Communications in Statistics - Simulation and Computation

Relative Efficiency of Higher Order Norm-Based Estimator Over the Least Squares Estimator

Submission ID	256430339
Article Type	Original Article
Keywords	Least Squares, Higher Order Loss Function, Rel ative Efficiency
Authors	Gopal K Basak, Samarjit Das, Arijit De, Atanu B iswas

Relative Efficiency of Higher Order Norm-Based Estimator Over the Least Squares Estimator

Gopal K Basak* Samarjit Das† Arijit De ‡ and Atanu Biswas §¶

Indian Statistical Institute

Abstract

In this article, we study the performance of the estimator that minimizes $L_{2k}-$ based loss function (for $k \geq 2$) against the estimator which minimizes the L_2- order loss function (or the least squares estimator). Commonly occurring examples illustrate the differences in efficiency between L_{2k} and L_2- based estimators. We derive an empirically testable condition under which the L_{2k} estimator is more efficient than the least squares estimator. We construct a simple decision rule to choose between L_{2k} and L_2 estimator. Special emphasis is provided to study L_4 based estimator. A detailed simulation study verifies the effectiveness of this decision rule. Also, the superiority of the L_4 based estimator is demonstrated in a real life data set. An artificially constructed example that mimics many real life scenarios shows the superiority of L_4 over L_2.

Key Words: Least Squares, Higher Order Loss Function.

JEL Classification: C01, C13

*Stat-Math Unit, Indian Statistical Institute, Kolkata, India, email: gkb@isical.ac.in

†Corresponding author: Economic Research Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata-700108, India, Tel: +91-33-2575-2627, email: samarjit@isical.ac.in

‡Former Masters student of Indian Statistical Institute

§Applied Statistics Unit, Indian Statistical Institute, Kolkata, India, email: atanu@isical.ac.in

¶We are thankful to Probal Choudhury, Jyoti Sarkar for their constructive inputs; and also Sattwik Santra for providing us the data.

1 Introduction

The genesis of this article comes from a simple question from a student. Recently, one of the authors gave a course on 'Introductory Econometrics' and tried to introduce the idea of least squares. The teacher followed the path he had inherited from his former mentors as an undergraduate student. One beautiful mind from the audience asked, "Why are we minimizing the residual sum of squares? Why not the higher order loss function"? The teacher talked about the impact of outliers, etc. Then the next question from the student: "Let us assume that there are no outliers; differentiability of the loss function is assumed; is there any relative advantage of using a higher order-based loss function compared to least squares?" It was a bad day for the teacher. The teacher tried to find an answer from Google (assuming that Google knows everything!) and from a few learned colleagues. Surprisingly, the teacher found that many researchers asked similar questions on Google.

The least squares (LS hereafter) method is possibly the most popular method of estimation routinely used to estimate the underlying (regression) parameters. Stigler (1981) rightly said: "The method of least squares is the automobile of modern statistical analysis: Despite its limitations, occasional accidents, and incidental pollution, it and its numerous variations, extensions, and related conveyances carry the bulk of statistical analysis, and are known and valued by nearly all". Such an overwhelming popularity of the LS may be due to its simplicity, optimal properties and robustness to any distributional assumption. Moreover, it leads to the best (minimum variance) estimator under normality. Laplace used the name "most advantageous method". However, it appears to us that such an irresistible popularity of the LS may have impeded the exploration of other *smooth* loss functions. Comparative computational difficulty might be another reason that such exploration was not favoured by pioneers such as Gauss, Laplace and others. Whereas, a large literature to incorporate *non-smooth* loss functions in order to address (outlier) robustness have been developed. Unfortunately and surprisingly, whole statistical literature is somewhat mute on the possible use of *smooth* higher order loss functions. Therefore, it is a pertinent question to ask: Are there any relative advantages in using higher order smooth loss function compared to the omnipresent least squares? Our aim here is to study an appropriate higher order estimator and compare its efficiency against the LS. In the regression set up, we find a significantly large and useful class of error distributions for which a higher order loss function is more efficient than the LS. In this paper, we give an empirically testable condition under which a higher order smooth loss functions lead to a more efficient estimator than the LS. We also provide a simple but pragmatic decision rule to make a choice between L_{2k} and L_2; $k = 2$ is emphasized. A detailed simulation study shows the effectiveness of such a decision rule. For simulation study L_4 is considered.

Studying the efficacy of higher order normed based estimator is important on its own right; not necessarily in comparison to least square. It opens up the possibility to consider a convex combination of loss functions of various degrees. Arthanari and Dodge (1981) considers convex combination of L_1 and L_2 norms; and studies its properties. Convex combination of $L_1, \ldots, L_p$ may lead to more useful estimator; and the resultant estimator is expected to be robust to any distributional assumption. Earlier also, such an use of higher order loss functions is attempted. Turner (1960) heuristically touch upon the possible use of a higher order loss function in the context of estimation of the location parameter. He discusses several kinds of general PDFs; and advices in the case of the double exponential, to minimize the sum of the absolute deviations; in the case of the normal, to minimize the sum of the squared deviations (least squares); and in the case of the q-th power distribution, to minimize the sum of the q-th power of the deviations (least q-th's). Attempts are also made to define a general class of likelihood to derive a robust parameter estimates, robust to distributional assumption. For example, Zeckhauser and Thompson (1970) defines a general class of distribution; and empirically found its suitability.

In Section 2, we describe the very basic model and develop the methodology needed to compare the efficiency the proposed loss functions. Various classes of error distributions are considered in Section 3. Here we attempt to examine the performance of L_{2k} if the model error distribution deviates from normality. Here we provide a very general classes of parametric distributions on finite support to illustrate the enormous scope of applicability of the L_4- based loss functions. In Section 4, we provide a decision rule along with its asymptotic properties. Section 5 summarizes the results of simulation study on mixture distributions. Section 6 gives an application to a real life data. In this section, we study an artificially constructed example that mimics many real life scenarios shows. This constructed example also shows the superiority of L_4 over L_2. Section 7 ends with some concluding remarks and identifies possible future directions of research.

2 Model and Methodology

Consider a linear regression set up

$$Y = X\theta + \varepsilon, \tag{2.1}$$

where Y is an $n \times 1$ vector of observations, X is an $n \times k$ design matrix and F is the cumulative distribution function (cdf) corresponding to the error vector ε. We assume the usual Guass-Markov set-up [1].

Now consider the following two estimators as

$$\widehat{\theta}_{L_2} = \widehat{\theta}_{ols} = argmin \sum_{i=1}^{n} \varepsilon_i^2;$$

and

$$\widehat{\theta}_{L_4} = argmin \sum_{i=1}^{n} \varepsilon_i^4.$$

Our objective is to compare $\widehat{\theta}_{OLS}$ and $\widehat{\theta}_{L_4}$, and find conditions under which the latter performs better than the former. Both of these estimators are M-estimators. So they possess the properties such as consistency and asymptotic normality under some standard conditions (see, for example; Stefanski and Boos (2002)).

For the OLS estimator,

$$\sqrt{n}\left(\widehat{\theta}_{OLS} - \theta_0\right) \xrightarrow{d} N\left(0, \sigma^2 S^{-1}\right), \tag{2.2}$$

where

$S = X'X$ where

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}.$$

[1]We borrow the theory of m-estimation. It needs many more conditions along with moment conditions. We will not elaborate all these assumptions for brevity. Consistency and asymptotic normality are assumed here.

We exhibit that $\widehat{\theta}_{L_4}$ satisfies the following result.

Lemma 1.

$$\sqrt{n}\left(\widehat{\theta}_{L_4}-\theta_0\right)\xrightarrow{d} N\left(0,\frac{\mu_6-\mu_3^2}{9\mu_2^3}S^{-1}\right). \tag{2.3}$$

Proof: For the L_4 estimator, using the M-estimator property, we have

$$\sqrt{n}\left(\widehat{\theta}_{L_4}-\theta_0\right)\xrightarrow{d} N\left(0,V(\theta_0)\right),$$

where

$$V(\theta_0)=A(\theta_0)^{-1}B(\theta_0)[A(\theta_0)^{-1}]',$$

$$\begin{aligned} B(\theta_0) &= E[\psi(y,\theta_0)\psi(y,\theta_0)']-E[\psi(y,\theta_0)]E[\psi(y,\theta_0)]', \\ A(\theta_0) &= E\left(\frac{\delta}{\delta\theta}\psi(y,\theta_0)\right), \end{aligned}$$

and

$$\psi=\frac{\delta S_4}{\delta\theta};$$

where

$$S_4=\sum_{i=1}^{n}\varepsilon_i^4.$$

Let μ_k denote the kth order central moment corresponding to the distribution of ε.

Consequently, we get

$$E[\psi(y,\theta_0)]E[\psi(y,\theta_0)]'=16\mu_3^2R$$

where

$$R=\begin{pmatrix} n^2 & n\sum_{i=1}^n x_{1i} & n\sum_{i=1}^n x_{2i} & \cdots & n\sum_{i=1}^n x_{ki} \\ n\sum_{i=1}^n x_{1i} & (\sum_{i=1}^n x_{1i})^2 & (\sum_{i=1}^n x_{1i})(\sum_{i=1}^n x_{2i}) & \cdots & (\sum_{i\neq j} x_{1i})(\sum_{i=1}^n x_{ki}) \\ n\sum_{i=1}^n x_{2i} & (\sum_{i=1}^n x_{2i})(\sum_{i=1}^n x_{1i})x_{1j} & (\sum_{i=1}^n x_{2i})^2 & \cdots & (\sum_{i=1}^n x_{2i})(\sum_{i=1}^n x_{2i}x_{ki}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n\sum_{i=1}^n x_{ki} & (\sum_{i=1}^n x_{ki})(\sum_{i=1}^n x_{1i}) & (\sum_{i=1}^n x_{ki})(\sum_{i=1}^n x_{2i}) & \cdots & (\sum_{i=1}^n x_{ki})^2 \end{pmatrix}.$$

Also, $E[\psi(y,\theta_0)\psi(y,\theta_0)']=16(\mu_6S+\mu_3^2Q)$, where

$$Q = \begin{pmatrix} n(n-1) & (n-1)\sum_{i=1}^{n} x_{1i} & (n-1)\sum_{i=1}^{n} x_{2i} & \cdots & n\sum_{i=1}^{n} x_{ki} \\ (n-1)\sum_{i=1}^{n} x_{1i} & \sum_{i\neq j} x_{1i}x_{1j} & \sum_{i\neq j} x_{1i}x_{2j} & \cdots & \sum_{i\neq j} x_{1i}x_{kj} \\ (n-1)\sum_{i=1}^{n} x_{2i} & \sum_{i\neq j} x_{2i}x_{1j} & \sum_{i\neq j} x_{2i}x_{2j} & \cdots & \sum_{i\neq j} x_{2i}x_{kj} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (n-1)\sum_{i=1}^{n} x_{ki} & \sum_{i\neq j} x_{ki}x_{1j} & \sum_{i\neq j} x_{ki}x_{2j} & \cdots & \sum_{i\neq j} x_{ki}x_{kj} \end{pmatrix},$$

Thus, $B(\theta_0) = 16\mu_6 S + 16\mu_3^2(Q - R)$. Since $Q - R = -S$

$$B(\theta_0) = 16S(\mu_6 - \mu_3^2) = 16SVar(\varepsilon^3).$$

Similarly, one can simplify

$$A(\theta_0) = 12\mu_2.S$$

Then

$$V(\theta_0) = A(\theta_0)^{-1}B(\theta_0)[A(\theta_0)^{-1}]' = \frac{1}{12\mu_2}S^{-1}(16S(\mu_6 - \mu_3^2)))\frac{1}{12\mu_2}S^{-1} = \frac{\mu_6 - \mu_3^2}{9\mu_2^2}S^{-1}.$$

Hence the proof. ▮

Theorem 1. The L_4 estimator performs better than the OLS estimator in terms of precision iff

$$\frac{\mu_6 - \mu_3^2}{\mu_2^3} < 9 \tag{2.4}$$

Proof. Proof follows by comparing 2.2 and 2.3. ▮

Note: For symmetric distribution of ε, or whenever $\mu_3 = 0$, the criterion will be

$$\frac{\mu_6}{\mu_2^3} < 9. \tag{2.5}$$

It may be interesting to examine the performance of L_{2k} relative to that of L_2. The following corollary is presented to this end.

Corollary 1. L_{2k} estimator better than LS (i.e., L_2) iff

$$\frac{Var(\varepsilon^{2k-1})}{(2k-1)^2(Var(\varepsilon))^{2k-1}} < 1.$$

Proof. The proof is analogous to that of Theorem 1. ▮

3 OLS versus L_4 for some selected distributions

In this section, we consider few important parametric error distributions to illustrate the vast scope of applicability of L_4 based loss function. The list of distributions considered is no way exhaustive, but certainly shows the immense opportunity of applications in diverse areas. Now, we check if the aforementioned condition (2.4) holds for different distributions of ε.

3.0.1 U-Shaped Distribution

Consider a simple U-shaped distribution:

$$f(x) = dx^{2k}; \quad -c \leq x \leq c, \text{ where } k \text{ is a positive integer.}$$

Note that $d = \frac{2k+1}{2c^{2k+1}}$. It is easy to calculate

$$\frac{\mu_6}{9\mu_2^3} = \frac{(2k+3)^3}{9(2k+1)^2(2k+7)} < 1.$$

Note that for $k = 1$, $\frac{\mu_6}{9\mu_2^3} = \frac{125}{729}$, and in the limit, $\frac{\mu_6}{9\mu_2^3} = \frac{1}{9}$.

A U-shaped distribution has two modes; and can be looked upon as a mixture of two (J-shaped) distributions - a mixture of an extreme positively skewed and another extreme negatively skewed distributions. One popular applied example of a U-shaped distribution is the number of deaths at various ages. Several more examples can be found in B. S. Everitt (2005).

3.0.2 Uniform$(-a, a)$

This is a symmetric distribution. Here

$$\mu_r = \frac{a^{r+1} - (-a)^{r+1}}{(r+1)\{a - (-a)\}},$$

and consequently

$$\frac{\mu_6}{9\mu_2^3} = \frac{3}{7} < 1.$$

Hence, L_4 estimator is better than the OLS estimator, when the error component has uniform distribution.

3.0.3 Normal(μ, σ^2)

This is again a symmetric distribution where

$$\mu_{2r} = \sigma^{2r}(2r-1) \times (2r-3) \times \cdots \times 5 \times 3 \times 1,$$

and hence

$$\frac{\mu_6}{9\mu_2^3} = \frac{15}{9} > 1.$$

Hence, for normally distributed errors, the OLS estimator is always preferred over the L_4 estimator.

3.0.4 Laplace(λ)

Here we have

$$\mu_r = \lambda^r \Gamma(r+1),$$

if r is even, and, consequently,

$$\frac{\mu_6}{9\mu_2^3} = \frac{6!}{72} > 1.$$

Hence, when ε follows Laplace distribution, the OLS estimator is preferred over the L_4 estimator.

3.0.5 Beta(a, b)

The beta distribution is a family of continuous probability distributions defined on the interval $[0, 1]$ parametrized by two positive shape parameters, denoted by a and b, that appear as exponents of the random variable and control the shape of the distribution. This class of distributions include a variety of symmetric, bell-shaped, positively skewed, negatively skewed, uniform, and 'U-shaped' distributions. The general form of the central moments of the beta distribution are quite complicated. So we will start with the raw moments and obtain the forms of μ_6, μ_3 and μ_2. Note that, here

$$\mu_r' = \prod_{i=0}^{r-1} \left\{ \frac{a+i}{a+b+i} \right\}.$$

The figure 4 provides a huge range of parameters for which L_4 is better than L_2.

3.0.6 Gaussian mixture distribution

Suppose that $F(x) = \frac{1}{2}N(\xi_1, \sigma^2) + \frac{1}{2}N(\xi_2, \sigma^2)$. We assume, for simplicity, common σ^2 for both the components. Here

$$\mu_r = \frac{1}{2}\sum_{i=0}^{r}\binom{r}{i}(\xi_1 - \xi)^{r-i}\mu_{1i} + \frac{1}{2}\sum_{i=0}^{r}\binom{r}{i}(\xi_2 - \xi)^{r-i}\mu_{2i},$$

where $\xi = \frac{1}{2}(\xi_1 + \xi_2)$ and μ_{ji} is the ith central moment of $N(\xi_j, \sigma^2)$ distribution, $j = 1, 2$. Let $\xi_2 - \xi = \frac{1}{2}(\xi_2 - \xi_1) = -(\xi_1 - \xi) = a_0$. Then (2.5) reduces to

$$6 + 18c^2 - 12c^4 - 8c^6 < 0, \tag{3.6}$$

where $c = \frac{\xi_2 - \xi_1}{2\sigma}$. The left hand side expression of (3.6), which is a function of c, is plotted in Figure 1. From the plot we can see for $|c| > 1.058$, this function assumes values less than zero and then it decreases rapidly. This means that the condition (2.5) of superiority of L_4 estimators will be satisfied if the means of the two component of the mixture distribution are more than 1.058σ distance. A mixture of more than two Gaussian distributions will behave similarly with respect to this condition. Also the case of unequal σ^2 can be tackled similarly.

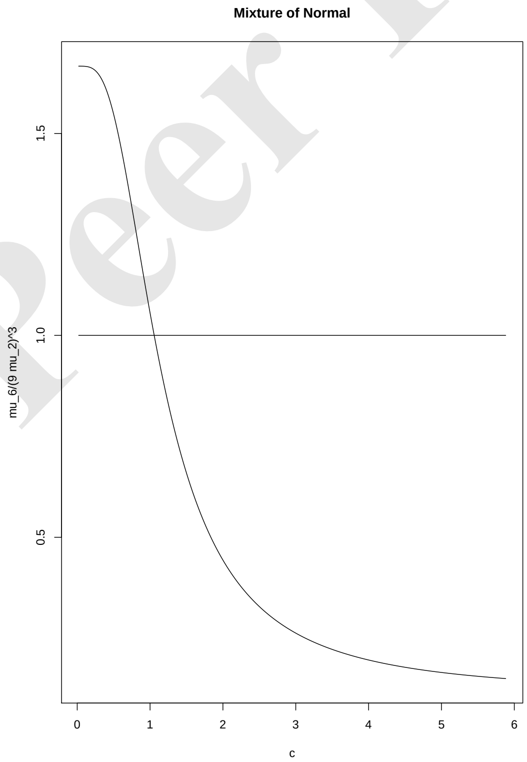

Figure 1: line plot as a function of c

3.0.7 Truncated Normal distribution

Consider (for simplicity) the both side truncated standard normal distribution. The even order moments are

$$\mu_{2k}^c = \frac{2}{d}\int_0^c \frac{x^{2k} exp(-\frac{x^2}{2})}{\sqrt{2\pi}}dx.$$

Define $\Delta = \sqrt{2\pi}(\Phi(c) - 0.5)$. Therefore,

$$\mu_{2k}^c = \frac{c^{2k-1} exp(-\frac{c^2}{2})}{\Delta} + \mu_{2k-2}^c(2k-1).$$

Now it is easy to calculate $\mu_6^c = 15 - \frac{c^5 exp(-\frac{c^2}{2})}{\Delta} - \frac{5c^3 exp(-\frac{c^2}{2})}{\Delta} - \frac{15c\, exp(-\frac{c^2}{2})}{\Delta}$, and $\mu_4^c = 3 - \frac{c^3 exp(-\frac{c^2}{2})}{\Delta} - \frac{3c\, exp(-\frac{c^2}{2})}{\Delta}$, and $\mu_2^c = 1 - \frac{c\, exp(-\frac{c^2}{2})}{\Delta}$. Now one can see that as long as $c \in (0, 2.33)$, L_4 performs better than L_2. One implication of this result is that 97 percent times L_4 performs better than L_2.

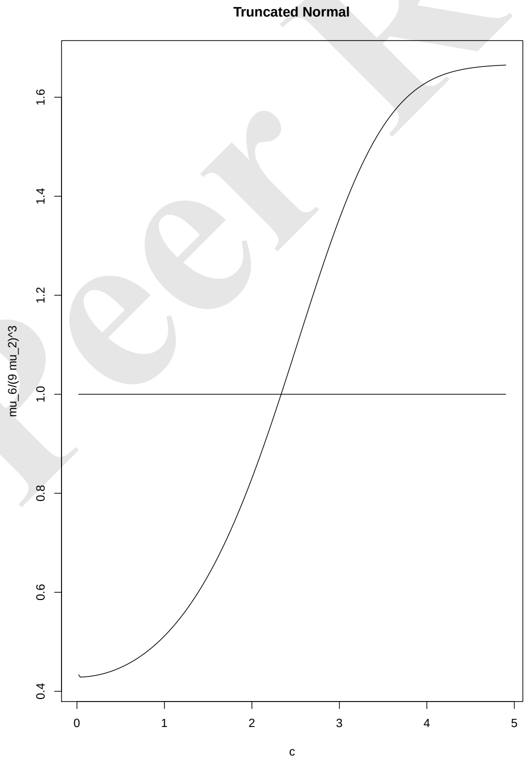

Figure 2: Truncated Normal plot

3.0.8 Raised cosine distribution

If X follows a raised cosine distribution with parameters a and b, denoted by $X \sim COR(a,b)$, then the probability density function (pdf) is given by

$$f(x) = \frac{1}{2b}\left[1+\cos\left(\pi\frac{x-a}{b}\right)\right], \quad a-b \leq x \leq a+b,\ a \in \mathbb{R},\ b>0.$$

The form of this distribution resembles that of a normal distribution except for the fact that it has finite tails. Suppose it can be assumed that the value of systematic errors lies in some known interval; and manufacturer has aimed to make device as accurate as possible. In such circumstances, Raised Cosine distribution may be appropriate. Another popular application is in circular data. See Rinne (2010, pp. 116). Other properties like the cdf, moment generating function (mgf), characteristic functions, raw moments up to order 4, and the kurtosis are available in Rinne (2010, pp. 116-118). It is observed that the distribution has a kurtosis of 2.1938, less than that of normal distribution. It has a thin tail. Here, using the mgf, we have

$$\begin{aligned}\mu_6 &= \frac{b^6\left(\pi^6-42\pi^4+840\pi^2-5040\right)}{7\pi^6},\\ \mu^2 &= \frac{b^2\left(\pi^2-6\right)}{3\pi^2},\end{aligned}$$

and hence

$$\mu_6/(9\mu_2^3) = \frac{3}{7} - \frac{72\pi^4-2196\pi^2+14472}{7(\pi^2-6)^3} = 0.8926 < 1.$$

Hence, for this distribution, (2.5) is satisfied, and consequently L_4 is preferred for parameter estimation.

3.1 A Sub-Gaussian family of distributions

Sub-Gaussian family of distributions is a well-studied family of distribution whose tail is dominated by the normal distribution. As we observed that the L_4 estimators are preferred for a distribution for which the tail is thinner than that of a normal distribution, here we discuss about a relatively uncommon distribution and validity of the condition with respect to this distribution. Consider a distribution with pdf of the form

$$f(x) = c\exp(-x^{2k}),$$

where k is an integer and c is the normalizing constant, which gives

$$c = \frac{k}{\Gamma\left(\frac{1}{2k}\right)}.$$

For various values of k, the pdf of the distribution is drawn in Figure 3. Note that $k = 1$ provides the normal curve. As k becomes larger and larger, tail of the distribution tend to collapse. For extremely large k, the distribution resembles a symmetric curve in a finite support. It is interesting to consider the peaks of all drawn curves. The first plot (the density plot) of this panel shows that L_4 performs better than L_2 for all those curves for which peaks are below the red curve. Here it may be mentioned that the red curve is drawn for $k = 1.45$. For the second plot of the panel, various values of k are given in the $x-$axis; and values of the test statistic are given in the $y-$axis. The parallel line, parallel to $x-$axis, shows the cut-off point, which is 1 . The second plot of the panel shows that when the value of k is greater than 1.45, L_4 performs better than L_2.

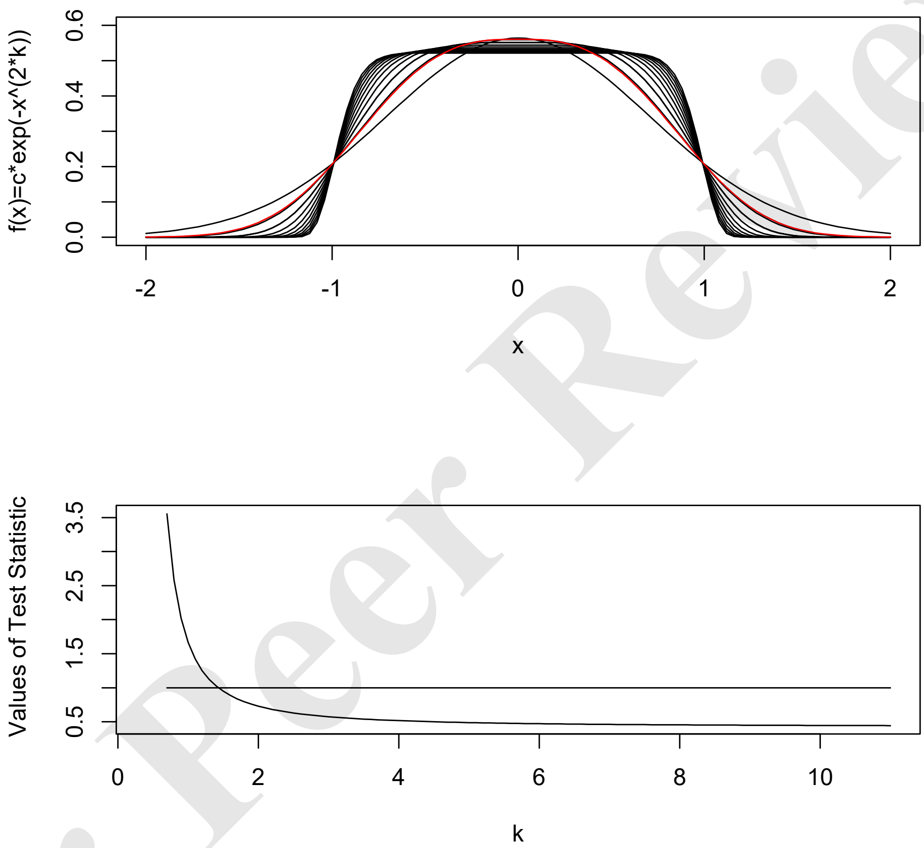

Figure 3: The shape of the various Sub-Gaussian distributions ($f(x) = c\exp(-x^{2k})$ for a range of values k where L_4 is better than L_2. The red curve is for $k = 1.45$.

Here $\mu_r = 0$ if r is odd. When r is even, we have

$$\mu_r = c\int_{-\infty}^{\infty} x^r \exp(-x^{2k}) = \frac{c}{k}\Gamma\left(\frac{r+1}{k}\right).$$

We immediately get

$$\frac{\mu_6}{9\mu_2^3} = \frac{\left(\Gamma\left(\frac{1}{2k}\right)\right)^2 \Gamma\left(\frac{7}{2k}\right)}{9\left(\Gamma\left(\frac{3}{2k}\right)\right)^3}.$$

The values of the test statistic against various values of K is drawn in the bottom part in Figure 3. We observe that for $k \geqslant 2$, $\frac{\mu_6}{9\mu_2^3}$ assumes values less than 1.

Now, according to Zeckhauser and Thompson (1970); Turner (1960) and Box and Tiao (1962), for a distribution with pdf

$$f(u) = k(\sigma, m)\exp\left(-\left|\frac{u}{\sigma}\right|^m\right), \quad \sigma > 0, \; m > 0,$$

where $k(\sigma, m) = \left[2\sigma\Gamma\left(1+\frac{1}{m}\right)\right]^{-1}$, L_m estimators dominate all $L_{m'}$ estimators where $m' < m$. Note that, for $m = 4$ and $\sigma = 1$, we obtain the distribution displayed in Figure 3. So L_4 estimator performs better than the corresponding L_2 estimator, that is the OLS estimator, which is entirely in agreement to what we derived above.

It may be noted that for the normal distribution $m = 2$; $m = 1$ gives the double exponential distribution; where m tends to ∞, the distribution tends to the rectangular.

The article of Zeckhauser and Thompson (1970) considers four empirical examples to find that there is a sizable gains in likelihood if m is estimated rather than pre-specified equal to 2. All of the evidence they found leads them to the conclusion that if accurate estimation of a linear regression line is important, it will usually be desirable to estimate not only the coefficients of the regression line, but also the parameters of the power distribution that generated the errors about the regression line. The effect on the estimates of regression coefficients may not be small.

3.2 A Pearsonian Family

In this section we consider a very general class of parametric distributions on finite support (this assumption is made to ease plot drawing) to illustrate the enormous scope of applicability of L_4 based loss function. Consider the class of distribution:

$$f = d(1+x^2)^a, \; a \in R \; d > 0, \; |x| \leq 1, \; .$$

Here d depends on a to make the f a density. The first plot of the panel depicts the shape of the density for different values of a. Depending on the value of a, this class of distributions includes various 'U-shaped' (for $a > 0$) and 'Bell-Shaped' (for $a < 0$) distributions. It is interesting to consider the peaks of all drawn curves. The first plot (the density plot) of this panel shows that L_4 performs better than L_2 for all those curves for which peaks are below the red curve. Here it may be mentioned that the red curve is drawn for $a = -3.2$. For the second plot of the panel, various values of a are given in the $x-$axis; and values of the test statistic are given in the $y-$axis. The second plot of the panel shows that when the value of a is greater than -3.2, L_4 performs better than L_2, in all shape going from low deep to hump till it reaches certain level.

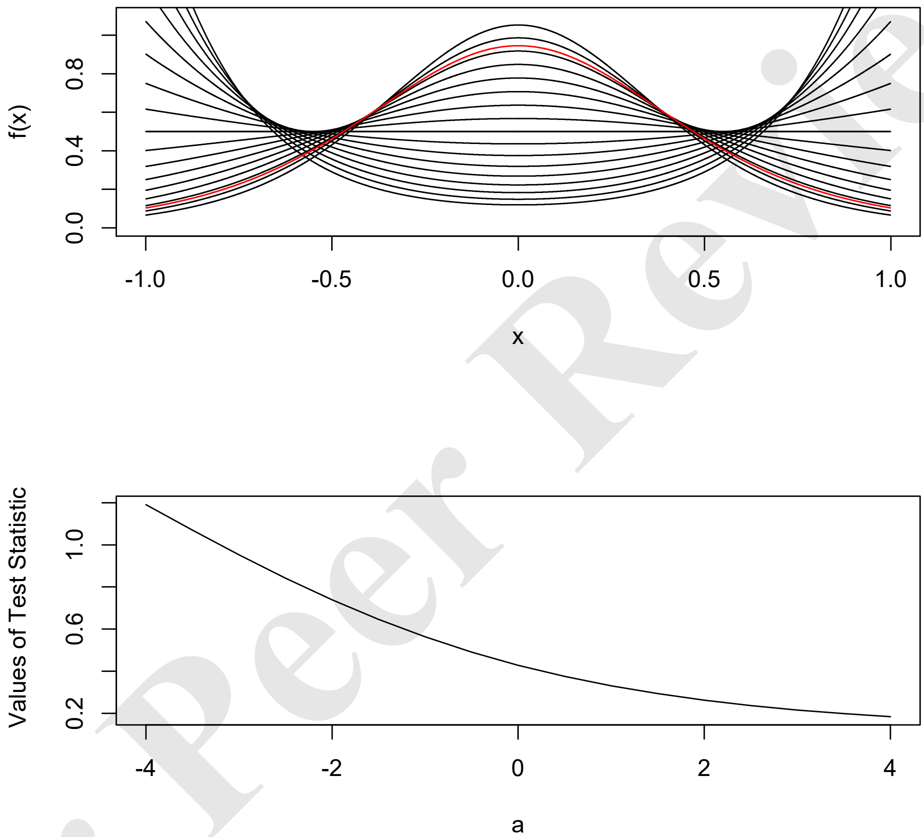

Figure 4: The shape of the distributions ($f = d(1+x^2)^a,\ a \in R,\ d > 0,\ |x| \leq 1$) for a range of values a where L_4 is better than L_2.

Plots of another parametric family of distributions (belongs to the Pearsonian family, Type II), given by

$$f = d(1-x^2)^a, \; a > -1 \; d > 0, \quad \text{function of } a, \; |x| < 1$$

are shown below. It may be noted that this particular distribution is linked to *Beta* distribution as well. To see this, let $Y \sim Beta(\alpha_1, \alpha_2), Y \in (0,1)$. Let $X = a + (b-a)Y$. Then

$$f(x) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{(x-a)^{\alpha_1-1}(b-x)^{\alpha_2-1}}{(b-a)^{\alpha_1+\alpha_2-1}}.$$

To see the equivalence, set $a = -b = -1$, and $\alpha_1 = \alpha_2 = \alpha + 1$.

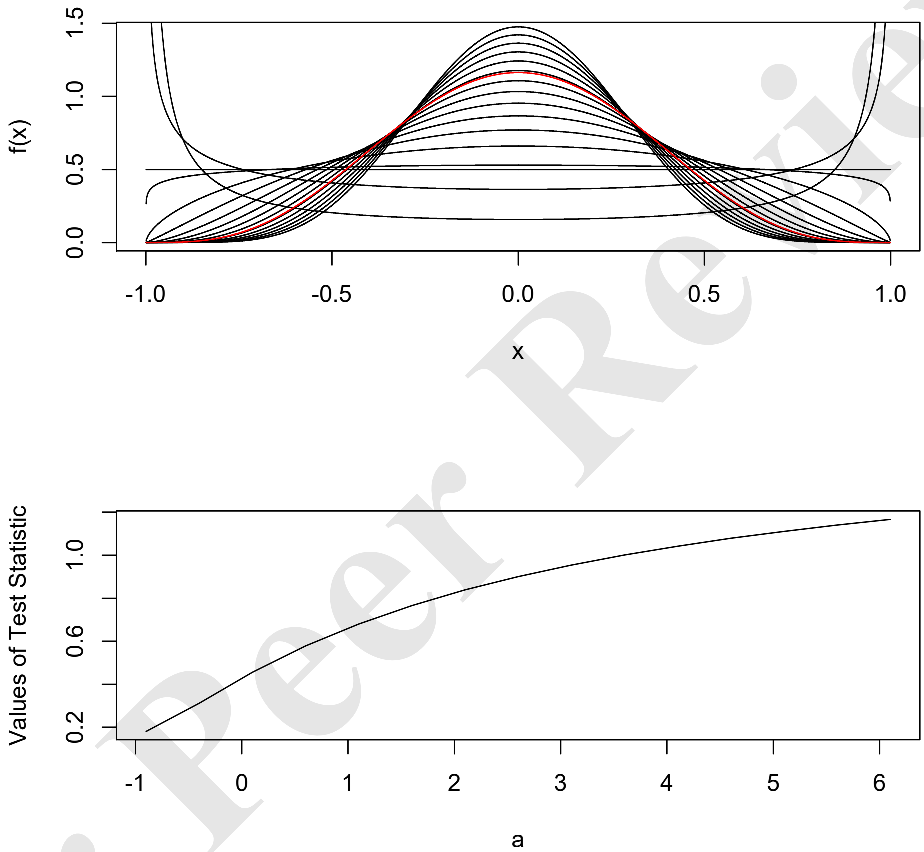

Figure 5: The shape of the distributions ($f = d(1-x^2)^a,\ a > -1,\ d > 0,\ |x| \leq 1$) for a range of values a where L_4 is better than L_2.

Figure 5 also depicts the same feature as in Figure 4. Depending on the value of a, this group of parametric of distributions depicts various 'U-shaped' and 'Bell-Shaped'. It is interesting to consider the peaks of all drawn curves. The first plot (the density plot) of this panel shows that L_4 performs better than L_2 for all those curves for which peaks are below the red curve. The red curve is drawn for $a = 3.5$. The second plot of the panel shows that when the value of a is greater than 3.5, L_4 performs better than L_2, in all shape going from low deep to hump till it reaches certain level.

These two distributions illustrate the enormous possibility of use of L_4 based loss function. Future studies will investigate whether this is a general phenomenon for other Pearsonian family of errors distributions.

4 Decision rule: OLS versus L_4

In this Section we derive a decision rule based on the criterion from Section 2 to decide whether OLS or L_4 estimator is preferred for some data.

Lemma 2. Suppose X follows a distribution for which μ_r exists for all r. Then

$$\sqrt{n}\widehat{\mu}_r = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \overline{x})^r = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu)^r - r\mu_{r-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu) + o_p(1).$$

Proof: Observe that

$$\begin{aligned}\sqrt{n}\widehat{\mu}_r &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \overline{x})^r = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu + \mu - \overline{x})^r \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu)^r + r\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^{r-1}\sqrt{n}(\mu - \overline{x}) \\ &\quad + \binom{r}{2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^{r-2}\sqrt{n}(\mu - \overline{x})^2 + \cdots . \end{aligned} \tag{4.7}$$

Now all the terms other than the first and second term of (4.7) are of the order $o_p(1)$ because $\sqrt{n}(\mu - \overline{x})$ is $O_p(1)$, and hence $\sqrt{n}(\mu - \overline{x})^k$, $k \geq 2$, is $o_p(1)$. Also $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^{r-1} = \mu_{r-1} + o_p(1)$. Therefore

$$\sqrt{n}\widehat{\mu}_r = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \overline{x})^r = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu)^r - r\mu_{r-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu) + o_p(1).$$

□

Furthermore, by delta method,

$$\sqrt{n}\left((\widehat{\sigma}^2)^{\frac{r}{2}} - (\sigma^2)^{\frac{r}{2}}\right) = \frac{r}{2}(\sigma^2)^{\frac{r}{2}-1}\sqrt{n}(\widehat{\sigma}^2 - \sigma^2) + o_p(1).$$

Then, we have the following Theorem.

Theorem 2. Let $v = \frac{\mu_6 - \mu_3^2}{\sigma^6}$. Suppose μ_{12} exists for distribution of X. Then

$$\sqrt{n}(\widehat{v} - v) = \frac{\alpha_0}{\widehat{\sigma}^6}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i + o_p(1),$$

where $\alpha_0 = (1, \quad -(6\mu_5 - 3\mu_2\mu_3), \quad -\mu_3, \quad -3\sigma^4 v)$ and $Z_i = \begin{pmatrix} (x_i-\mu)^6 - \mu_6 \\ (x_i - \mu) \\ (x_i-\mu)^3 - \mu_3 \\ (x_i-\mu)^2 - \sigma^2 \end{pmatrix}$.

Proof of this theorem is given in the Appendix B. ▮

Given the form of v, we can now develop a test rule or a decision rule to examine if one should use L_2 based loss function or L_4. The **Theorem 1** sates that the L_4 estimator performs better than the OLS estimator in terms of precision iff

$$v = \frac{\mu_6 - \mu_3^2}{\mu_2^3} < 9 \tag{4.8}$$

Therefore we can write down our hypotheses as

$$H_0 : v \geq 9 \qquad \text{versus} \qquad H_1 : v < 9.$$

Under H_0; L_2 is preferred.
Here it may be noted that the null hypothesis is a composite hypothesis, to tackle this problem, we use 0/1 loss function and calculate the risk. Zero-one loss is the simplest loss function which literally counts how many mistakes an hypothesis function makes on a data set. For every single example it suffers a loss of 1 if it is mispredicted, and 0 otherwise.
Hence our decision theoretic problem reduces to

$$A_0 : v \geq 9 \qquad \text{versus} \qquad A_1 : v < 9.$$

We use the statistic

$$T = \frac{\sqrt{n}(\widehat{v} - 9)}{s},$$

where

$$s^2 = \widehat{V(v)}.$$

Corollary 2. T is asymptotically $N(0,1)$, under A_0. ∎

Proof. Proof follows from the proof of **Theorem 2**, See Appendix B. ∎

5 Simulation study

We carry out the decision making procedure under 0-1 loss function and calculate the risk function, which is the expected loss. Here we generate data from three types of distribution, one for which L_4 is always better than OLS estimator, one where OLS is better than L_4, and the third one is near the boundary. The values of the calculated risk are given in Table 1.

Table 1 given in Appendix A should be here

Simulation study is based on 10000 iterations; and with sample sizes of 100, 200, 500, 1000, 2000, 5000.

The first panel of Table 1 is based on mixture of two T distributions with 6 degrees of freedom (DF) each. Mean of each components are set at $(5,-5),(4,-4),(3,-3),(2,-2)$. Here it may be mentioned that our test needs existence of 6th order moments. To this end, we need t distribution with at least 7 df. DF 6 is considered to examine the performance of our decision rule even when moments do not exist. Mixture coefficients are taken from $U(0,1)$ distribution. From this part of the table, it is clear that decision is more certain as sample size increases; more importantly, it is so when the distance between the two components are more.

The second panel of the Table is based on mixture of two T distribution with 10 df each. Here findings corroborate with the first panel. The third panel is based on mixture of two T distribution with 20 df each.

The 4th panel of the table is based on mixture of two asymmetric *Beta* distributions. Mixture coefficients are taken from $U(0,1)$ distribution. The fifth panel is based on two symmetric beta distributions with weight from $U(0,1)$. The first column, in Panel 5 needs special attention. The parameter combination $((4,4;4,4))$ is chosen such that it is in the neighborhood of the boundary the test statistic. Here it shows that test does not favour (for the large sample case, n=5000) any one, as expected. Here the risk is near 50% .

The 6th panel of the Table is based on mixture of two normal distributions. Here also findings are on the expected line. As sample size increases, test correctly discriminates between L_4 and L_2.

6 Empirical Illustration

In this section, we provide two illustrations. One is based on a constructed data set which resembles many real life scenarios; and the second one is based on a real life data set.

6.1 Constructed Example

Data often contains rounding errors. Variables (like heights or weights, age in years, or birth weight in ounces.) that by their very nature are continuous are, nevertheless, typically measured in a discrete manner. People feel more comfortable to report their age as mid forty, mid fifty and so on. They are rounded to a certain level of accuracy, often to some preassigned decimal point of a measuring scale (e.g., to multiples of 10 cm, 1 cm, or 0.1 cm) or simply our preference of some numbers over other numbers. The reason may be the avoidance of costs associated with a fine measurement or the imprecise nature of the measuring instrument. The German military, for example, measures the height of recruits to the nearest 1 cm. Even if precise measurements are available, they are sometimes recorded in a coarsened way in order to preserve confidentiality or to compress the data into an easy to grasp frequency table.

Here we consider the linear regression where the dependent variable is rounded to nearest integer; independent variables are free of any such errors. The true but unobserved dependent variable (y_{st}) is generated as

$$y_{st} = 8 + 1 \times x_1 + 2 \times x_2 \quad \text{where } x_1 = 1.3 \times sample.int(10); x_2 = 2.32 * sample(10:18)\ [2].$$

[2]sample.int(10) randomly arranged 1 to 10 integers; sample(10:18) randomly arranged 10 to 18 integers; floor $(y_{st}/5)$ is the largest integer less than or equal to y_{st}.

However, assume that we do not observe Y_{st} but observe

$$Y = 5 \times floor(y_{st}/5).$$

Now we are regressing Y on X_1 and X_2. For this example, we set a moderate sample size of 40. We consider 5000 replications.The output is summarized as follows:

Table 2: Average Estimates Based on the Constructed Data.

	Parameters		
Estimates	Intercept=5.5	Slope 1 =1	Slope 2=2
Average (L_2)	7.013	1.115	1.924
Average (L_4)	6.548	1.021	1.962

It is observed that 90 percent times L_4 is preferred over L_2 based on our proposed decision rule.

After estimation of the model, it may be of interest to know which set of estimators provides the best fit. In the present context it is a tricky problem to find an appropriate 'goodness of fit' measure. Likelihood based methods are not tenable. Similarly, residual sum of square or R^2 are not useful to compare the performances of these two set of parameter estimates. The RSS or R^2 provides advantage to the least square estimators as the least squares estimator is derived by minimizing the RSS. Here we suggest to apply the idea of Pseudo R^2 (See Cameron and Trivedi, 2005 for details, page No. 311).

Let $Q(\theta)$ denotes the objective function being maximized, Q_0 denotes its value in the intercept-only model, Q_{fit} denotes the value in the fitted model, and Q_{max} denotes the largest possible value of $Q(\theta)$. Then the maximum potential gain in the objective function resulting from inclusion of regressors is $Q_{max} - Q_0$ and the actual gain is $Q_{fit} - Q_0$. This suggests the measure

$$R^2_{RG} = \frac{Q_{fit} - Q_0}{Q_{max} - Q_0}.$$

where the subscript RG means relative gain. Note that, for least squares, $R^2 = R^2_{RG}$. For both the loss functions, $Q_{max} = 0$.

We also calculated the number of times the Pseudo R^2 for L_4 is numerically greater than that of L_2. It is astonishing to see that 100 percent times the Pseudo R^2 for L_4 is numerically greater than that of L_2.

6.2 Real Life Example

For our empirical analysis, we use the data provided by the National Sample Survey Organization of India viz. the NSSO 68th round all India unit level survey on consumption expenditure (Schedule1.0, Type 1 and 2) conducted during July 2011 to June 2012. This dataset is a nationally representative sample of household and individual characteristics based on a stratified sampling of households. For this round, the dataset is comprised of 1,68,880 household level observations. The dataset provides a detailed list of various household and individual specific characteristics along with the consumption expenditures of the households. In addition to this, data is also provided on the households' localization which includes the sector (Rural or Urban), the district and the state/union territory (henceforth, the union territories will be referred to as states). For our analysis, we use the amount of land possessed (in logarithm form) by the households as our principal (dependent) variable along side various demographic variables as controls (independent variables). The kernel density plot clearly suggest that amount of land possession by rural households does have a bimodal distribution [3]. The plot clearly indicates that India is suffering from "vanishing middle-class syndrome," only marginal and rich farmers are there. The empirical analysis demands some routine and rudimentary summary statistics as given in Table 2. We regress the amount of land possessed (Y) on six explanatory variables, [4] viz, Median age of a household (mage), the number of children below 15 years of age (chlt15), the number of old people above 60 years of age (Ogt60), the number of male member in the households (male), the number of female member in the households (female); and finally the number of member with education level above 10th standard (highedu). We then estimate [5] the linear regression model based on L_2 and L_4. The estimated results are summarized as below:

[3]The same phenomenon is also seen for all-India households (rural and urban together). The plot presented here is for rural household excluding the households with no land. It is interesting to note that bi-modality is observed both for (1) households with non-zero amount of land ; and (2) with all households. All the results presented here are based on rural household with non-zero lands. Number of rural households with non-zero amount of land is 98483. Whole study is based on This set of 98483 observations.

[4]We also tried with many other explanatory variables available in our master file. We also repeated the same exercise for all-india (rural and urban together) households. It is need less to mention that overall findings are same across all models we attempted.

[5]In this paper we do not pursue the endogeneity issue, if any.

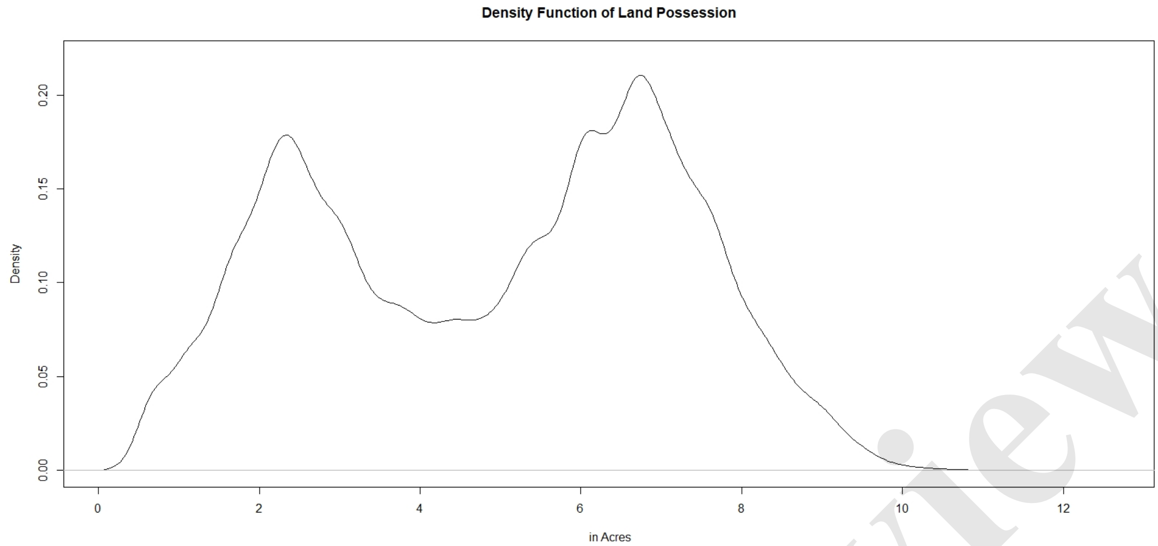

Table 3: Summary Statistics.

Mean	Median	SE	Min	Max	Kurtosis
4.945	5.352	2.290	0.693	12.007	1.944

Note: (i) All the results presented here are based on rural household with non-zero lands. Number of rural households with non-zero amount of land is 98483. (ii) This table is based on non-logarithm data.

Table 4: Model Estimates and Standard Errors .

Variables	L_2	L_4
Intercept	3.03493710	3.47252382
	(0.0342302942)	(0.0120790850)
mage	0.01240624	0.01077902
	(0.0007863559)	(0.0002774869)
chlt15	-0.14759963	-0.09232877
	(0.0082999771)	(0.0029288714)
Ogt60	0.06754135	0.04168707
	(0.0135886351)	(0.0047951174)
male	0.36720544	0.24403590
	(0.0064274676)	(0.0022681058)
female	0.31067372	0.21197564
	(0.0071582759)	(0.0025259912)
highedu	0.18669815	0.16207843
	(0.0074214316)	(0.0026188528)

Note: (i) All the results presented here are based on rural household with non-zero lands. Number of rural households with non-zero amount of land is 98483. Whole study is based on this set of 98483 observations. Logarithm transformation is taken to reduce degree of heteroscedasticity. (ii) Standard errors are provided in the parenthesis. (iii) Least-squares based estimates are used as an initial estimates for L_4 estimation. (iv) 'Rootsolver' in R is used to obtain the L_4 estimates.

It can be noted that Standard errors (SE) of the parameter estimates are provided in the parenthesis. It is to observe, as expected, that SE of L_4 based estimates are significantly and uniformly less than that of L_2 based estimates. The value of our proposed test statistic is 5.30311871 which lies beyond 95 per cent confidence interval (8.90603838 9.09396162) suggesting that L_4 based estimates are more efficient than that of L_2 . The pseudo R^2 for L_4 is 0.14779372 and the same for L_2 is 0.09794736 . The pseudo R^2 also clearly suggests the supremacy of L_4 over L_2.

7 Discussion

This paper tried to give answer to the unassailable question: Does higher order loss function based estimator perform better than the omnipresent least squares? We presume that teacher, student face this question on the first-day class on regression analysis. We tried to show that, in several real life situations, *smooth* higher order loss function based estimator may lead to more efficient estimator as compared to universal least squares. It is true that least squares has one unassailable advantages, its simplicity. It may also be computationally less intensive. However, with the advent of modern computing power, computational issues may hardly be relevant.

Further work may commence in the following directions. A generalized version of the condition similar to the one derived in section 4 may be useful for comparing L_{2k} and L_{2k+2} estimators. This may be obtained following a similar approach i.e. by obtaining the variance of these two estimators using the expression of variance of m-estimators and comparing them. However, estimation of higher moments may have impact on the performance of the proposed decision rule. It may be useful to study the impact of outliers on the parameter estimates coming from higher order based loss function. Comparison of *break-down point* of L_2, L_4 estimators may be very useful. It may also be interesting to find robust standard errors for L_4 for non *i.i.d* set up.

It may be extremely useful to consider a convex combination of loss functions of various degrees. Arthanari and Dodge (1981) considers convex combination of L_1 and L_2 norms; and studies its properties. Convex combination of $L_1, \ldots, L_p$ may lead to more useful estimator; and the resultant estimator is expected to be robust to any distributional assumption. Such combination may give answer to the omnipresent question: What is the optimal loss function for a given data set? The choice and design of loss functions is important in any practical application (see Hennig and Kutlukaya, 2007). Future research will shed light in this direction.

Appendix A

Table 1: Simulated Risk

Sample size	Mixture of T-distribution	(5, -5)	(4,- 4)	(3, -3)	(2, -2)
100	Mixture of T-distn (df=6)	9940	9834	9200	5896
200		9952	9838	9256	5571
500		9961	9844	9309	4861
1000		9965	9856	9337	4309
2000		9975	9896	9348	3763
5000		9973	9912	9445	2833
100	Mixture of T-distn (df=10)	10000	9986	9860	8092
200		9999	9992	9900	8329
500		10000	9997	9959	8842
1000		10000	9996	9973	9041
2000		10000	10000	9991	9368
5000		10000	9999	9988	9666
100	Mixture of T-distn (df=20)	10000	9999	9981	9155
200		10000	10000	9996	9585
500		10000	10000	10000	9846
1000		10000	10000	10000	9950
2000		10000	10000	10000	9984
5000		10000	10000	10000	9999
	Mixture of Beta-distn (asym)	(4,10; 10,4)	(1,4; 4,1)	(2,4; 4,2)	(3,4; 4,3)
100		10000	10000	9691	3632
200		10000	10000	9992	4791
500		10000	10000	10000	7277
1000		10000	10000	10000	9241
2000		10000	10000	10000	9957
5000		10000	10000	10000	10000
	Mixture of Beta-distn (sym)	(4,4; 4,4)	(3,3; 3,3)	(2,2; 2,2)	(1,1; 1,1)

Continued to next page...

...Continued from previous page					
100		2007	3740	7665	9987
200		1881	4748	9445	10000
500		2089	7178	9993	10000
1000		2574	9200	10000	10000
2000		3464	9957	10000	10000
5000		5810	10000	10000	10000
	Mixture of two normal distributions	(3, -3)	(2, -2)	(1, -1)	(0, 0)
100		9999	9806	1862	211
200		10000	9959	1330	26
500		10000	10000	815	0
1000		10000	10000	478	0
2000		10000	10000	265	0
5000		10000	10000	63	0

Appendix B

Proof of Theorem 2: We write

$$\begin{aligned}\widehat{v} - v &= \frac{\widehat{\mu}_6 - \widehat{\mu}_3{}^2}{\widehat{\sigma}^6} - \frac{\mu_6 - \mu_3^2}{\sigma^6} = \frac{\widehat{\mu}_6 - \widehat{\mu}_3^2 - \mu_6 + \mu_3^2}{\widehat{\sigma}^6} - v\frac{\widehat{\sigma}^6 - \sigma^6}{\widehat{\sigma}^6} \\ &= \left[\frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^6 - \left(\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^3\right)^2 - \mu_6 + \mu_3^2}{\widehat{\sigma}^6}\right] - v\left[\frac{(\widehat{\sigma}^2)^3 - (\sigma^2)^3}{\widehat{\sigma}^6}\right],\end{aligned}$$

and hence

$$
\begin{aligned}
& \sqrt{n}(\widehat{v}-v) \\
= \;& \frac{1}{\hat{\sigma}^6}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\overline{x})^6-\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\overline{x})^3\right)\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3\right)-\sqrt{n}(\mu_6-\mu_3^2)\right) \\
& -v\sqrt{n}\left[\frac{(\widehat{\sigma}^2)^3-(\sigma^2)^3}{\widehat{\sigma}^6}\right] \\
= \;& \frac{1}{\hat{\sigma}^6}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\mu)^6-6\mu_5\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\mu)-\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\mu)^3-3\mu_2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\mu)\right)\mu_3\right. \\
& \left.-\sqrt{n}(\mu_6-\mu_3^2)-3\sigma^4 v\sqrt{n}(\widehat{\sigma}^2-\sigma^2)\right]+o_p(1) \\
= \;& \frac{1}{\hat{\sigma}^6}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}((x_i-\mu)^6-\mu_6)-(6\mu_5-3\mu_2\mu_3)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\mu)-\mu_3\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}((x_i-\mu)^3-\mu_3)\right)\right. \\
& \left.-3\sigma^4 v\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}((x_i-\mu)^2-\sigma^2)\right)\right]+o_{p(1)} \\
= \;& \frac{1}{\hat{\sigma}^6}\begin{pmatrix} 1, & -(6\mu_5-3\mu_2\mu_3), & -\mu_3, & -3\sigma^4 v\end{pmatrix}\frac{1}{\sqrt{n}}\begin{pmatrix}\sum[(x_i-\mu)^6-\mu_6] \\ \sum(x_i-\mu) \\ \sum[(x_i-\mu)^3-\mu_3] \\ \sum[(x_i-\mu)^2-\sigma^2]\end{pmatrix}+o_p(1) \\
= \;& \frac{\alpha_0}{\hat{\sigma}^6}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i+o_p(1).
\end{aligned}
$$

□

Note that $\alpha_0=(1,\quad-(6\mu_5-3\mu_2\mu_3),\quad-\mu_3,\quad-3\sigma^4 v)$.

Now,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i=\sqrt{n}\overline{Z}\xrightarrow{d}N(0,\Gamma),$$

where $\Gamma=\lim_{n\to\infty}nE(\overline{Z}\overline{Z}')$ and $\overline{Z}=\frac{1}{n}\sum_{i=1}^{n}Z_i$. Here

$$\overline{Z}=\begin{pmatrix} m_6^o-\mu_6 \\ m_1^o \\ m_3^o-\mu_3 \\ m_2^o-\sigma^2\end{pmatrix},$$

with

$$m_r^o=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^r$$

and

$$\begin{aligned} E(m_r^o) &= \mu_r, \\ V(m_r^o) &= \frac{\mu_{2r} - \mu_r^2}{n}, \\ Cov(m_r^o, m_s^o) &= \frac{\mu_{r+s} - \mu_r \mu_s}{n}, \end{aligned}$$

and hence

$$\Gamma = \lim_{n\to\infty} nE(\overline{ZZ}') = \begin{pmatrix} \mu_{12} - \mu_6^2 & \mu_7 & \mu_9 - \mu_3\mu_6 & \mu_8 - \mu_2\mu_6 \\ \mu_7 & \mu_2 & \mu_4 & \mu_3 \\ \mu_9 - \mu_3\mu_6 & \mu_4 & \mu_6 - \mu_3^2 & \mu_5 - \mu_2\mu_3 \\ \mu_8 - \mu_2\mu_6 & \mu_3 & \mu_5 - \mu_2\mu_3 & \mu_4 - \mu_2^2 \end{pmatrix}.$$

Consequently, we get

$$\sqrt{n}(\widehat{v} - v) \xrightarrow{d} N\left(0, \frac{\alpha_0 \Gamma \alpha_0'}{\sigma^{12}}\right).$$

Let

$$s^2 = \widehat{V(v)} = \frac{\widehat{\alpha}_0 \widehat{\Gamma} \widehat{\alpha}_0'}{\widehat{\sigma}^{12}}.$$

Hence our decision theoretic problem reduces to

$$A_0 : v \geq 9 \quad \text{versus} \quad A_1 : v < 9.$$

We use the statistic

$$T = \frac{\sqrt{n}(\widehat{v} - 9)}{s},$$

which is asymptotically $N(0,1)$, under A_0. ∎

8 Reference

1. Arthanari, T. S. and Y. Dodge (1981): Mathematical Programming in Statistics, New York: John Wiley

2. Box, G. E. P. and G. C. Tiao (1962): A Further Look at Robustness via Bayes's Theorem, *Biometrika* , 49, 419-432.

3. Cameron, A. C. and P. Trivedi (2005): Microeconometrics, Cambridge.

4. Everitt, B. S. (2005): U-Shaped Distribution, Encyclopedia of Biostatistics, 8, John Wiley & Sons.

5. Hennig, C. and M. Kutlukaya (2007): Some Thoughts about the Design of Loss Functions, *REVSTAT – Statistical Journal* 5, 1, 19–39.

6. Rinne, H. (2010): Location-Scale Distributions – Linear Estimation and Probability Plotting Using MATLAB. *Preprint.*

7. Stefanski, L. A. and D. D. Boos (2002): The Calculus of M-Estimation, *The American Statistician*, 56, 1, 29-38.

8. Stigler, S. (1981): Gauss and the Invention of Least Square, *The Annals of Statistics*, 9,3, 465-474.

9. Turner, M. C. (1960): On Heuristic Estimation Methods, *Biometrics* , 16, 299-301.

10. Zeckhauser, R. and M. Thompson (1970): Linear Regression with Non-Normal Error Terms. *The Review of Economics and Statistics*, Vol. 52, No. 3, 280-286