TJMM
2 (2010), No. 2, 159-163

THE OPTIMIZATION OF SIMPLE REGRESSION MODELS

CONSTANTIN ZĂVOIANU AND FELICIA ZĂVOIANU

Abstract. This paper contains an optimization algorithm for simple regression models, which has the purpose of selecting the best regression equation, if, for the same set of statistical data, several statistically valid models exist. Furthermore, the paper presents a method of determining the confidence intervals of the parameters of the non-linear simple regression models by operating, simultaneously, both with the original model as well as with the corresponding linearized model.

1. Formulating the Problem

Let there $y_1, y_2, \ldots, y_n$ be a set of observations made upon the random variable Y, that correspond to the n experimental values of the factor X, marked up with $x_1, x_2, \ldots, x_n$, viz.:

X	$x_1\ x_2\ x_3\ \ldots\ x_i\ \ldots\ x_n$
Y	$y_1\ y_2\ y_3\ \ldots\ y_i\ \ldots\ y_n$

and $Y^{(k)} = f^{(k)}(\alpha_0^{(k)}, \alpha_1^{(k)}, X, u^{(k)})$, $k = \overline{1,7}$, a family of simple regression models, where $\alpha_0^{(k)}, \alpha_1^{(k)}, u^{(k)} \in R$, $u^{(k)}$ being a residual variable with a normal distribution, with $M(u^{(k)}) = 0$, $D^2(u^{(k)}) = \left(\sigma^{(k)}\right)^2$, i.e. $u^{(k)} \sim N(0, \sigma^{(k)})$. The type of each model (linear or non-linear) depends on the expressions of the functions included in the analysis, as shown in the next definitions:

$$f^{(k)}(\alpha_0^{(k)}, \alpha_1^{(k)}, X, u^{(k)}) =$$

$$\begin{cases} \alpha_0^{(1)} + \alpha_1^{(1)} X + u^{(1)} & , \text{for } k = 1 \quad (\text{ linear model }) \\ \alpha_0^{(2)} + \alpha_1^{(2)} \log_a X + u^{(2)},\ a > 0 & , \text{for } k = 2 \quad (\text{ logarithmical model }) \\ \alpha_0^{(3)} X^{\alpha_1^{(3)}} 10^{u^{(3)}} & , \text{for } k = 3 \quad (\text{ power model }) \\ \alpha_0^{(4)} + \alpha_1^{(4)} \frac{1}{X} + u^{(4)} & , \text{for } k = 4 \quad (\text{ hyperbolical model }) \\ e^{\alpha_0^{(5)} + \alpha_1^{(5)} X + u^{(5)}} & , \text{for } k = 5 \quad (\text{ exponential model }) \\ \frac{b}{1 + e^{\alpha_0^{(6)} + \alpha_1^{(6)} X + u^{(6)}}} & , \text{for } k = 6 \quad (\text{ logistical model }) \\ \frac{1 + e^{\alpha_0^{(7)} + \alpha_1^{(7)} X + u^{(7)}}}{c},\ c \neq 0 & , \text{for } k = 7 \quad (\text{ logistical model }) \end{cases}$$

The task at hand is to determine, in the case of the two variables being strongly correlated, the best estimates, marked up $a_0^{(k)}$and $a_1^{(k)}$, for the statistical parameters $\alpha_0^{(k)}$ and $\alpha_1^{(k)}$, by starting with the values of the two variables, and implicitly the function or functions $f_k(a_0^{(k)}, a_{(1}^{(k)}, X)$, so that the regression equation or equations $\hat{Y}^{(k)} = f_k(a_0^{(k)}, a_1^{(k)}, X)$ can be efficiently used both in analysis as well as in prognosis.

2010 *Mathematics Subject Classification.* 62j02, 62j05.

Key words and phrases. regression models, parameters, linear model, non-linear model, confidence intervals, optimization criteria.

2. The Description of the Optimization Algorithm

In order to answer the requirements of the optimization problem one should follow the next steps:

Step 1. *Determining the correlation coefficient* $r_{Y/X}$ of the two variables, with the $r_{Y/X} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{ns_X s_Y}$ formula, where $\bar{x}$ and $\bar{y}$ are the simple arithmetic averages of the random variables X and Y, obtained using the observed data, and s_X, s_Y are the average square deviations of the random variables X and Y, viz.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,;\, s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}} \,;\, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \,;\, s_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}\,.$$

Step 2. *Checking the intensity of the statistical connection between the two variables.* If $r_{Y/X} \geq 0,5$, the connection has an average or a high intensity, therefore one can move on to Step 3, otherwise, if $r_{Y/X} < 0,5$ the connection is weak and the algorithm will stop.

Step 3. *The linearization of the models and the determination of the correlation coefficients* $r^{(k)}_{W/Z}$. By means of substituting the X and Y variables, all the models taken into consideration can be converted into linear models with the form: $W^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} Z^{(k)} + u^{(k)}$. The expressions of the parameters $\beta_0^{(k)}$ and $\beta_1^{(k)}$ as well as the individual values $w_i^{(k)}$, $z_i^{(k)}$, $i = \overline{1,n}$, of the $W^{(k)}$ and $Z^{(k)}$ variables, are obtained depending on the $\alpha_0^{(k)}$ and $\alpha_1^{(k)}$ parameters and the individual values y_i, x_i, $i = \overline{1,n}$, of the Y and X variables, taking into account the expressions of the $f^{(k)}$, $k = \overline{1,7}$ functions, after making some variable changes (table 1), thus:

Table 1.

k	1	2	3	4	5	6	7
$f^{(k)}$	$f^{(1)}$	$f^{(2)}$	$f^{(3)}$	$f^{(4)}$	$f^{(5)}$	$f^{(6)}$	$f^{(7)}$
$w_i^{(k)}$	y_i	y_i	$\lg y_i$	y_i	$\ln y_i$	$\ln\left(\frac{b}{y_i}-1\right)$	$\ln(c\cdot y_i - 1)$
$z_i^{(k)}$	x_i	$\log_a x_i$	$\lg x_i$	$\frac{1}{x_i}$	x_i	x_i	x_i
$\beta_0^{(k)}$	$\alpha_0^{(1)}$	$\alpha_0^{(2)}$	$\lg\alpha_0^{(3)}$	$\alpha_0^{(4)}$	$\alpha_0^{(5)}$	$\alpha_0^{(6)}$	$\alpha_0^{(7)}$
$\beta_1^{(k)}$	$\alpha_1^{(1)}$	$\alpha_1^{(2)}$	$\alpha_1^{(3)}$	$\alpha_1^{(4)}$	$\alpha_1^{(5)}$	$\alpha_1^{(6)}$	$\alpha_1^{(7)}$

The value of the correlation coefficient for the linearized model number k is:

$$r^{(k)}_{W^{(k)}/Z^{(k)}} = \frac{\sum_{i=1}^{n}\left(z_i^{(k)}-\bar{z}^{(k)}\right)\left(w_i^{(k)}-\bar{w}^{(k)}\right)}{nS_{Z^{(k)}}S_{W^{(k)}}}$$

where

$$\bar{z}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} z_i^{(k)}$$

$$S_{Z^{(k)}} = \sqrt{\frac{\sum_{i=1}^{n}\left(z_i^{(k)}-\bar{z}^{(k)}\right)^2}{n}} \,;\, \bar{w}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} w_i^{(k)} \,;\, S_{W^{(k)}} = \sqrt{\frac{\sum_{i=1}^{n}\left(w_i^{(k)}-\bar{w}^{(k)}\right)^2}{n}}.$$

Step 4. *Solving the linearized models.* For each linearized model, whose correlation coefficient $r^{(k)}_{W^{(k)}/Z^{(k)}}$ satisfies the relation $r^{(k)}_{W^{(k)}/Z^{(k)}} \geq 0,75$, the coefficients of the regression line are determined as follows:

$$\begin{cases} b_1^{(k)} = r_{W^{(k)}/Z^{(k)}}\sqrt{\frac{\sum_{i=1}^{n}\left(w_i^{(k)}-\bar{w}^{(k)}\right)^2}{\sum_{i=1}^{n}\left(z_i^{(k)}-\bar{z}^{(k)}\right)^2}} \\ b_0^{(k)} = \bar{w}^{(k)} - b_1^{(k)}\cdot\bar{z}^{(k)} \end{cases}.$$

Step 5. *Determining the regression equations corresponding to the original models.* For the models selected at the previous step, the values of the $a_0^{(k)}$ and $a_1^{(k)}$ coefficients are determined from the regression equations, corresponding to the original models, depending on the values of the $b_0^{(k)}$ and $b_1^{(k)}$ coefficients from the linearized models, as follows (table 2):

Table 2.

k	1	2	3	4	5	6	7
$f^{(k)}$	$f^{(1)}$	$f^{(2)}$	$f^{(3)}$	$f^{(4)}$	$f^{(5)}$	$f^{(6)}$	$f^{(7)}$
$a_0^{(k)}$	$b_0^{(1)}$	$b_0^{(2)}$	$10^{b_0^{(3)}}$	$b_0^{(4)}$	$b_0^{(5)}$	$b_0^{(6)}$	$b_0^{(7)}$
$a_1^{(k)}$	$b_1^{(1)}$	$b_1^{(2)}$	$b_1^{(3)}$	$b_1^{(4)}$	$b_1^{(5)}$	$b_1^{(6)}$	$b_1^{(7)}$

Step 6. *Checking the validness of the selected models.* The validness check can be made both on the *linearized models* as well as on the *original models.*

a. For each selected linear model $W^{(k)} = b_0^{(k)} + b_1^{(k)} Z^{(k)}$, we calculate:
   - the sum of the squares of the residua $SP_{rez}^{(k)} = \sum_{i=1}^{n} \left(u_i^{(k)}\right)^2 = \sum_{i=1}^{n} (w_i^{(k)} - \hat{w}_i^{(k)})^2$;
   - the sum of the squares of the regression $SP_{reg}^{(k)} = \sum_{i=1}^{n} (\hat{w}_i^{(k)} - \bar{w}^{(k)})^2$;
   - the $F^{*(k)} = \frac{SP_{reg}^{(k)}}{SP_{rez}^{(k)}} \cdot \frac{n-2}{1}$ statistics. If $F^{*(k)} > 4 \cdot F_{\alpha;\,1\,,\,(n-2)}$ then the $\hat{W}^{(k)}$ model is valid from a statistical point of view.

b. For the original models $\hat{Y}^{(k)} = f_k(a_0^{(k)}, a_1^{(k)}, X)$ that were obtained from the information provided by the linear models, we determine:
   - the sum of the squares of the residua $SP_{rez}^{(k)} = \sum_{i=1}^{n} \left(u_i^{(k)}\right)^2$, where the values of $u_i^{(k)}$ residua are determined as shown in table 3;
   - the sum of the squares of the regression $SP_{reg}^{(k)} = SP_{total} - SP_{rez}^{(k)}$, where $SP_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$;
   - the $F^{*(k)} = \frac{SP_{reg}^{(k)}}{SP_{rez}^{(k)}} \cdot \frac{n-2}{1}$ statistics. If $F^{*(k)} > 4 \cdot F_{\alpha;\,1\,,\,(n-2)}$ then the $\hat{Y}^{(k)}$ model is statistically valid.

Table 3.

k	$y_i = f^{(k)}(a_0^{(k)}, a_1^{(k)}, x_i, u_i^{(k)})$	$\hat{y}_i^{(k)} = f_k(a_0^{(k)}, a_1^{(k)}, x_i)$	The $u_i^{(k)}$ residua expressions
1	$a_0^{(1)} + a_1^{(1)} x_i + u_i^{(1)}$	$a_0^{(1)} + a_1^{(1)} x_i$	$u_i^{(1)} = y_i - \hat{y}_i^{(1)}$
2	$a_0^{(2)} + a_1^{(2)} \log_a x_i + u_i^{(2)}$	$a_0^{(2)} + a_1^{(2)} \log_a x_i$	$u_i^{(2)} = y_i - \hat{y}_i^{(2)}$
3	$a_0^{(3)} x_i^{a_0^{(3)}}\, 10^{u_i^{(3)}}$	$a_0^{(3)} x_i^{a_0^{(3)}}$	$\frac{y_i}{\hat{y}_i^{(3)}} = 10^{u_i^{(3)}}$ $u_i^{(3)} = \lg y_i - \lg \hat{y}_i^{(3)}$
4	$a_0^{(4)} + a_1^{(4)} \frac{1}{x_i} + u_i^{(4)}$	$a_0^{(4)} + a_1^{(4)} \frac{1}{x_i}$	$u_i^{(4)} = y_i - \hat{y}_i^{(4)}$
5	$e^{a_0^{(5)} + a_1^{(5)} x_i + u_i^{(5)}}$	$e^{a_0^{(5)} + a_1^{(5)} x_i}$	$\frac{y_i}{\hat{y}_i^{(5)}} = e^{u_i^{(5)}}$ $u_i^{(5)} = \ln y_i - \ln \hat{y}_i^{(5)}$
6	$\frac{b}{1+e^{a_0^{(6)} + a_1^{(6)} x_i + u_i^{(6)}}}$	$\frac{b}{1+e^{a_0^{(6)} + a_1^{(6)} x_i}}$	$\frac{y_i}{\hat{y}_i^{(6)}} = \frac{1+e^{a_0^{(6)} + a_1^{(6)} x_i}}{1+e^{a_0^{(6)} + a_1^{(6)} x_i} e^{u_i^{(6)}}}$ from where $e^{u_i^{(6)}} = \frac{\left(1+e^{a_0^{(6)} + a_1^{(6)} x_i}\right) \hat{y}_i^{(6)} - y_i}{y_i e^{a_0^{(6)} + a_1^{(6)} x_i}}$, and $u_i^{(6)} = \ln\left[\left(1 + e^{a_0^{(6)} + a_1^{(6)} x_i}\right) \hat{y}_i^{(6)} - y_i\right] - \ln y_i - \left(a_0^{(6)} + a_1^{(6)} x_i\right)$
7	$\frac{1+e^{a_0^{(7)} + a_1^{(7)} x_i + u_i^{(7)}}}{c}$	$\frac{1+e^{a_0^{(7)} + a_1^{(7)} x_i}}{c}$	$\frac{y_i}{\hat{y}_i^{(7)}} = \frac{1+e^{a_0^{(7)} + a_1^{(7)} x_i + u_i^{(7)}}}{1+e^{a_0^{(7)} + a_1^{(7)} x_i}}$ and we deduce that $u_i^{(7)} = \ln\left[\left(1 + e^{a_0^{(7)} + a_1^{(7)} x_i}\right) y_i - \hat{y}_i^{(7)}\right] - \ln \hat{y}_i^{(7)} - \left(a_0^{(7)} + a_1^{(7)} x_i\right)$

Observation. Proceeding as previously mentioned, in both cases and for each pair of models $\left(\hat{W}^{(k)}, \hat{Y}^{(k)}\right)$, one should obtain the same values for the sum of the squares of the residua, the sum of the squares of the regression and, implicitly, the $F^{*(k)}$ statistics, thus, double checking the correct determination of the $SP_{rez}^{(k)}$ and $SP_{reg}^{(k)}$ values.

Step 7. *Choosing the best regression equation.* As it is possible to have various valid models, in order to select the best regression equation (the optimum simple regression model) the following criteria can be used:

(1) – the corresponding equation of the model for which the correlation ratio $R_{Y/X}^{(k)} = \sqrt{1 - \frac{SP_{rez}^{(k)}}{SP_{total}}}$ has the highest value;
(2) – the corresponding equation of the model for which the order of size of the $\frac{F^{*(k)}}{F_{\alpha;\,1\,,\,(n-2)}}$ratio is the highest;
(3) – the corresponding equation of the model for which $SP_{rez}^{(k)}$ has the minimum value;
(4) – combinations of the three previously mentioned criteria that also take into consideration the possibility of analyzing and determining the veracity of the results according to the specifics of the studied phenomena.

3. Finding the Confidence Intervals For the Parameters of Simple Regression Models

In order to determine the confidence intervals for the α_0 and α_1 parameters, as well as the confidence intervals for $\hat{y}_i$, $M(\hat{y}_i)$ and $M(\hat{y}_0)$, from a simple regression model, it is necessary to work both with the original model as well as with the linearized model that corresponds to it. As there can be several particularities between any two given models, we will presume that the linearized model of an original model has the form: $W = \beta_0 + \beta_1 Z + u$ and it can be adjusted through $\hat{W} = b_0 + b_1 Z$. The confidence intervals for the parameters of the linear model are the ones in table 4:

Table 4.

Parameter	The lower limit of the interval	The upper limit of the interval
β_0	$p^{(0)} = b_0 - t_{1-\frac{\alpha}{2},n-2} \cdot s_u\sqrt{\frac{\sum_{i=1}^n z_i^2}{n\sum_{i=1}^n (z_i-\bar{z})^2}}$	$q^{(0)} = b_0 + t_{1-\frac{\alpha}{2},n-2} \cdot s_u\sqrt{\frac{\sum_{i=1}^n z_i^2}{n\sum_{i=1}^n (z_i-\bar{z})^2}}$
β_1	$p^{(1)} = b_1 - t_{1-\frac{\alpha}{2},n-2} \cdot s_u\sqrt{\frac{1}{\sum_{i=1}^n (z_i-\bar{z})^2}}$	$q^{(1)} = b_1 + t_{1-\frac{\alpha}{2},n-2} \cdot s_u\sqrt{\frac{1}{\sum_{i=1}^n (z_i-\bar{z})^2}}$
$\hat{w}_i$	$p_i = w_i - t_{1-\frac{\alpha}{2},n-2} \cdot s_u$	$q_i = w_i + t_{1-\frac{\alpha}{2},n-2} \cdot s_u$
$M(\hat{w}_i)$	$\hat{p}_i = \hat{w}_i - t_{1-\frac{\alpha}{2},n-2} \cdot s_u\sqrt{\frac{1}{n} + \frac{(z_i-\bar{z})^2}{\sum_{i=1}^n (z_i-\bar{z})^2}}$	$\hat{q}_i = \hat{w}_i + t_{1-\frac{\alpha}{2},n-2} \cdot s_u\sqrt{\frac{1}{n} + \frac{(z_i-\bar{z})^2}{\sum_{i=1}^n (z_i-\bar{z})^2}}$
$M(\hat{w}_0)$	$p_0 = \hat{w}_0 - t_{1-\frac{\alpha}{2},n-2} \cdot s_u\sqrt{1 + \frac{1}{n} + \frac{(z_0-\bar{z})^2}{\sum_{i=1}^n (z_i-\bar{z})^2}}$	$q_0 = \hat{w}_0 + t_{1-\frac{\alpha}{2},n-2} \cdot s_u\sqrt{1 + \frac{1}{n} + \frac{(z_0-\bar{z})^2}{\sum_{i=1}^n (z_i-\bar{z})^2}}$

In order to determine the confidence intervals for the parameters of the original model, the limits of the confidence intervals that correspond to the parameters from the linearized model and the transformations made upon the original model in order to linearize it, need to be taken into account. For each case, the following intervals are obtained for the parameters of the original model (table 5):

Table 5.

Original model	The (p, q) confidence intervals for the parameters of the model				
	α_0	α_1	$\hat{y}_i$	$M(\hat{y}_i)$	$M(\hat{y}_0)$
$Y = \alpha_0 + \alpha_1 X + u$	$\left(p^{(0)}, q^{(0)}\right)$	$\left(p^{(1)}, q^{(1)}\right)$	(p_i, q_i)	$(\hat{p}_i, \hat{q}_i)$	(p_0, q_0)
$Y = \alpha_0 + \alpha_1 \log_{\ a} X + u$	$\left(p^{(0)}, q^{(0)}\right)$	$\left(p^{(1)}, q^{(1)}\right)$	(p_i, q_i)	$(\hat{p}_i, \hat{q}_i)$	(p_0, q_0)
$Y = \alpha_0 X^{a_1} 10^u$	$\left(10^{p^{(0)}}, 10^{q^{(0)}}\right)$	$\left(p^{(1)}, q^{(1)}\right)$	$(10^{p_i}, 10^{q_i})$	$\left(10^{\hat{p}_i}, 10^{\hat{q}_i}\right)$	$(10^{p_0}, 10^{q_0})$
$Y = \alpha_0 + \alpha_1 \frac{1}{X} + u$	$\left(p^{(0)}, q^{(0)}\right)$	$\left(p^{(1)}, q^{(1)}\right)$	(p_i, q_i)	$(\hat{p}_i, \hat{q}_i)$	(p_0, q_0)
$Y = e^{\alpha_0 + \alpha_1 X + u}$	$\left(p^{(0)}, q^{(0)}\right)$	$\left(p^{(1)}, q^{(1)}\right)$	(e^{p_i}, e^{q_i})	$(e^{\hat{p}_i}, e^{\hat{q}_i})$	(e^{p_0}, e^{q_0})
$Y = \frac{b}{1+e^{\alpha_0+\alpha_1 X+u}}$; $b > 0$	$\left(p^{(0)}, q^{(0)}\right)$	$\left(p^{(1)}, q^{(1)}\right)$	$\left(\frac{b}{1+e^{q_i}}, \frac{b}{1+e^{p_i}}\right)$	$\left(\frac{b}{1+e^{\hat{q}_i}}, \frac{b}{1+e^{\hat{p}_i}}\right)$	$\left(\frac{b}{1+e^{q_0}}, \frac{b}{1+e^{p_0}}\right)$
$Y = \frac{1+e^{\alpha_0+\alpha_1 X+u}}{c}$; $c > 0$	$\left(p^{(0)}, q^{(0)}\right)$	$\left(p^{(1)}, q^{(1)}\right)$	$\left(\frac{1+e^{p_i}}{c}, \frac{1+e^{q_i}}{c}\right)$	$\left(\frac{1+e^{\hat{p}_i}}{c}, \frac{1+e^{\hat{q}_i}}{c}\right)$	$\left(\frac{1+e^{p_0}}{c}, \frac{1+e^{q_0}}{c}\right)$

References

[1] Harrell F.E.Jr., *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer, 2006.

[2] Isaic-Maniu Al., Mitru C., Voineagu V, *Statistică*, Editura Universitară, București, 2003.

[3] Panaretos J., www.stat-athens.aueb.gr/˜jpan/diatrives/Tsiptsis/ chapter3.pdf.

[4] Pecican E.St., *Econometrie*, Editura All, București, 1994.

[5] Rencher A.C., *Linear Models in Statistics*, Wiley-Interscience, 1999.

[6] Hans von Storch, Zwiers F.W., *Statistical Analysis in Climate Research*, Cambridge University Press, Cambridge, UK, 2002.

[7] Zăvoianu Felicia, *Models for evaluating the impact of economic restructuring at a territorial level*, Doctoral Thesis, A.S.E. București, 2007.

University of Petroșani
Department of Mathematics and Computer Science
Universității 20, 332006 Petroșani, România
E-mail address: constantin.zavoianu@yahoo.com

University of Petroșani
Department of Mathematics and Computer Science
Universității 20, 332006 Petroșani, România
E-mail address: fzavoianu@yahoo.com