APPLICATION OF ROBUST METHODS TO CAR OWNERSHIP TRENDS MODELLING IN JOHANNESBURG

John Kelly

City of Johannesburg, P.O. Box 30733, Braamfontein 2017

ABSTRACT

Robust regression and local polynomial smoothing are applied to the inverse problem for the logistic differential equation (DE) model, in order to develop a more objective, accurate and automatable trends model. A method of inferring the time shift parameter is proposed and applied, allowing the closed form solution of the DE to be used for the prediction of ownership levels in Johannesburg.

A simulation study is employed to verify and evaluate the application of non-linear regression to the inverse problem. It is demonstrated that considerable improvements in accuracy, over the transformation to linear-form method, can be obtained. However, on application to actual data, the non-linear regression algorithm fails to converge.

The appropriateness of the methods in the case of lower asymptote and early growth phase data, and heterogeneous populations are investigated by simulation.

1. BACKGROUND

The car ownership data available to the City is from the vehicle register. Typically, data from administrative sources contain a large proportion of outliers. Moreover, car ownership data are a time series with a significant degree of autocorrelation, in violation of lest squares assumptions. These deficiencies are usually overcome by the manual removal of outliers. The main shortcoming of such a procedure is that it is subjective. This can lead to *ad hoc* influences on inferences, most dangerously in the direction of preconceived beliefs.

Therefore, the methods of analysis used must be correspondingly robust and objective. They should not be too sensitive to departures from model assumptions or the presence of a substantial number of outliers. Given the contaminated nature of the data, ordinary least squares regression is not suitable. Robust regression techniques are recommended.

2. GOALS

1. To develop a more robust and objective analysis method. Specifically, to eliminate the need for the removal of outliers.
2. To develop a method of inferring the time parameter of the solution of the logistic DE.
3. To improve the accuracy of the numerical method.
4. To investigate the performance of the techniques on early growth data.
5. To investigate the performance of the techniques on heterogeneous population data.

3. THE LOGISTIC DIFFERENTIAL EQUATION

The logistic differential equation (DE) is traditionally used to model the growth in car ownership. The DE is used in a wide range of bounded growth applications, such as quantitative ecology and chemical kinetics (Arrowsmith and Place, 1992 p. 12). Although the logistic equation is a very simple non-linear DE (quadratic), it can give rise to a rich and complex dynamics that has made the logistic DE one of the most famous and studied DE in mathematics (Arrowsmith and Place, 1992 pp. 245-250). The Logistic DE can be written as:

$$\frac{dx}{dt} = \frac{\kappa}{\alpha} x(\alpha - x) \tag{1}$$

$$x(t_0) = x_0 \tag{2}$$

where

t is time.

The time unit used in this study is the year, and 0h00 1st January is t=0. Car ownership level (ownership per thousand (1000) of the population) at time t is designated as x(t). The initial ($t = t_0$) level of ownership is x_0. α is the saturation level of ownership parameter, and κ is the ownership growth rate parameter.

The car ownership level x(t) is called a state variable, and α and κ are referred to as parameters. A closed form solution for the above DE is:

$$x(t) = \alpha\left(1 + e^{-\kappa(t-\gamma)}\right)^{-1} \tag{3}$$

Where

γ is the time parameter arising, as a constant of integration, out of the initial conditions of the DE.

4. THE INVERSE PROBLEM

The problem of inferring the parameters of a DE from a sample of observations (x_i, t_i) constitutes what is known as an inverse problem. The standard method of inferring the values of the parameters α, κ is to apply a linearizing transformation. This changes the non-linear DE inverse problem to a linear regression problem. Applying the transformation

$$y_i = \frac{1}{x_i}\frac{x_{i+1} - x_i}{t_{i+1} - t_i} = \frac{\kappa}{\alpha}(\alpha - x_i) \tag{4}$$

to the data gives the regression equation:

$$y_i = \kappa - \frac{\kappa}{\alpha} x_i \tag{5}$$

The parameters can be obtained from the regression parameters β_0 (intercept) and β_1 (gradient), by

$$\alpha = -\frac{\beta_0}{\beta_1} \tag{6}$$

and

$$\kappa = \beta_0 \tag{7}$$

Note that the transformation method does not infer a value for the time parameter γ. As a consequence, future values of car ownership cannot be forecast. Neither can errors with respect to observed values be inferred.

5. METHOD OF VIRTUAL SAMPLES

Linear regression cannot be used directly to infer the time parameter. However, the inferred saturation and growth parameters can be used to construct a virtual sample of time parameters. The time parameter can then be inferred as the arithmetic mean of the virtual sample.

The predicted value of the ith observation of the car ownership level is:-

$$\hat{x}_i = \hat{\alpha}\left(1 + e^{-\hat{\kappa}(t_i - \hat{\gamma})}\right)^{-1} \ldots \tag{8}$$

solving for the time parameter gives

$$\hat{\gamma}_i = t_i + \frac{1}{\hat{\kappa}} \ln\left(\frac{\hat{\alpha}}{x_i} - 1\right) \tag{9}$$

$\{\hat{\gamma}_i\}_{i=1}^{n}$ will be called a virtual sample (for the time parameter). Since γ has a strong physical meaning, it is assumed that the mean and variance exist for the distribution of $\hat{\gamma}_i$, hence by the Law of Large Numbers:

$$\hat{\gamma} = \mu_\gamma = \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i \tag{10}$$

6. ROBUST REGRESSION

Three robust regression methods are described below.

6.1 Least Trimmed Squares (LTS)

Least Trimmed Squares (LTS) is a rather simple method, consisting of three steps. In step one, an ordinary least squares regression is performed. The second step is to calculate the confidence intervals corresponding to the data points, and remove any points lying outside the intervals. Thirdly, an ordinary least squares regression is performed on the reduced data set. LTS, despite its intuitive appeal, suffers from a serious disadvantage. There exists a substantial risk that the initial OLS will not approximate the dominant trend in the data.

6.2 MM-Robust Regression

MM-Robust Regression is performed in two steps. In the first step, the subset of observations constituting the dominant trend is identified by use of the S-estimate of location and scale. In the second step, the regression is performed with points further from the dominant trend having their influence discounted.

The kernel function ρ_c will give a greatly inflated value to points situated 'far' from the dominant location. When the sum of the kernels is minimized the distant points (outliers) will contribute large terms in the sum, and therefore will have little influence on the regression parameters. The kernel function used in this application of MM-Robust regression is the bi-square function ("S-Plus 2000 Guide to Statistics vol. 1" 1999 p. 283).

6.2.1 Finding the Dominant Trend

The S-estimate of location and scale is found by solving the equation:

$$\frac{1}{p}\sum_{i=1}^{p}\rho_c\left(\frac{y_i-\underline{\theta}^T x_i}{s}\right)=\frac{1}{2} \tag{11}$$

for a set of $n_J = \lceil (6.9)2^p \rceil$ sub-samples from the original sample of the points (x_{1i}, …. x_{pi}, y_i) for i= 1, …. n. The S-estimate of location $\underline{\theta}^0$ of the dominant trend is approximated by the smallest $\underline{\theta}$ in the re-samples, and the corresponding scale s is the S-estimate of scale ("S-Plus 2000 Guide to Statistics vol. 1" 1999 p. 282), (Yohai 1987 p. 646).

6.2.2 M-Estimators

An M-estimator (of $\underline{\theta}$) is a generalization of the maximum likelihood estimator. The M-estimator of $\underline{\theta}$ for the kernel function ρ_c is:

$$\underline{\hat{\theta}}=\underset{\underline{\theta}\in NN(\underline{\theta}^0)}{\arg\min}\frac{1}{n}\sum_{i=1}^{n}\rho_c\left(\frac{y_i-\underline{\theta}^T \underline{x}_i}{s}\right). \tag{12}$$

The notation $\underline{\theta}\in NN(\underline{\theta}^0)$ is taken to indicate that the appropriate value of $\underline{\theta}$, among the minimizing arguments. That is the one closest to $\underline{\theta}^0$ ("S-Plus 2000 Guide to Statistics vol. 1" 1999 pp. 282 & 283), (Yohai 1987 pp. 644 & 645).

6.3 Local Polynomial Smoothing

In order to transform the non-linear DE into a simple linear equation, the gradient at each observed point has to be calculated. The errors in calculating the gradient can be amplified by a small but opposite error in car ownership (x), at adjoining points. A substantial reduction in the errors incurred can be obtained by the use of local polynomial smoothing (LOcally WEighted Scatter plot Smoothing, i.e. LOWESS) a robust inference technique (Fan J. and Gijbels I.1996 pp. 24-26).

7. APPLICATION TO JOHANNESBURG DATA

7.1 Methodology

The data are those published in the "Moving SA" November 1997 (Appendix 10.2 "Car Registration Data" Table 4, Appendix 10.3 "Population Data" Table 4) car ownership study. For reasons discussed above, several regression methods will be applied. The regression algorithms are: 1. ordinary least squares, a non-robust method, 2. least trimmed squares, 3. robust regression (MM-regression). In addition, the result of smoothing the data prior to applying the linearizing transformation will be compared.

7.2 Comparing Regression Algorithms

The results of running the Moving SA data under the three regression algorithms are given in Table 1 below. The abbreviations DNE: does not exist, and Dev.: diverges are used.

Table 1 Comparison of inferred Values of Parameters

Regression Algorithm	No LOWESS				LOWESS			
	$\hat{\alpha}$	$\hat{\kappa}$	$\hat{\gamma}$	NLR	$\hat{\alpha}$	$\hat{\kappa}$	$\hat{\gamma}$	NLR
Ordinary Least Squares (OLS)	121	-0.054	DNE	Dev.	428.2	+0.065	1977.5	Dev.
Least Trimmed Squares (LTS)	33	-0.0046	DNE	Dev.	454.4	+0.059	1979.9	Dev.
MM-Robust Regression (MMRR)	30	-0.004	DNE	Dev.	484.7	+0.041	1985.3	Dev.

7.3 Conclusions

1. The values of the parameters, inferred using LOWESS, are more credible than those obtained without LOWESS. The regression algorithms, applied after LOWESS, make little difference to the credibility of the parameter values. Credibility is based on comparative values and professional judgment.
2. A more objective method is to use the goodness of fit. The scatter plot of the original un-smoothed data against the predicted curve (Figure 3) shows that MM-Robust regression with local polynomial smoothing gives a good fit to the data.
3. Figure 1 shows the convergence of the bandwidth parameter to a stable value allowing an objective choice of bandwidth to be made. Considering the closeness of the fit (Figure 3) it is unlikely that the convergence of $\hat{\alpha}$, $\hat{\kappa}$ are artefacts of the smoothing algorithm. A belief further supported by the similarity of both of the limiting values ($\hat{\alpha}$, $\hat{\kappa}$) of the bandwidths on convergence.
4. It has been shown that smoothing the data produces (Figure 2) objectively obtained results in an automated way.
5. These improved values of the parameters are not in the convergence region of the non-linear algorithm.
6. The virtual sample method enabled the time parameter to be inferred.

Table 2 Forecasts

Year	2010	2015	2020	2025	2030
Ownership (cars per 1000 persons)	355.6	374.0	390.6	405.1	417.9

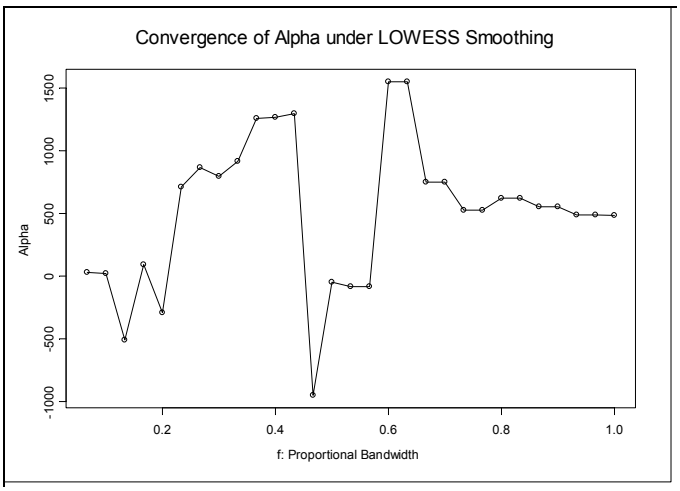

Figure 1

Figure 1 is the **plot of the saturation parameter against the proportional bandwidth**. The un-linearized data was smoothed at various bandwidths and the limiting bandwidth was selected for the smoothing transformation.

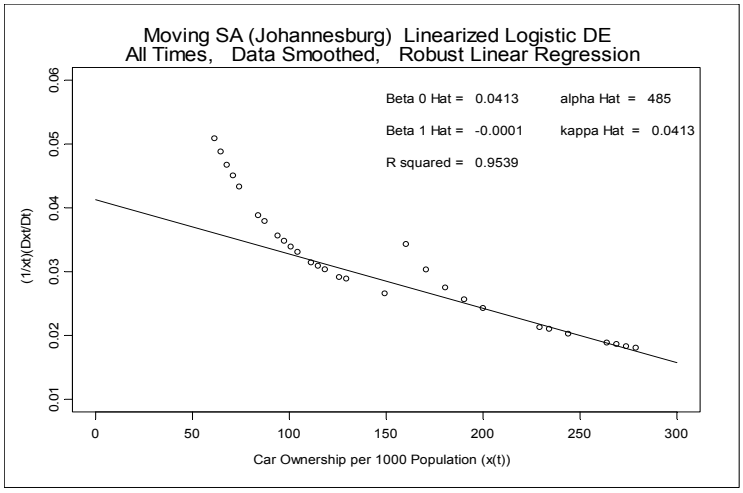

Figure 2

Figure 2 is the **scatter plot of the MM-robust regression fit** to the linearized smoothed data. The inferred values of the parameters: Beta 0 hat (intercept), and Beta 1 hat (gradient), are displayed in the upper left hand corner of the plot; as are the inferred logistic parameters alpha hat and kappa hat, calculated from the regression parameters. Note the excellent R squared value of 0.95.

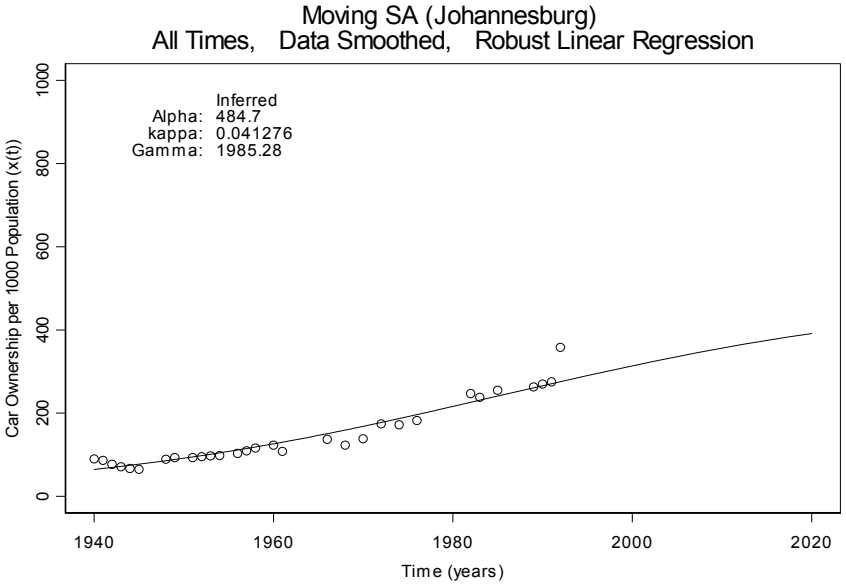

Figure 3

Figure 3: *Nota Bene*: Scatter diagram of the original un-smoothed data. Hence the diagram is a true refection of the fit of the model to the data.

8. SIMULATION STUDY OF THE APPLICATION OF NON-LINEAR REGRESSION

8.1 The Two Stage Method

Stage 0 Simulate Data: Stage 0.1: Prior values for the parameters are set. **Stage 0.2:** Random observation times at a given average observation rate are generated. **Stage 0.3** Using the generated observation times and the prior parameter values the car ownership levels are computed, from the logistic model (equation 3). **Stage 0.4:** Gaussian noise (error) is then added to the solution. A sample size of 108 and a standard deviation of 35 (vehicles per thousand persons) is used in all of the simulations. **Stage 1 Linear Regression:** Apply the linearization transformation to the logistic differential equation, from which is obtained, by robust regression, a value of the saturation parameter (α) and the growth parameter (κ). **Stage 2 non-Linear Regression:** Using the values obtained from the first stage as initial values; solve the nonlinear least-squares problem (algorithm: Gauss-Newton).

8.2 Conclusions

1. The good approximation of the inferred values to the pre-set, known values of the parameters, and the good fits shown in the scatter plot (Figure 4); verify that the coded algorithm is correct.
2. The saturation parameter (α) is accurately inferred by non-linear regression the mean absolute error (MAE) is approximately 1%. For MMRR the MAR is approximately 30%.
3. Both methods are lest accurate for the growth parameter (κ). For MMRR the MAE is in excess of 100%. In contrast, for the nonlinear method the MAE is about 3%.
4. The time parameter (γ) is most accurate (MAE 0.1%).
5. Non-linear regression clearly out-performs the other inference methods.
6. Nonlinear robust regression proper is worth investigating.

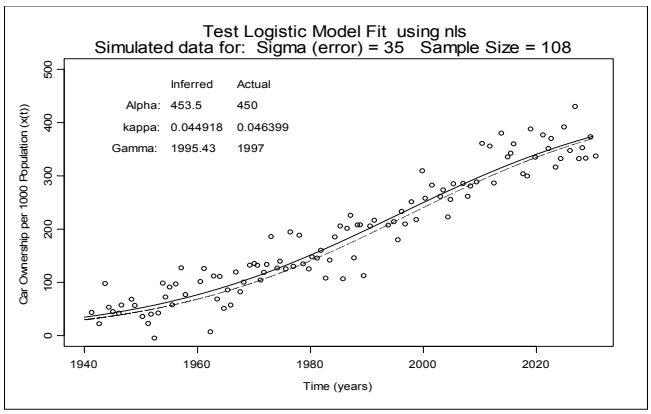

Figure 4

Figure 4 is the scatter plot for the non-linear regression.

1. The points are the untransformed simulated data.
2. The solid curve is the inferred ownership.
3. The dashed curve is the 'actual' ownership in terms of prior set values of the parameters.
4. the actual and inferred values of the parameters displayed in the upper left-hand corner.

9. SENSITIVITY TO EARLY LOW GROWTH PHASE DATA

9.1 Goal

To determine the effect on inferred values of the parameters of only having observations from the early growth phase of the logistic curve.

9.2 Method

The method used is similar to that used in the simulation study of non-linear regression.

1. Prior values for the parameters are set. **2.** The observation period (0, T) is defined as a proportion (T/ γ) of the period (0, γ). **3.** Random observation times at a given average observation rate are generated in the interval (0, T). **4.** Using the generated observation times and the prior parameter values, the car ownership levels are computed, from the logistic model (eq. 3). **5.** Gaussian noise is then added to the solution. **6.** The two-stage non-linear regression method is used to infer the parameters.

9.3 Conclusion

1. Using the restricted data there is a large under estimation of all three parameters. **2.** The error is greater the lower the proportion of (0, γ) that is used. **3.** Errors are substantially reduced as T approaches γ.

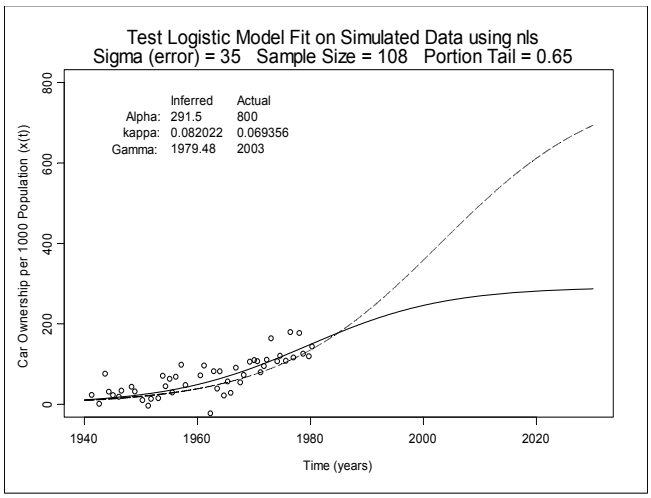

Figure 5

Figure 5 is a scatter plot of the non-linear regression fit to the simulation test on the early growth phase data.

1. The points are the truncated simulated data for 65% of way to the maximum growth point γ.
2. The solid curve represents the inferred ownership values.
3. The dashed curve is the actual ownership values.
4. the actual and inferred values of the parameters are displayed in the upper left-hand corner.

10. EFFECTS OF SUB-POPULATIONS

10.1 Goal

Due to the poverty resulting from previous discriminatory practices in South African society, certain groups were less able to attain car ownership. Subsequently various measures have been put in place to accelerate the socio-economic development of previously disadvantaged groups. Both the previous discrimination and the corrective measures possibly constitute systematic departures from the simple logistic growth model. The goal of the sub-populations effects investigation is to determine a possible range of variation due to the above effects.

10.2 Method

Once again the method used is similar to that used in simulation study of non-linear regression. **1.** Two populations are considered, a previously disadvantaged population, and an economically advantaged population. **2.** Two sets of prior values for the parameters are chosen. **3.** It is assumed that, in the long run, the socio-economic conditions of the two populations will equalize, producing a common saturation parameter (α=650). **4**. The difference in development is modelled using the time parameter (γ) and the growth parameter (κ). **5.** The sample size and standard deviations are as before. **6.** Values the car ownership levels are generated using the logistic model

$$(x(t)= x_1(t)+ x_2t \qquad (13))$$

where

$$x_i(t)= \alpha_i(1+e^{-\kappa_i(t-\gamma_i)})^{-1}) \qquad (14)$$

7. Gaussian noise is then added to the solution.

10.3 Conclusion

1. The departures from the simple homogeneous model are clearly apparent by inspection of the scatter plot (Figure 6). **2.** If the parameters are to be inferred, the simplest method would be to separately model the populations using the two stage method.

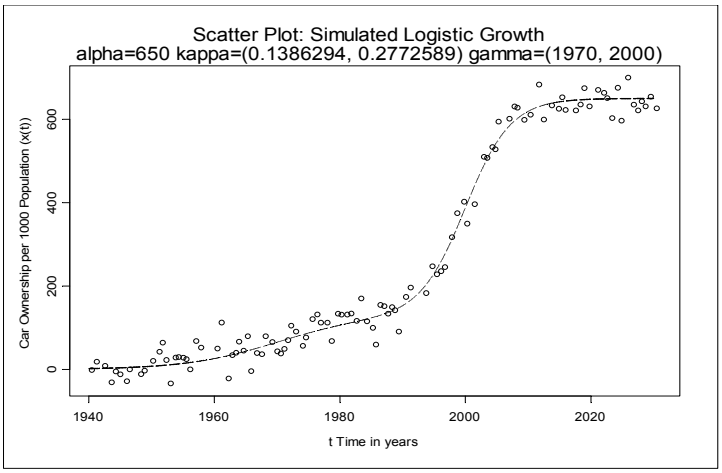

Figure 6

Figure 6 is the **scatter plot of the simulation of the non-homogeneous population**. This simulation assumes both a delay effect and a corrective intervention effect of doubling the growth rate parameter for the previously disadvantaged group.

1. Disadvantaged population: α=650, γ=2000, κ=0.2773;
2. Advantaged population: α=650, γ=1970, κ=0.1387.
3. The dashed curve is the simulated ownership values.

11. SUMMARY OF CONCLUSIONS

1. The difficulties presented by poor quality data can be mitigated by the use of LOWESS and robust regression. The procedure is fully objective and automatic. Consequently the need for the removal of outliers has been eliminated.
2. The virtual sample method has enabled the time parameter to be inferred and hence predictions of car ownership levels to be made.
3. Non-Linear regression has been shown, on simulated data, to be a more accurate numerical method.
4. Using simulated data, restricted to the early growth phase, it has been shown that there are large under estimates of all three parameters.
5. Significant departures from the Logistic DE model have been shown for simulated heterogeneous population data.

12. WAY FORWARD

1. The difficulties presented by poor quality data can be mitigated by the use of appropriate methods of analysis.
2. In the medium term consideration should be given to sample survey methods to supplement and audit the administrative data.
3. In the long term consideration should be given to bringing the administrative process more closely in line with the needs of transport planning while meeting the needs of other stakeholders.

13. REFERENCES

[1] Arrowsmith, DK and Place, CM, 1992. “Dynamical Systems”, Chapman & Hall

[2] Fan, J. and Gijbels, I. !996. “Local Polynomial Modelling and Its Applications”, Chapman & Hall

[3] Seber, GAF and Wild CJ, 1989. “Nonlinear Regression”, John Wiley & Sons

[4] “S-Plus 2000 Guide to Statistics vol. 1” 1999, MathSoft Inc.

[5] Yohai, V.J. 1987, “High Breakdown and High Efficiency Robust Estimates for Regression”,

[6] The Annals of Statistics vol. 15, no. 20, pp. 642-656