-362