Statistica Applicata, Vol. 4, n. 2, 1992

MAXIMUM LIKELIHOOD AND L_P–NORM ESTIMATORS

Gianna Agrò

Institute of Statistics, University of Palermo.

*In this paper the L_P–norm estimators are proposed to estimate the coefficients of a simple linear regression model under the theoretical assumption of a Normal of Order **p** error distribution.*

*The choice of **p** is solved by maximizing the likelihood function by means of an iterative procedure with the L_P–norm estimation of the simple linear regression parameters.*

1. INTRODUCTION

The L_P–norm estimators are considered among the alternatives to the least squares estimators since this method is sensitive to departure from normality in the residual distribution. It is well known that the performance of the least squares estimators is noticeably degraded when the residual distribution is not Normal.

The L_P–norm estimators minimize the sum of the $\mathbf{p}^{th}$ power of the absolute deviations of the observed points from the regression function; they are a generalization of the least squares technique (where **p**=2) and it has been shown that values of **p** different from 2 are more suitable for estimation when the error distribution is not Normal (see Forsyte (1972), Harter (1977), Money et al. (1982), Mineo (1986)).

In this paper, in the second and third section, two different theoretical approaches to the same problem of L_P–norm estimation are briefly recalled; in the fourth section a new method for estimating the suitable exponent **p** for the L_P–norm estimation of the simple linear regression model parameters is proposed.

Finally the simulation results confirm the good performance of the new method and justify other future developments of the proposed approach.

2. L_P–NORM ESTIMATORS OF REGRESSION PARAMETERS

When observing a sample of n pairs $(y_i, \underline{x}_i)$ where y_i is the dependent variable and $\underline{x}_i=(x_{1i}...x_{mi})'$ the independent variable vector, a general regression model is:

$$y_i = g(\underline{x}_i, \underline{\theta}) + \varepsilon_i \qquad (1)$$

where g(.) is a linear or non linear function in the parameter vector $\underline{\theta} = (\theta_0 \ldots \theta_m)'$ and ε_i is a random error.

To obtain the L_P–norm estimator of the unknown regression parameter vector we need to minimize the following function:

$$S_P(\underline{\theta}) = \sum_{i=1}^{n} |y_i - g(\underline{x}_i, \underline{\theta})|^{\mathbf{P}} \qquad 1 \leq \mathbf{p} < \infty. \qquad (2)$$

There are basically two associated problems to be considered when using the L_P–norm estimator. The first concerns the choice of **p,** while the second concerns the choice of the algorithm used to obtain the solution.

In the literature several algorithms are proposed to solve the second problem for linear or non linear function g(.) while the choice of **p** is solved through various selection rules: Harter (1977) proposed an adaptive procedure depending on the kurtosis β_2 of the error distribution: if $\beta_2 > 3.8$ use **p**=1 (the regression of the least absolute values) if $2.2 < \beta_2 < 3.8$ use **p**=2 (the regression of the least squares) and if $\beta_2 < 2.2$ use **p**=∞ (Chebychev or minimax regression); Money et al. (1982) and Sposito et al. (1983) suggested the following empirical relationships respectively:

$$\hat{\mathbf{p}} = 9 / \hat{\beta}_2^2 + 1 \qquad \text{for } 1 \leq \mathbf{p} < \infty \qquad (3)$$

$$\hat{\mathbf{p}} = 6 / \hat{\beta}_2 \qquad \text{for } 1 \leq \mathbf{p} < 2 \qquad (4)$$

where $\hat{\beta}_2$ is the sample residual kurtosis.

The above formulae derive from simulation results in which several symmetrical error distributions with different β_2 values were considered: Uniform (β_2=1.8), Parabolic (β_2=2.14) Triangular (β_2=2.4), Normal (β_2=3), Contaminated Normal (β_2=4), Contaminated Normal (β_2=5) and Laplace (β_2=6).

Both formulae (3) and (4) are *ad hoc* while another theoretically motivated formula is based on the $\mathbf{p}^{th}$ order exponential distribution which has the following density function:

$$f(z) = \frac{1}{\phi \ \Gamma(1+1/\mathbf{p}) 2^{1+1/\mathbf{p}}} \exp\left[-\frac{1}{2} \left| \frac{z-\mu}{\phi} \right|^{\mathbf{p}} \right] \qquad (5)$$

with μ a location parameter and ϕ a scale parameter (see Turner (1960)). If the residual distribution belongs to this class of exponential distributions and the exponent **p** is known then the maximum likelihood estimate of the parameter vector can be obtained by simply minimizing the sum of the $\mathbf{p}^{th}$ power of the absolute residuals.

It can be shown that the kurtosis β_2 of the distribution (5) is given by:

$$\beta_2 = \frac{\Gamma(5/\mathbf{p})\Gamma(1/\mathbf{p})}{\left[\Gamma(3/\mathbf{p})\right]^2} \tag{6}$$

In the above formulae **p** is function of the kurtosis β_2 which can be estimated on the residuals resulting after a least squares fitting of the model (1).

Instead of using a selection rule an *adaptive procedure* is suggested by Gonin and Money (1985): Fit a curve using least squares. Compute the sample kurtosis of the resulting residuals and make a prediction of the optimal exponent **p** using either formula (3) or (4) or (6). Use this estimate $\hat{\mathbf{p}}$ and fit a new curve to the data. Subsequently compute the sample kurtosis of the resulting residuals and make a new prediction of the true exponent **p.** Repeat the process until no further change in the value of $\hat{\mathbf{p}}$ is detected.

This algorithm consists of an inner iteration (minimization over the parameter) and an outer iteration (calculation of $\hat{\mathbf{p}}$) therefore it calculates the limit of the sequence $\hat{\mathbf{p}}_i$.

As it can be noted, the selection of the suitable exponent **p** needs some knowledge about the distribution of the errors at least in one important characteristic such as the kurtosis; nevertheless if the exponential power function (5) is assumed as error distribution in the regression model (1) then the link between the error distribution and optimal exponent **p** is even clearer. In fact the exponential power function is a family of distributions, specifically it is: a Laplace distribution when **p**=1; a Normal distribution and a Uniform distribution when **p**=2 and **p**=∞ respectively. When one of those distributions is specified, the maximum likelihood estimator of the regression coefficients can be obtained by L_P–norm estimator with **p**=2 if an underlying normal error distribution is assumed and with **p**=1 or **p**=∞ if a Laplace or a Uniform error distributions are assumed.

3. MAXIMUM LIKELIHOOD ESTIMATORS FOR THE REGRESSION PARAMETERS ASSUMING *A NORMAL OF ORDER P* ERROR DISTRIBUTION

The assumption of a random error model is the starting point of this approach: in the regression model (1) we can suppose the errors ε_i are independent, identically distributed variables according to a Normal of Order **p** distribution.

This function is a family of unimodal symmetric curves similar to the exponential power distribution (5). The Normal of Order **p** function (also called Generalized Normal) can be obtained from the density function (5) by expressing the scale parameter ϕ as function of the absolute central moment of order **p.** The

following function, defined on the whole real axis, was also obtained by Vianelli (1963) and Lunetta (1963):

$$f_P(z) = \frac{1}{2\ \mathbf{p}^{1/\mathbf{p}}\Gamma(1+1/\mathbf{p})\ \sigma_P}\ \exp\left[-\frac{1}{\mathbf{p}}\left|\frac{z-\mu}{\sigma_P}\right|^{\mathbf{p}}\right] \tag{7}$$

with $E(z)=\mu$ and $\sigma_P^{\mathbf{p}} = E|z-\mu|^{\mathbf{p}}$.

The above function is characterized by the location parameter μ, the scale parameter σ_P and the shape parameter **p.**

In fact as **p** varies from 0 to ∞, the (7) assumes different shapes with different length of tails and kurtosis:

$0<\mathbf{p}<1$: double exponential distributions. They are cuspidate, very long tailed and have $\beta_2>6$;

p=1: the Laplace distribution. It is cuspidate, long tailed and has $\beta_2=6$;

$1<\mathbf{p}<2$: leptokurtic distributions. They have long tails and $6 > \beta_2 > 3$;

p=2: the normal distribution with $\beta_2=3$;

p>2: platikurtic distributions with short tails and $3 > \beta_2 > 1.8$;

p–>∞: the rectangular distribution with $\beta_2=1.8$.

For a sample of n observed points $(y_i, \underline{x}_i)$, the logarithm of the likelihood function of the model (1) is given by:

$$L = -n\ \lg\left[2\ \mathbf{p}^{1/\mathbf{p}}\sigma_P\ \Gamma(1+1/\mathbf{p})\right] - \frac{1}{\mathbf{p}\ \sigma_P^{\mathbf{p}}}\left[\sum_i^n |y_i - g(\underline{x}_i, \underline{\theta})|^{\mathbf{p}}\right] \tag{8}$$

where $y_i=z$ and $g(\underline{x}_i, \underline{\theta}) = \mu$, for analogy with the relationship (7).

If the shape parameter **p** is specified, we have to calculate the first partial derivatives only respect to θ_i with i=1...m: $\partial L/\partial\theta_i = \mathbf{p}\ \Sigma|y_i - g(\underline{x}_i, \underline{\theta})|^{\mathbf{p}-1}\partial g(x_i, \underline{\theta})/\partial\theta_i$.

Equating to zero the above derivatives, a system of m non linear equations has to be solved in order to obtain the maximum likelihood estimators of the regression coefficients. When the order **p** is specified all the terms in the function (8) are constant except the last part containing the vector $\underline{\theta}$, consequently the maximum likelihood estimators are the L_P–norm estimators, as explained in the previous paragraph. Then in L_P–norm estimators the optimal exponent **p** is the shape parameter of the Normal of Order p function assumed as underlying error distribution.

When the value of **p** is unknown it can be estimated by a particular index of the Normal of Order **p** function. This index, called *Generalized Kurtosis,* is given by Mineo (1978):

$$\beta_P = \frac{\mu_{2P}}{\left[\mu_P\right]^2} = \mathbf{p}+1 \tag{9}$$

where μ_r is the theoretical r^{th} absolute central moment.

A method to obtain maximum likelihood estimates or L_P–norm estimates for the regression coefficients of a simple linear regression model was proposed by Mineo (1989).

The value of **p** is estimated from the index (9) while the iterative procedure is summarized in the following steps:

Step 0: Set i=0 with $\hat{\mathbf{p}}_0 = 2$;

Step 1: Fit the model to the data by using $\hat{\mathbf{p}}_i$;

Step 2: Compute the estimated Generalized Kurtosis $\hat{\beta}_{\hat{p}}$ of the resulting residuals and calculate the value $\hat{\mathbf{p}}_{i+1}$ by solving the equation: $\hat{\beta}_{\hat{p}} = \hat{\mathbf{p}} + 1$;

Step 3: Repeat steps 1 and 2 until $\hat{\mathbf{p}}_i$ converges.

The method supplies an estimate $\hat{\mathbf{p}}$ of the shape parameter **p** and the L_P–norm estimates, or the maximum likelihood estimates, for the regression parameters on the basis of $\hat{\mathbf{p}}$.

4. A NEW PROCEDURE: MAXIMUM LIKELIHOOD ESTIMATION OF THE SHAPE PARAMETER *P*

The new method is derived from the second approach. Consider the simple linear regression model:

$$y_i = \alpha + \beta x_j + \varepsilon_j \qquad j=1,\ldots,n \qquad (10)$$

for a sample of the observed data (x_j, y_j) where the random errors ε_i are independent variables, identically distributed according to (7) with $\mu=0$, $\sigma_P^{\mathbf{p}}$ constant and $1 \leq \mathbf{p} < \infty$.

To simplify the procedure in this study we assumed $\sigma_P^{\mathbf{p}} = 1$ so that it will not appear in the following expressions.

Now the logarithm of the likelihood function, for the observed sample and the model (10), is given by:

$$L(\alpha,\beta,\mathbf{p}) = -n \log\left[2\ \mathbf{p}^{1/\mathbf{p}}\ \Gamma(1+1/\mathbf{p})\right] - \left[(1/\mathbf{p})\ \Sigma\left|y_j - \alpha - \beta x_j\right|^{\mathbf{p}}\right]. \qquad (11)$$

The maximum likelihood estimators $\hat{\mathbf{p}}$, a, b of the parameters **p**, α and β can be jointly calculated by solving a non linear system of the first partial derivatives equated to zero, or by using one of the direct optimization algorithms without derivatives.

The new iterative procedure summarized in the following steps is an alternative in order to minimize (11) (reversing the sign on the right–side):

Step 0: Set i=0 with $\hat{\mathbf{p}}_0 = 2$;

Step 1: Fit the model to the data using $\hat{\mathbf{p}}_i$ and find the L_P–norm estimates of α and β at this step: a_i and b_i;

Step 2: Compute the estimated residuals $\hat{\varepsilon}_j = y_j - a_i - b_i x_j$ and put them in the (11);

Step 3: Minimize the likelihood function (11) to calculate an estimate $\hat{\mathbf{p}}_{i+1}$ of the shape parameter **p** and save the likelihood value L_{i+1};

Step 4: Compare the last likelihood value to the previous one and if $L_{i+1} < L_i$ then repeat from step 1 or else;

Step 5: Put $\hat{\mathbf{p}} = \hat{\mathbf{p}}_i$, $a=a_i$ and $b=b_i$ and use them as maximum likelihood estimates of **p,** α and β parameters.

The above algorithm doesn't produce joint estimates for p, α and β because it follows a zig–zag procedure.

In step 1 a L_P–norm estimation of the regression parameters is made; for a multiple linear regression model the existence of the solution is assured by a full rank matrix of the independent variables and the uniqueness is assured by the exponent $1<\mathbf{p}<\infty$.

The convergences at steps 3 and 4 have not been theoretically proved although they have been empirically verified by simulation tests on samples of different sizes and with different values of **p.**

5. SIMULATION STUDY AND RESULTS

The performance of the new method has been tested through a simulation study in order to verify empirically the unbiasedness and the asymptotic behavior of the maximum likelihood estimators for the simple linear regression model parameters and for the shape parameter **p.**

To accomplish the Monte Carlo simulation, 500 samples of size n=50 have been generated from Normal of Order **p** distributions with six different values of **p.** The algorithm for generating pseudo–random standardized deviates (for $\mathbf{p}\geq 1$) was proposed by Chiodi (1986) and the considered values of **p** were: **p**=(1.0 (0.5) 3.5). Each sample was made of variable pairs (x_j, ε_j) and the values of y_j were obtained from the model (10) with α=1.0 and β=1.0.

The same simulation plan was repeated for samples of size n=100 and n=200.

The conjugated gradient algorithm was used to calculate the L_P–norm estimates in step 1. This was done to avoid the second derivatives of the residual $\mathbf{p}^{th}$ power sum since a zero residual causes an indefinite form when the exponent is $\mathbf{p}<2$.

An algorithm based on a parabolic interpolation method was used to minimize the likelihood function in step 3.

The main results of the simulation experiment are summarized in table I: Mean and variance of the estimates $\hat{p}$ a and b; mean and variance for the estimate $S_{\hat{P}} = \left[\sum |y_j - a - bx_j|^{\hat{p}} / n\right]^{1/\hat{p}}$ of the scale parameter σ_P; $K_{M(a)}$ and $K_{M(b)}$ the empirical standardized deviates of M(a) and M(b) from the parent values α=1 and β=1 of the model (10).

Tab. I: Mean and Variance of $\hat{p}$, a, b, $S_{\hat{P}}$ estimated for a simple linear regression model (α=1, β=1, σ_P=1) on 500 samples of size n=50, 100, 200 generated from *Normal of Order p* distribution with different values of p.

p	M($\hat{p}$)	V($\hat{p}$)	M(a)	V(a)	M(b)	V(b)	$K_{M(a)}$	$K_{M(b)}$	M($S_{\hat{P}}$)	V($S_{\hat{P}}$)
					n=50					
1.0	1.0472	0.0815	0.9862	.0268	0.9980	.0170	–1.8749	–0.3374	0.9620	.01313
1.5	1.6522	0.2066	1.0048	.0224	0.9958	.0222	0.7250	–0.6173	0.9921	.00732
2.0	2.2437	0.3086	1.0025	.0214	1.0058	.0245	0.3952	0.8352	0.9946	.00475
2.5	2.9041	0.6801	1.0051	.0188	0.9868	.0205	0.8333	–2.0523	1.0017	.00264
3.0	3.5001	0.9797	0.9948	.0142	0.9970	.0216	–0.9609	–0.4457	1.0033	.00168
3.5	4.2043	1.4715	0.9973	.0130	0.9848	.0175	–0.5270	–2.5613	1.0030	.00131
					n=100					
1.0	1.0195	0.0285	0.9972	.0108	1.0024	.0072	–0.5841	0.6428	0.9830	.00558
1.5	1.5547	0.0626	1.0071	.0117	0.9989	.0091	1.4681	–0.2418	0.9993	.00296
2.0	2.0929	0.1232	0.9992	.0099	0.9978	.0108	–0.1707	–0.4634	0.9995	.00212
2.5	2.6747	0.2206	1.0036	.0085	0.9896	.0096	0.8554	–2.3480	1.0003	.00147
3.0	3.1975	0.2687	0.9990	.0069	0.9981	.0094	–0.2630	–0.4190	1.0012	.00098
3.5	3.7637	0.3939	1.0003	.0055	0.9949	.0085	0.1094	–1.2305	1.0009	.00076
					n=200					
1.0	1.0106	0.0140	1.0006	.0057	1.0061	.0032	0.1945	2.4373	0.9897	.00294
1.5	1.5221	0.0272	1.0030	.0057	0.9976	.0045	0.9129	–0.7785	1.0017	.00159
2.0	2.0308	0.0498	1.0011	.0051	0.9964	.0052	0.3652	–1.0885	0.9994	.00104
2.5	2.5588	0.0805	1.0002	.0042	0.9931	.0049	0.0797	–2.1267	0.9998	.00078
3.0	3.0777	0.1054	0.9996	.0038	0.9975	.0048	–0.1189	–0.7835	1.0013	.00048
3.5	3.6198	0.1697	0.9989	.0030	0.9964	.0039	–0.4326	–1.2667	1.0002	.00040

For each fixed value of **p** and each sample size the maximum likelihood estimates *a* and *b* seem unbiased and have a very small variance.

The values $K_{M(a)}$ and $K_{M(b)}$ are calculated under the hypothesis that a and b have an asymtotical Normal distribution for each fixed **p**; they confirm the goodness of the estimates of the regression parameters since the empirical deviates are almost always very close to the mean of standardized normal.

Also the estimate $S_{\hat{P}}$ seems unbiased for each combination of **p** and n, and the estimate variance $V\left(S_{\hat{P}}\right)$ is very small.

It can be also noted that the variances of a, b and $S_{\hat{p}}$ decrease in a roughly proportional way as the sample size n increases.

The new important result is the performance of the maximum likelihood estimator of **p:** it is biased and specifically it overestimates the true value of **p.** This bias is more evident for increasing **p** but for greater sample sizes the estimator seems unbiased; this is evident in the result obtained for n=200.

Also the variance of **p** increases for increasing values of **p** but decreases for large sample sizes.

In the figures 1, 2 and 3 the empirical frequency distributions of $\hat{\mathbf{p}}$ obtained for each pair (**p,** n) are shown.

The graphical display confirms the above considerations on the $M(\hat{\mathbf{p}})$ and $V(\hat{\mathbf{p}})$ and shows the skewed frequency distributions. The skewness decreases as the sample size increases, but a possible asymptotic convergence to a Normal distribution seems to be quite slow.

As already stated before, these are the initial results of some studies on maximum likelihood estimators of the regression parameters when the residuals follow a Normal distribution of Order **p.** In a research in progress it will be attempted to further investigate the analytic characteristics of the sampling distribution of $\hat{\mathbf{p}}$ and the simultaneous estimate of σ_P. The empirical distributions of $\hat{\mathbf{p}}$ here presented and the others resulting from the simulation performed by Mineo with his procedure will also be compared to draw some statistical and numerical conclusions.

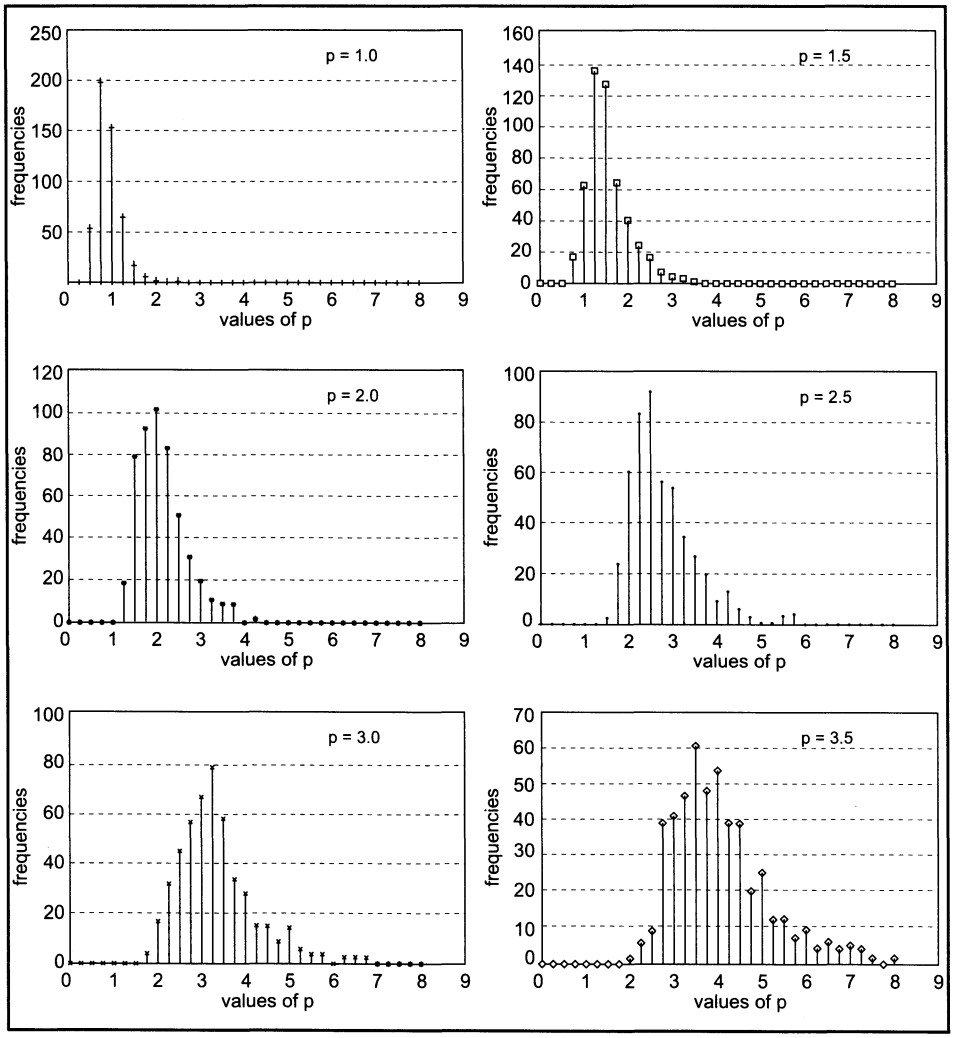

Fig. 1: Empirical frequency distributions of $\hat{p}$ estimated on 500 samples of size n=50 generated from a normal of order p distribution for six different values of p: (3.5 (0.5) 1.0).

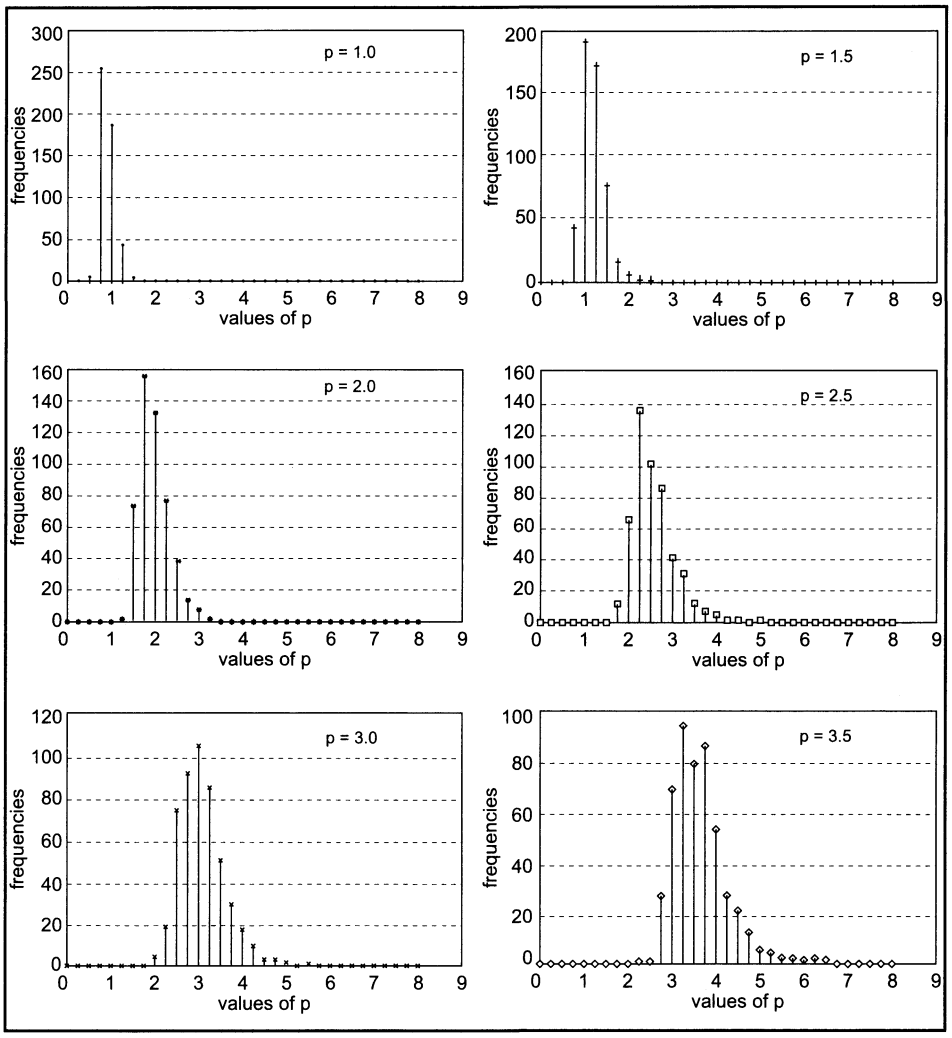

Fig. 2: Empirical frequency distributions of $\hat{p}$ estimated on 500 samples of size n=100 generated from a normal of order p distribution for six different values of p: (3.5 (0.5) 1.0).

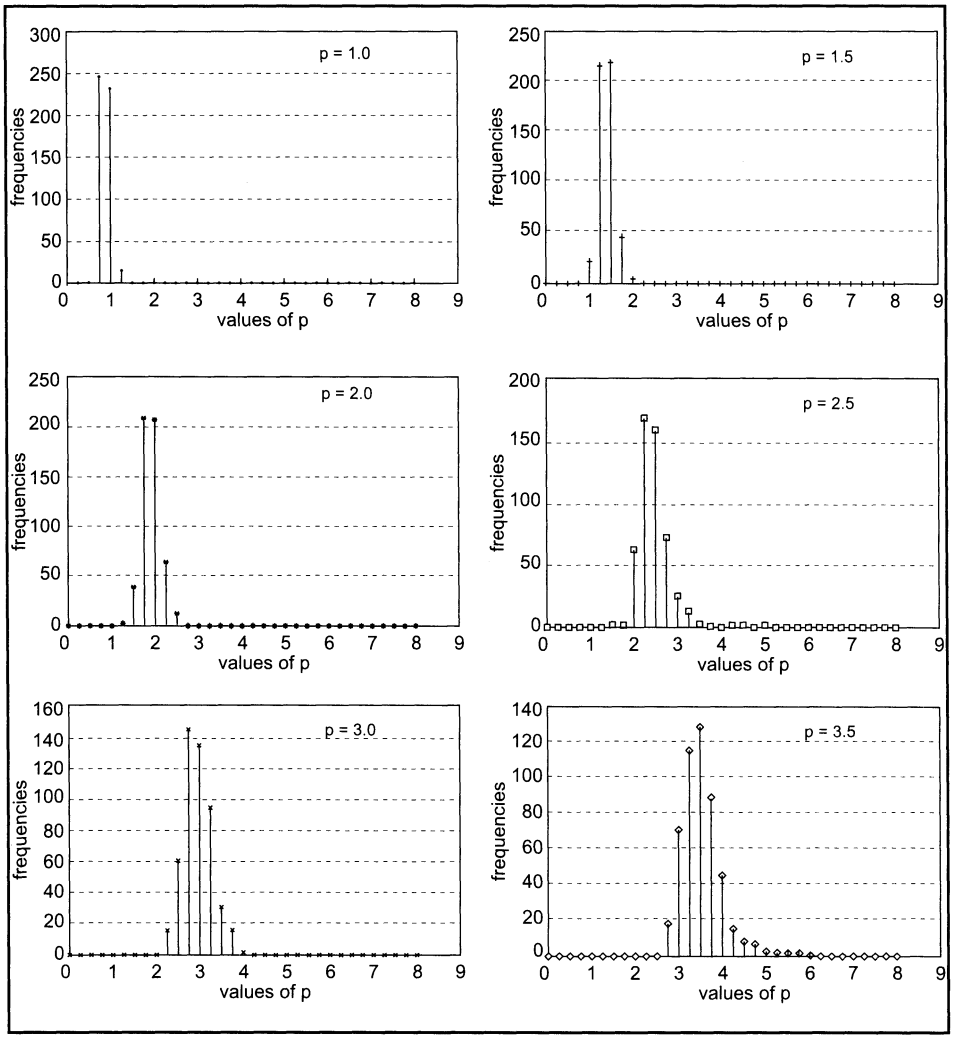

Fig. 3: Empirical frequency distributions of $\hat{p}$ estimated on 500 samples of size n=200 generated from a normal of order p distribution for six different values of p: (3.5 (0.5) 1.0).

REFERENCES

Chiodi M.: *Procedures for generating pseudo–random numbers from a normal distribution of order p,* Rivista di Statistica Applicata, 19, 7–26; 1986.

Forsythe A.B.: *Robust estimation of straight line regression coefficients by minimizing pth power deviations,* Technometrics, 14, 159–156; 1972.

Gonin R. and Money A.H.: *Nonlinear L_p–norm estimation: part II: The asymptotic distribution of the exponent p as a function of the sample kurtosis,* Comm. Statist.– Theor. Meth., 14, 841–849; 1985.

Harter H.L.: *Nonuniqueness of least absolute values regression,* Comm. Statist.– Theor. Meth, A6, 829–838; 1977.

Lunetta G.: *Di una generalizzazione dello schema della curva normale,* Annali della Facoltà di Economia e Commercio di Palermo, 17, 2, 235–244; 1963.

Mineo A.: *Prontuari delle probabilità integrali delle curve normali di ordine p comprese fra $\pm k_p s_p$ e i criteri per la loro valutazione ed il loro impiego,* Istituto di Statistica dell'Università, Palermo, 1978.

Mineo A.: *La migliore combinazione delle osservazioni per la stima dei parametri di intensità e dispersione,* Studi in onore di Silvio Vianelli, Istituto di Statistica dell'Università di Palermo, 1986.

Mineo A.: *The norm–p estimation of location, scale and simple linear regression parameters,* Lecture notes in statistics, Statistical modelling, Procedings Trento 222–233; 1989.

Money A.H. et al.: *The linear regression model: L_p–norm estimation and the choice of p,* Comm. Statist.– Simula. Computa, 11, 89–109; 1982.

Sposito V.A. et al.: *On the efficiency of using the sample kurtosis in selecting optimal L_p–norm estimators,* Comm. Statist.– Simula. Computa., 12, 265–272; 1983.

Turner M.E.: *On heuristic estimation method,* Biometrics, 16, 299–301; 1960.

Vianelli S.: *La misura della variabilità condizionata in uno schema generale delle curva normali di frequenza,* Statistica, 1963.

RIASSUNTO

*Gli stimatori di minima norma–**p**, L_p–norm, vengono qui riproposti per la stima dei parametri di un modello di regressione lineare semplice nell'ipotesi che gli errori si distribuiscano secondo una curva Normale di Ordine **p**.*

*Il nuovo metodo fornisce una stima di massima verosimiglianza del parametro di struttura **p** della distribuzione Normale di Ordine **p**, parametro che è anche l'esponente **p** del metodo di stima di minima norma–**p** nell'ipotesi distribuzionale su specificata.*