Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 20

5-1-2007

Optimal Lp-Metric for Minimizing Powered Deviations in Regression

Stan Lipovetsky
GfK Custom Research North America, Minneapolis, MN, stan.lipovetsky@gfk.com

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 219-227

1538 – 9472/07/$95.00

Optimal Lp-Metric for Minimizing Powered Deviations in Regression

Stan Lipovetsky
GfK Custom Research North America
Minneapolis, Minnesota

Minimizations by least squares or by least absolute deviations are well known criteria in regression modeling. In this work the criterion of generalized mean by powered deviations is suggested. If the parameter of the generalized mean equals one or two, the fitting corresponds to the least absolute or the least squared deviations, respectively. Varying the power parameter yields an optimum value for the objective with a minimum possible residual error. Estimation of a most favorable value of the generalized mean parameter shows that it almost does not depend on data. The optimal power always occurs to be close to 1.7, so these powered deviations should be used for a better regression fit.

Key words: Regression, absolute and squared deviations, *Lp*-metric, gamma-function.

Introduction

The criterion of generalized mean by powered deviations is considered for regression modeling. Usually regressions are constructed by minimization of squared deviations of the observations to a theoretical surface, although some other measures, particularly, absolute deviations are also applied in regression, multidimensional scaling, clustering, and other distance-based techniques (Armstrong & Frome, 1976; Hastie & Tibshirani, 1990; McCullagh & Nelder, 1997; Venables & Ripley, 1997). Robust regression modeling and kernel smoothing use different measures of distance for smaller and bigger deviations (Huber, 1972, 1981; Hill & Holland, 1977; Hampel et al., 1986; Ripley, 1996). Particularly, the *Lp*-metric, or the generalized mean, is widely used as so called *M*-estimator (Maximum likelihood) for robust evaluations (Ramsay, 1977; Sposito, 1982).

In other fields it is also called *Lp*-metric for operators spaces, vector and matrix norms, Hölder's mean, power mean, exponential mean, Kolmogorov's mean, or Minkowski distance (Hardy, Littelwood, & Polya, 1934; Daykin & Eliezer, 1969; Borwein & Borwein, 1987; Korn & Korn, 1988; Alvarez, 1992; Rooij & Heiser, 2005). Power means are related to Box-Cox transformation often used in applied statistics aims (Weisberg, 1985; McCullagh & Nelder, 1997; Tishler & Lipovetsky, 1997, 2000; Lipovetsky & Conklin, 2000).

If the parameter of the generalized mean equals one or two, p=1 or p=2, the fitting corresponds to the least absolute L_1 or the least squared L_2 deviations, respectively. Theoretical properties of the *Lp*-metrics in the range from 1 to 2 were studied in works on approximation theory, Banach's conjecture, and random processes (Breiman, 1968; Fletcher et al., 1971; Kanter, 1973). It is also known due to Jensen's inequality that a generalized mean of a lower power is smaller than a generalized mean of a larger power (Beckenbach, 1946; Korn & Korn, 1988) that is true for the constant set of the averaging values. However, the estimates of the model parameters and the corresponding residual errors depend on a power parameter, so the better generalized power mean can be reached for a smaller power value. In the literature, known numerical simulations indicated that the minimal residuals correspond to the p-powered deviations close to $L_{1.5}$ or $L_{1.8}$ metrics (Gentleman, 1965; Forsythe, 1972; Ramsay, 1977).

Dr. Stan Lipovetsky, GfK Custom Research GfK Custom Research North America, 8401 Golden Valley Road, Minneapolis, Minnesota 55427-0900. Email address: stan.lipovetsky@gfk.com

In the current work, trying an objective of least powered deviations in a wide range of the power parameter, it was possible to find an optimum value for the objective by minimizing the residual error. Numerical estimation of an optimum value of the generalized mean parameter indicates a remarkable outcome – this optimum value is almost a constant that does not depend on the data. Analytical derivation shows that the optimal metric parameter is defined via the gamma function of this parameter, and the optimal value occurs to be close to $p \approx 1.7$. Thus, the optimum metric for fitting any data can be suggested – it is neither the mostly used squared deviations L_2, nor the absolute deviations L_1, but the intermediate powered deviations of $L_{1.7}$.

Powered Deviations in Regression Modeling

Consider a multiple linear regression model of the dependent variable y by n independent variables $x_1, x_2, ..., x_n$:

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_n x_{in} + \varepsilon_i \ , \quad (1)$$

where i denotes observations (i = 1, 2, …, N), and ε_i are deviations of the empirical values y_i from the theoretical model. Least squares minimization corresponds to the objective:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - a_0 - a_1 x_{i1} - ... - a_n x_{in}\right)^2 . \quad (2)$$

This distance is equivalent to the squared Euclidean norm of the errors, or the L_2 metric. Absolute deviations minimization corresponds to the objective of the mean module:

$$S^1 = \frac{1}{N}\sum_{i=1}^{N}\left|\varepsilon_i\right| = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - a_0 - a_1 x_{i1} - ... - a_n x_{in}\right| . \quad (3)$$

It is the Hamming distance (also known as Manhattan, or taxi-driver distance), or L_1 metric.

Generalized powered mean of the deviations can be expressed as follows:

$$S^{2q} = \frac{1}{N}\sum_{i=1}^{N}(\varepsilon_i^2)^q = \frac{1}{N}\sum_{i=1}^{N}\left((y_i - a_0 - a_1 x_{i1} - ... - a_n x_{in})^2\right)^q . \quad (4)$$

In this definition, if power parameter q equals one, than the generalized mean (4) is reducing to the squared mean (2). If q equals one half, the generalized mean (4) is presented as a square root of squared deviation that coincides with absolute value of the deviations in the objective (3). The definition (4) emphasizes that only positive items are summed, and the parameter p of Lp metric equals doubled q-parameter. Then (4) can be simplified by using $2q$ parameter, and represented as the power-mean deviation itself:

$$S = \left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^{2q}\right)^{\frac{1}{2q}} = \left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sum_{j=0}^{n}a_j x_{ij}\right)^{2q}\right)^{\frac{1}{2q}} . \quad (5)$$

where the intercept's variable x_0 identically equals one.

For a given value of power parameter q, minimization of the objective (5) by the parameters of regressions yields a system of the first order partial derivatives:

$$U_k = \frac{\partial S}{\partial a_k} = \left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^{2q}\right)^{\frac{1}{2q}-1}\left[\frac{-1}{N}\sum_{i=1}^{N}x_{ik}\varepsilon_i^{2q-1}\right] = 0 , \quad (6)$$

with errors defined as in (5):

$$\varepsilon_i = y_i - \sum_{j=0}^{n}x_{ij}a_j . \quad (7)$$

Non-linear system of equations (6) can be solved numerically by the Newton-Raphson procedure in the Iteratively Re-Weighted Least Squares (IRLS) approach (Bender, 2000; Lipovetsky & Conklin, 2005). For this algorithm the elements of Hessian, or the matrix of second derivatives, are constructed using the derivatives of (6):

$$\begin{aligned} H_{mk} &= \frac{\partial^2 S}{\partial a_m \partial a_k} \\ &= \frac{\partial U_k}{\partial a_m} \\ &= \frac{2q-1}{N}\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^{2q}\right)^{\frac{1}{2q}-1} G_{mk} \end{aligned}, \tag{8}$$

where the elements G_{mk} are defined by the expression:

$$G_{mk} = \sum_{i=1}^{N} x_{im} x_{ik} \varepsilon_i^{2q-2} - \left(\sum_{i=1}^{N}\varepsilon_i^{2q}\right)^{-1}\left(\sum_{i=1}^{N} x_{im}\varepsilon_i^{2q-1}\right)\left(\sum_{i=1}^{N} x_{ik}\varepsilon_i^{2q-1}\right). \tag{9}$$

Newton-Raphson procedure for finding vector of coefficients a (5) can be presented as:

$$a^{(t+1)} = a^{(t)} - H^{-1}U, \tag{10}$$

where t denotes iteration steps, H^{-1} is the inverted Hessian, and U is the gradient-vector with the elements (6). The round parentheses in (6) and in (8) contain the same constant that is canceled in the expression (10), and also the constant N is canceled, so (10) can be reduced to:

$$a^{(t+1)} = a^{(t)} + (2q-1)^{-1} G^{-1} X'\varepsilon^{2q-1}, \tag{11}$$

where G^{-1} is the inverted matrix of elements (9), X' denotes the transposed matrix of all the regressors in (5), and $X'\varepsilon^{2q-1}$ is matrix notation for the sum in the squared parentheses (6).

It is convenient to introduce a diagonal matrix of powered errors by all observations:

$$\begin{aligned} W &= \mathrm{diag}(\varepsilon^{2q-2}) \\ &= \mathrm{diag}(\varepsilon_1^{2q-2}, \varepsilon_2^{2q-2}, \ldots, \varepsilon_N^{2q-2})' \end{aligned} \tag{12}$$

where ε is the N-th order vector-column of the deviations (7). Then (9) in the matrix form is:

$$G = X'WX - \frac{1}{\varepsilon' W \varepsilon}\left(X'W\varepsilon\right)\left(X'W\varepsilon\right)'. \tag{13}$$

The subtracted outer product in (13) is arranged of the vector $X'W\varepsilon$ of the weighted product of regressors and residuals. Such a product is always close to zero due to the relations of orthogonality between regressors x and residual errors ε. This property is exact for linear and approximate for a nonlinear regression (Lipovetsky & Conklin, 2006).

It is always advisable to keep in only the stable part of the Hessian (Becker & Le Cun, 1988), so it makes sense to reduce (13) to the main first item of the weighted second moment matrix $X'WX$. Then the solution (11) can be simplified to:

$$a^{(t+1)} = a^{(t)} + (2q-1)^{-1}(X'WX)^{-1}X'W\varepsilon, \tag{14}$$

where due to (12) the equality $X'\varepsilon^{2q-1} = X'W\varepsilon$ is used. It is interesting to note that the exact expression (14) yields if instead of the mean deviation objective S (5) the powered-deviation S^{2q} objective (4) is minimized. With (7) in the matrix form, the expression (14) becomes:

$$\begin{aligned} a^{(t+1)} &= (X'WX)^{-1}(X'WX)a^{(t)} \\ &\quad + (2q-1)^{-1}(X'WX)^{-1}X'W(y - Xa^{(t)}), \\ &= (X'WX)^{-1}X'Wz^{(t)} \end{aligned} \tag{15}$$

where the working variable is denoted as:

$$\begin{aligned} z^{(t)} &= Xa^{(t)} \\ &+(2q-1)^{-1}(y-Xa^{(t)}) \\ &=(2q-1)^{-1}\left(y+(2q-2)Xa^{(t)}\right) \end{aligned} . \quad (16)$$

The working variable (16) is a combination of the empirical dependent variable (vector *y*) and the predicted values of the dependent variable (vector $Xa^{(t)}$) at any *t*-th iteration step. The right-hand side (15) shows that the solution is presented as a weighted linear regression of the dependent variable $z^{(t)}$ by all the predictors, so (15)-(16) define the IRLS algorithm.

It is interesting to note that if *q*=1 then $z^{(t)}$ (16) is reducing to the constant vector *y*, and *W* (12) is reducing to the scalar matrix of identical ones, so the problem (5) and solution (15) coincide with a regular linear regression. For *q*=0.5 the Hessian (8) degenerates to zero, so the approach (10) does not work, and the methods of linear programming are mostly applied. The process of minimization (5)-(16) can include the power parameter *q* as well. However, the residuals are usually only weakly dependable on this parameter. So, it is better to find parameters of regression for each fixed *q*, trying *q* in a wide range of its values.

To explain the results on stability of the power parameter that yields the minimum residual errors in regression modeling, assume the normal distribution for the residual errors using the probability density function:

$$f(\varepsilon)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\varepsilon^2}{2\sigma^2}\right), \quad (17)$$

where ε are the residuals (7) and σ is the standard error. For a new random variable of the powered error $\delta=\varepsilon^{2q}$, its probability density function can be defined by the technique of variables transformation (Hogg & Craig, 1969), that yields:

$$f(\delta)=\frac{1}{\sqrt{2\pi}\sigma\cdot 2q}\delta^{\frac{1}{2q}-1}\exp\left(\frac{-\delta^{1/q}}{2\sigma^2}\right). \quad (18)$$

Such a distribution corresponds to a badness of fit function for *M*-estimates in robust regression (Huber, 1972, 1981; Ramsay, 1977). Approximation of the generalized powered mean (4) by the integral of the random variable $\delta=\varepsilon^{2q}$ (18), can be expressed as follows:

$$\begin{aligned} S^{2q} &= \frac{1}{N}\sum_{i=1}^{N}(\varepsilon_i^2)^q = \frac{1}{N}\sum_{i=1}^{N}\delta_i \approx \int_{-\infty}^{\infty}\delta f(\delta)d\delta \\ &=\frac{1}{\sqrt{2\pi}\sigma q}\int_0^{\infty}\delta^{v-1}\exp\left(-\mu\delta^b\right)d\delta \end{aligned} , \quad (19)$$

with the parameters denoted as:

$$v=\frac{1}{2q}+1,\quad \mu=\frac{1}{2\sigma^2},\quad b=\frac{1}{q}. \quad (20)$$

The integral in (19) can be expressed via gamma function (Gradshteyn & Ryzhik, 1965; Gordon, 1994):

$$\int_0^{\infty}\delta^{v-1}\exp\left(-\mu\delta^b\right)d\delta=\frac{1}{|b|}\mu^{-v/b}\Gamma\left(\frac{v}{b}\right), \quad (21)$$

so (19) can be simplified to:

$$\begin{aligned} S^{2q} &= \frac{1}{\sqrt{2\pi}\sigma q}\cdot q(2\sigma^2)^{q+1/2}\Gamma\left(q+\frac{1}{2}\right) \\ &=\frac{2^q\sigma^{2q}}{\sqrt{\pi}}\Gamma\left(q+\frac{1}{2}\right) \end{aligned} . \quad (22)$$

For the case *q*=1, when the generalized power mean (4) is reducing to the least squares, the expression (22) is simplifying to:

$$\begin{aligned} S^2 &= \frac{2\sigma^2}{\sqrt{\pi}}\Gamma\left(1+\frac{1}{2}\right) \\ &=\frac{2\sigma^2}{\sqrt{\pi}}\cdot\frac{1}{2}\Gamma\left(\frac{1}{2}\right), \\ &=\frac{2\sigma^2}{\sqrt{\pi}}\cdot\frac{\sqrt{\pi}}{2} \\ &=\sigma^2 \end{aligned} \quad (23)$$

where the properties $\Gamma(1+x)=x\Gamma(x)$ and $\Gamma(1/2)=\sqrt{\pi}$ of gamma function are applied (Abramowitz & Stegun, 1974). The result (23) proves that the residual mean error estimates the theoretical standard error of the distribution (17). For the case q=1/2, when the generalized power mean (4) reduces to the least absolute deviations, the expression (22) is:

$$S=\frac{\sqrt{2}\sigma}{\sqrt{\pi}}\Gamma(1)=\sqrt{\frac{2}{\pi}}\sigma\cong 0.8\sigma. \tag{24}$$

It is the mean absolute deviation that equals about 80% of the standard deviation (see Abraham & Ledolter, 1983, p. 133). For a positive x, gamma function reaches its minimum $\Gamma(x)=0.886$ at the point x=1.462 (Abramowitz & Stegun, 1974). The q value (22) is by 0.5 less at this point, or q =0.962, so p=2q=1.924 suggests a better powered approximation than the least squares with p=2. Taking the 2q-th root of the expression (22) shows that the generalized residual mean S is proportional to the value of the standard error σ itself. The residual mean S in the units of σ, can be presented up to a constant as the 2q-th root of the gamma function:

$$\frac{S}{\sigma}=\left(\Gamma\left(q+\frac{1}{2}\right)\right)^{\frac{1}{2q}}. \tag{25}$$

This function reaches its minimum at the value $q\approx 0.83$. A difference between theoretical estimate and empirical numerical trying for the best power parameter can be explained by a not exactly normal distribution of the empirical residual errors assumed in the theoretical derivation. Thus, the metric of the smallest residual deviation (4) or (22) equals $p=2q\approx 1.7$. Although the evaluation via gamma function is a rough approximation, but it supports the empirical results that not the least-squares but a slightly-less-than-least-squares powered deviations produce minimum residual error estimations.

Numerical Example

For an illustration of the regular numerical output the data on cars technological solutions is used. This data is given in (Chambers & Hastie, 1992), and is available in the statistical package (*S-PLUS'2000*, 1999, cu.summary file). The data contains the following variables of dimensions and mechanical specifications of 111 various cars, supplied by manufacturers or measured by Consumers Union reports: Weight (y) – pounds (considered in hundreds); Length (x_1) – inches; WheelBase (x_2) – length of wheelbase, inches; Width (x_3) – inches; Height (x_4) – height of car, inches; FrontHd (x_5) – distance between the car's head-liner and the head of a 5ft. 9in. front seat passenger, inches; RearHd (x_6) – a similar distance for the rear seat passenger, inches; FrtLegRoom (x_7) – maximum front leg room, inches; RearSeating (x_8) – rear fore-and-aft seating room, inches; FrtShld (x_9) – front shoulder room, inches; RearShld (x_{10}) – rear shoulder room, inches; Turning (x_{11}) – radius of the turning circle, feet; Disp (x_{12}) – the engine displacement, cubic inches; HP (x_{13}) – the net horsepower; Tank (x_{14}) – fuel refill capacity, gallons; HPrevs (x_{15}) – the red line, or the maximum safe engine speed, rpm. The weight can be considered as an aggregate that has a strong impact on a car's cumulative characteristics, such as mileage per gallon (correlation with weight equals –0.87), and price (correlation with weight equals 0.70).

Regressions were constructed by powered deviations (5) with various values of the parameter q. Several best by the residual characteristics models are presented in Table 1. Each column of Table 1 corresponds to a particular value of q-parameter and contains the coefficients of regression (beginning from the

Table 1. Regressions by several minimized powered deviations.

q	**0.83**	**0.84**	**0.85**	**0.86**	**0.87**	**0.88**	**0.89**	**0.9**	**1.0**	**1.1**
a_0	-50.076	-48.400	-47.617	-49.880	-50.114	-50.275	-50.458	-50.641	-52.388	-54.017
a_1	0.147	0.160	0.150	0.135	0.135	0.135	0.135	0.135	0.135	0.135
a_2	0.081	0.071	0.053	0.086	0.087	0.087	0.087	0.088	0.089	0.089
a_3	0.279	0.288	0.353	0.321	0.319	0.318	0.317	0.317	0.314	0.315
a_4	0.259	0.324	0.355	0.361	0.362	0.363	0.364	0.364	0.370	0.375
a_5	-0.431	-0.098	-0.283	-0.331	-0.330	-0.327	-0.323	-0.319	-0.286	-0.256
a_6	0.708	0.238	0.091	0.098	0.092	0.088	0.083	0.079	0.042	0.011
a_7	0.348	0.305	0.137	0.169	0.170	0.170	0.171	0.172	0.181	0.190
a_8	-0.142	-0.135	-0.123	-0.129	-0.129	-0.129	-0.129	-0.129	-0.129	-0.129
a_9	0.018	-0.107	-0.105	-0.075	-0.073	-0.071	-0.069	-0.067	-0.054	-0.045
a_{10}	-0.001	0.008	0.019	0.017	0.017	0.017	0.017	0.017	0.018	0.019
a_{11}	-0.029	0.002	0.040	0.066	0.068	0.068	0.069	0.069	0.073	0.077
a_{12}	-0.013	-0.019	-0.009	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008
a_{13}	0.060	0.060	0.050	0.049	0.048	0.048	0.048	0.048	0.049	0.049
a_{14}	0.123	0.226	0.159	0.138	0.138	0.138	0.138	0.137	0.132	0.125
a_{15}	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
S^{2q}	1.845	1.430	1.280	1.266	1.279	1.292	1.305	1.319	1.468	1.646
S	1.446	1.237	1.157	1.147	1.152	1.157	1.161	1.166	1.212	1.254
Sabs	1.234	1.015	0.966	0.963	0.963	0.963	0.964	0.964	0.970	0.976
Ssqr	1.543	1.327	1.229	1.213	1.213	1.213	1.213	1.212	1.212	1.212
Mean	-0.733	-0.371	-0.122	0.000	0.011	0.012	0.011	0.010	0.000	-0.010
S^{2q} ***cent***	1.465	1.327	1.269	1.266	1.279	1.292	1.305	1.319	1.468	1.646
S cent	1.259	1.183	1.150	1.147	1.152	1.157	1.161	1.166	1.212	1.254
Sabs cent	1.043	0.965	0.966	0.963	0.963	0.964	0.964	0.965	0.970	0.975
Ssqr cent	1.357	1.274	1.223	1.213	1.213	1.213	1.213	1.212	1.212	1.212

intercept a_0) that are slowly varying across the power parameter q values. Below the coefficients, several estimates for the residual errors are presented: the powered residual S^{2q} (4), the residual deviation S (5), the absolute residual S*abs* (3), and the residual standard error *Ssqr* (corresponds to square root of (2) for mean square root deviation). Note that the last two estimates are obtained by the corresponding set of the regression coefficients. The three of the residual error measures – S^{2q}, S , and *Sabs* – have minimum at the value around q=0.86. The residual mean square root error *Ssqr*, of course, reaches its minimum at the point q=1 that corresponds the least square solution (2). Behavior of these four error measures is shown in Figure 1 in a wide range of q. After initial decreasing and oscillating for q below 0.86, the S^{2q}, S, and *Sabs* curves reach their minima, and then with q increase they grow as well. The residual mean square root error *Ssqr* is very flat beginning from the same threshold q=0.86.

The bottom section of Table 1 presents the estimate of mean value of the deviations (7), and all four residual error estimates centered by this mean value (the error estimates are denoted as S^{2q} *cent*, *S cent*, *Sabs cent*, and *Ssqr cent*). It is

interesting to see that the mean of the deviations is at first negative, than for bigger q values the mean grows and reaches zero at about q=0.86, then it stays positive till the next reach of zero at the value q=1. So, these two values of q produce minimum centered residual error estimates. The mean deviation and the four centered measures of the residual errors are shown in Figure 2 in a range of q values. The behavior of the residual mean stabilizes with q above 0.86. All centered error measures change similarly but more flatly than those of non-centered measures from the previous graph, also with a threshold at the point of about q=0.86. The obtained results on the minimum of S^{2q}, S , and *Sabs* errors in the vicinity of the parameter value about 0.83-0.87 are amazingly constant. In numerous regressions by different data sets the same power region of q is obtained for the minimum residual errors by the powered deviations.

Fig.1: Residual error estimates

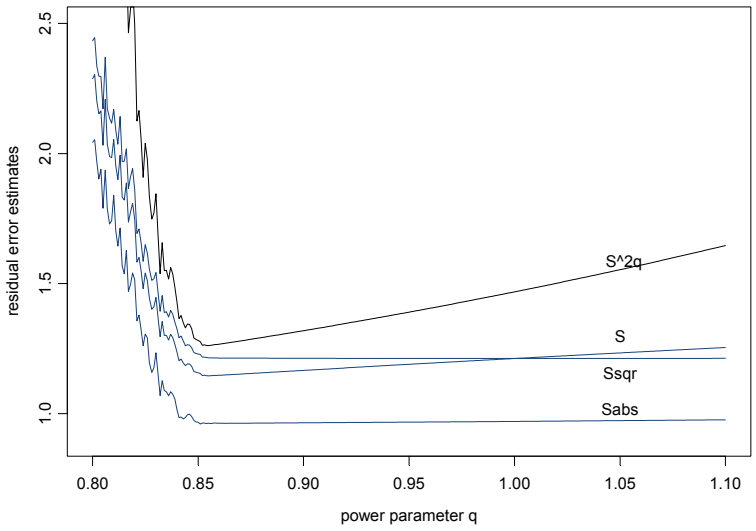
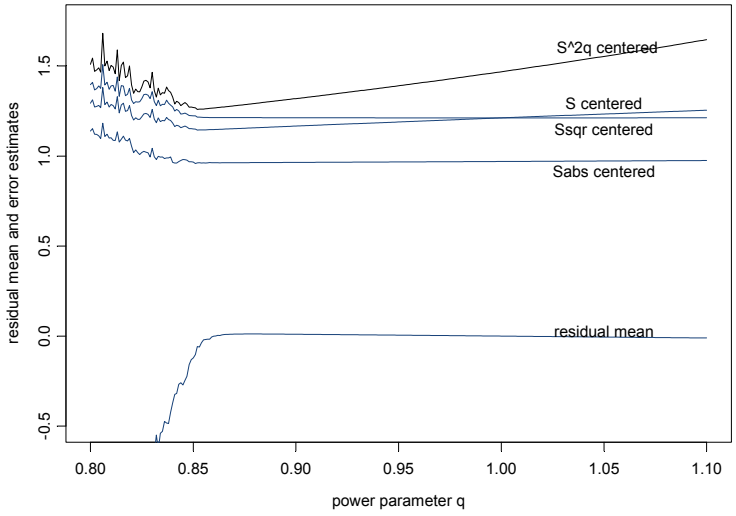


Fig.2: Residual mean and centered error estimates



Conclusion

The generalized powered deviations were considered to estimate minimum possible residual error and the corresponding value of the power parameter. Numerical estimations performed in the work support the analytical result that the best optimization objective corresponds to the metric in the vicinity of $L_{1.7}$. Although change of coefficients and residuals in regressions by different power parameter is moderate, a metric close to the optimum $L_{1.7}$ can be applied for tuning the model. The objective of powered deviations can serve both to the theoretical investigation and practical application in numerous problems of regression modeling.

References

Abraham, B. & Ledolter, J. (1983). *Statistical methods for forecasting*. New York, N.Y.: Wiley.

Abramowitz, M. & Stegun, I. A. (Eds.), (1974). *Handbook of mathematical functions*. National Bureau of Standards. New York, N.Y.: Dover.

Alvarez, S. A. (1992). L^p Arithmetic, *The American Mathematical Monthly*, *99*, 556-662.

Armstrong, R. D. & Frome, E. L. (1976). A comparison of two algorithms for absolute deviation curve fitting. *Journal of the American Statistical Association*, *71*, 328-330.

Becker, S. & Le Cun, Y. (1988). Improving the convergence of back-propagation learning with second order methods. In: Touretzky, D. S., Hinton, G. E., & Sejnowski, T. J. (eds.), *Proceedings of the 1988 Connectionist Models Summer School*, 29-37, Morgan Kaufmann, San Mateo, CA.

Beckenbach, E. F. (1946). An inequality of Jensen. *The American Mathematical Monthly*, *53*, 501-505.

Bender, E. A. (2000). *Mathematical methods in artificial intelligence*, IEEE Computer Society Press, Los Alamitos, CA.

Borwein, J. M. & Borwein, P. B. (1987). The way of all means. *The American Mathematical Monthly*, *94*, 519-522.

Breiman, L. (1968). *Probability*. Reading, M.A.: Addison-Wesley.

Chambers, J. M. & Hastie, T. J. (1992). *Statistical models in S*. Wadsworth & Brooks, Pacific Grove, CA.

Daykin, D. E. & Eliezer, C. J. (1969). Elementary proofs of basic inequalities. *The American Mathematical Monthly*, *76*, 543-546.

Fletcher, R., Grant, J. A., & Hebden, M. D. (1971). The calculation of linear best *Lp* approximations. *Computer Journal*, *14*, 276-279.

Forsythe, A. B. (1972). Robust estimation of straight line regression coefficients by minimizing pth power deviations. *Technometrics*, *14*, 159-166.

Gentleman, W. M. (1965). *Robust estimation of multivariate location by minimizing p-th power deviations*, Ph.D. thesis, Dept. of Mathematics, Princeton University.

Gordon, L. (1994). A stochastic approach to the gamma function. *The American Mathematical Monthly*, *101*, 858-865.

Gradshteyn, I. S. & Ryzhik, I. M. (1965). *Table of integrals, series, and products*. London: Academic Press.

Hampel, F., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York, N.Y.: Wiley.

Hardy, G. H., Littelwood, J. E., & Polya, G. (1934). *Inequalities*. Cambridge: Cambridge University Press.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman and Hall.

Hill, R. W. & Holland, P. W. (1977). Two robust alternatives to least-squares regression. *Journal of the American Statistical Association*, *72*, 828-833.

Hogg, R. V. & Craig, A. T. (1969). *Introduction to mathematical statistics*. New York, N.Y.: Macmillan.

Huber, P. J. (1972). Robust statistics: A review. *Annals of Mathematical Statistics*, *43*, 1041-1067.

Huber, P. J. (1981). *Robust statistics*. New York, N.Y.: Wiley.

Kanter, M. (1973). Stable laws and the imbedding of L^p spaces. *The American Mathematical Monthly*, *80*, 403-407.

Korn, G. A. & Korn, T. M. (1988). *Mathematical handbook for scientists and engineers*. New York, N.Y.: McGraw-Hill.

Lipovetsky, S. & Conklin, M. (2000). Box-Cox generalization of logistic and algebraic binary response models. *International Journal of Operations and Quantitative Management*, *6*, 276-285.

Lipovetsky, S. & Conklin, M. (2005). Latent class regression model in IRLS approach. *Mathematical and Computer Modelling*, *42*, 301-312.

Lipovetsky, S. & Conklin, M. (2005). Ridge regression in two parameter solution. *Applied Stochastic Models in Business and Industry, 21*, 525-540.

McCullagh, P. & Nelder, J. A. (1997). *Generalized linear models*. London: Chapman and Hall.

Ramsay, J. O. (1977). Comparative study of several robust estimates of slope, intercept, and scale in linear regression. *Journal of the American Statistical Association*, *72*, 608-615.

Ripley, B. D. (1996). Pattern recognition and neural networks. Cambridge: Cambridge University Press.

Rooij de, M. & Heiser, W. J. (2005). Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika*, *70*, 99-122.

Sposito, V. A. (1982). On unbiased *Lp* regression estimators. *Journal of the American Statistical Association*, *77*, 652-653.

S-PLUS'2000 (1999). MathSoft Inc., Seattle, WA.

Tishler, A. & Lipovetsky, S. (1997). The flexible CES-GBC family of cost functions: Derivation and application. *The Review of Economics and Statistics*, *LXXIX*, 638-646.

Tishler, A. & Lipovetsky, S. (2000). A globally concave, monotone and flexible cost function: Derivation and application. *Applied Stochastic Models in Business and Industry*, *16*, 279-296.

Weisberg, S. (1985). *Applied Linear Regression*. New York, N.Y.: Wiley.

Venables, W. N. & Ripley, B. D. (1997). *Modern applied statistics with S-PLUS*. New York, N.Y.: Springer.